Boris Otto
Michael ten Hompel
Stefan Wrobel *Editors*

Designing Data Spaces

The Ecosystem Approach to Competitive Advantage

OPEN ACCESS

Springer

Designing Data Spaces

Boris Otto • Michael ten Hompel • Stefan Wrobel
Editors

Designing Data Spaces

The Ecosystem Approach to Competitive Advantage

Springer

Editors
Boris Otto
Fraunhofer-Institut für Software- und Systemtechnik ISST
Dortmund, Germany

Michael ten Hompel
Fraunhofer-Institut für Materialfluss und Logistik IML
Dortmund, Germany

Stefan Wrobel
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Sankt Augustin, Germany



ISBN 978-3-030-93974-8 ISBN 978-3-030-93975-5 (eBook)
https://doi.org/10.1007/978-3-030-93975-5

© The Editor(s) (if applicable) and The Author(s) 2022. This book is an open access publication.
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Digital transformation is changing the way we live and work with incredible speed. Data is one driver of this change and opens up entirely new opportunities. This makes data the most significant resource for our future. Being able to access, use, and combine data is the key to prosperity and innovation in Germany and worldwide. Safeguarding people's sovereignty over their own data is and will be vital for creating trust and confidence among actors in the data economy.

This is why the Federal Ministry of Education and Research (BMBF) has been funding research projects on the secure and sovereign exchange of data in the private sector since 2015. The Fraunhofer initiative for International Data Spaces (IDS) is an outstanding example in this context. This initiative provides the technological basis and fundamental infrastructure for secure data exchange and facilitates the development of new data-driven business models and applications. It has resulted in the successful foundation of the International Data Spaces Association (IDSA).

The fair and sovereign exchange of data has since increased in importance across all sectors of the economy. It has also raised the visibility of the enormous potential of data sharing for spurring cross-sector innovation and the development of new business models. IDSA now counts more than 130 members from over 20 countries and is an important partner in the BMBF-supported Gaia-X initiative that is aimed at establishing a distributed data infrastructure in Europe. These two initiatives, IDS and Gaia-X, are vital for creating domain-specific data spaces and for their interconnection in a European data and infrastructure ecosystem that is characterized by sovereignty and interoperability.

It therefore does not come as a surprise that the concept of the IDS has already been successfully applied in a number of initiatives and data spaces that are being developed under the umbrella of Gaia-X. Examples include the Catena-X ecosystem and the Mobility Data Space, which enables the self-determined exchange of data in the transport sector in order to encourage the development of innovative, eco-friendly, and user-friendly transport schemes. These projects contribute to Germany's and Europe's technological sovereignty and demonstrate what cooperation between equal partners along the entire value chain can look like—that is,

between manufacturers and suppliers, dealers' associations and equipment manufacturers, and providers of applications, platforms, and infrastructure.

There is also enormous further potential that we want to take advantage of. We can take data spaces and data ecosystems in the digital age to a new level of knowledge and technology transfer. To this end, we need substantive, interoperable data spaces in industry and science that can be interlinked flexibly, based on demand and the respective business model, and enable reliable and sustainable data-driven economic activity. This is the objective of the BMBF's FAIR Data Spaces project launched in 2021. It is aimed at establishing a common cloud-based data space for science and industry. The Gaia-X initiative and the National Research Data Infrastructure (NFDI) will collaborate even more closely to achieve this. After all, research data is a key resource not only for science but also for industry and society. By further expanding and interlinking data spaces, we are laying an important foundation for progress and innovation.

All this shows that we are on the right path and have already achieved several successes.

It is equally clear that we still have work to do, particularly with regard to research and development of technologies we need for operating and using data spaces. New technologies for the collection, processing, and provision of data will play a major part in this context. However, we must also retain our bright minds in the data economy and ensure large-scale training and skills development for more excellent professionals.

This will allow Germany to remain a modern country of innovation in a competitive European Union.

Federal Ministry of Education and Research Wolf-Dieter Lukas
Berlin, Germany
24 September 2021

Preface

Data is a strategic resource for business innovation and societal prosperity. The good news is that this resource is growing. The volumes of data produced by smart devices which determine all areas of our daily lives are exponentially increasing. Making use of these amounts of data bears a huge value potential. However, today we don't make optimal use of the data. Reasons for that are multifold and comprise data protection and regulatory aspects, lack of trust between data providers and data users, as well as limited accessibility and availability of data.

Data platforms address these challenges as they form a marketplace in which data providers and data users meet. Following the logic of network effects, data platform providers are keen in maximizing the data offering in their marketplace. Thus, they are investing large amounts of money to access data from various sources to increase their attractiveness for data users.

It is a matter of fact that most large data platform providers come from outside Europe. Because of the inherent business model logic of data platforms—e.g., winner-takes-it-all effects and gatekeeping functionality—typical values of the Single European Market such as freedom of choice and a fair compensation for the data provider in secondary use scenarios are at risk in the data economy.

Thus, Europe should act now to make sure the European data economy is designed according to what German Chancellor Angela Merkel once called the Social Data Market Economy. This model balances individual interests of the data provider and common interests in using the data. It rests on principles such as sovereignty over data, trust among market participants, and interoperability and portability of data.

Initiatives such as GAIA-X and International Data Spaces (IDS) are promising endeavors toward a European way in the data economy as they set standards and provide software architectures for fair data sharing and trusted data spaces. GAIA-X is a not-for-profit industry association headquartered in Brussels which aims at a federated data and service infrastructure for Europe. The IDS Association contributes to GAIA-X by providing architectures and open-source software implementations for data sovereignty technologies.

Back in 2015, the idea was born by a consortium of different stakeholders from industry, government, and research. From its beginning, IDS focused on facilitating trustworthy and sovereign data sharing in business ecosystems. The German Federal Ministry for Education and Research (BMBF) funded a research project to design the architecture of the IDS. In parallel, the IDS Association was founded to bundle the interest of the different industrial domains, to provide use cases to test the architecture design, and to foster adoption and standardization activities. A first version of the Industrial Data Space Reference Architecture Model (IDS RAM) was publicly presented at Hannover Messe 2017, where 26 research institutions and companies already showed first demonstrators of solutions based on aspects of the reference architecture. In total, the IDS Association currently has over 130 organizational members from 22 countries.

Today, data spaces are a central means for the implementation of the European strategy on data. From a computer sciences standpoint, they represent a federated data architecture approach which requires no central data store but provides software functionality for data sovereignty and trust. From a business point of view, data spaces are a multilateral form of collaboration on data which is why communities in the mobility or manufacturing domain form and take ownership of domain-specific data spaces as called for in the European strategy on data. Finally, data spaces are also objects of regulation. The European Data Governance Act defines provisions for data sharing services offered through and via dataspaces.

It is important to understand that domain-specific data spaces in Europe will not be implemented in a top-down approach with one large initiative taking care of all the various stakeholder interest and then deliver everything as one dedicated project. In contrast, the European data space for a certain domain will materialize as the entirety of a variety of different bottom-up endeavors which share the same principles and are interoperable.

To accomplish a Social Data Market Economy in Europe, it is therefore essential to put a common data infrastructure into place with interoperable and compatible identity management, data cataloguing schemes, and data sovereignty technologies.

This edited book gives room for elaborating on the various perspectives of data spaces. It presents implementations and business cases of data spaces and gives an outlook to future developments. The book addresses readers both from the scientific and the practitioners' community. In doing so, it aims at proliferating the vision of a social data market economy based on data spaces which embrace trust and data sovereignty.

Dortmund, Germany — Boris Otto
Dortmund, Germany — Michael ten Hompel
Sankt Augustin, Germany — Stefan Wrobel

Contents

Part I Foundations and Context

1 The Evolution of Data Spaces 3
Boris Otto

2 How to Build, Run, and Govern Data Spaces 17
Lars Nagel and Douwe Lycklama

3 International Data Spaces in a Nutshell 29
Heinrich Pettenpohl, Markus Spiekermann, and Jan Ruben Both

4 Role of Gaia-X in the European Data Space Ecosystem 41
Hubert Tardieu

5 Legal Aspects of IDS: Data Sovereignty—What Does It Imply? 61
Alexander Duisberg

6 Tokenomics: Decentralized Incentivization in the Context of Data Spaces 91
Jan Jürjens, Simon Scheider, Furkan Yildirim, and Michael Henke

Part II Data Space Technologies

7 The IDS Information Model: A Semantic Vocabulary for Sovereign Data Exchange 111
Christoph Lange, Jörg Langkau, and Sebastian Bader

8 Data Usage Control 129
Christian Jung and Jörg Dörr

9 Building Trust in Data Spaces 147
Monika Huber, Sascha Wessel, Gerd Brost, and Nadja Menz

10 Blockchain Technology and International Data Spaces 165
Wolfgang Prinz, Thomas Rose, and Nils Urbach

11 Federated Data Integration in Data Spaces 181
Matthias Jarke and Christoph Quix

12 Semantic Integration and Interoperability 195
Sören Auer

13 Data Ecosystems: A New Dimension of Value Creation Using AI and Machine Learning 211
Dirk Hecker, Angelika Voss, and Stefan Wrobel

14 IDS as a Foundation for Open Data Ecosystems 225
Fabian Kirstein and Vincent Bohlen

15 Defining Platform Research Infrastructure as a Service (PRIaaS) for Future Scientific Data Infrastructure 241
Yuri Demchenko, Cees de Laat, Wouter Los, and Leon Gommans

Part III Use Cases and Data Ecosystems

16 Silicon Economy: Logistics as the Natural Data Ecosystem 263
Michael ten Hompel and Michael Schmidt

17 Agricultural Data Space 279
Ralf Kalmar, Bernd Rauch, Jörg Dörr, and Peter Liggesmeyer

18 Medical Data Spaces in Healthcare Data Ecosystems 291
Thomas Berlage, Carsten Claussen, Sandra Geisler, Carlos A. Velasco, and Stefan Decker

19 Industrial Data Spaces 313
Thomas Usländer and Andreas Teuscher

20 Energy Data Space 329
Volker Berkhout, Carsten Frey, Philipp Hertweck, David Nestle, and Manuel Wickert

21 Mobility Data Space 343
Sebastian Pretzsch, Holger Drees, and Lutz Rittershaus

Part IV Solutions and Applications

22 Data Sharing Spaces: The BDVA Perspective 365
Edward Curry, Tuomo Tuikka, Andreas Metzger, Sonja Zillner, Natalie Bertels, Charlotte Ducuing, Davide Dalle Carbonare, Sergio Gusmeroli, Simon Scerri, Irene López de Vallejo, and Ana García Robles

23 Data Platform Solutions 383
Fabrice Tocco and Laurent Lafaye

24 FIWARE for Data Spaces 395
Ulrich Ahle and Juan Jose Hierro

25 Sovereign Cloud Technologies for Scalable Data Spaces 419
Wieland Holfelder, Andreas Mayer, and Thomas Baumgart

26 Data Space Based on Mass Customization Model 437
Lucheng Chen, Haiqin Xie, Wen Yang, and Lin Xiao

27 Huawei and International Data Spaces 451
Martin A. O'Brien, David Mohally, Götz P. Brasche, and Andrea G. Sanfilippo

28 International Collaboration Between Data Spaces and Carrier Networks 471
Akira Sakaino

29 From Linear Supply Chains to Open Supply Ecosystems 485
Fabian Biegel and Nemrude Verzano

30 Data Spaces: First Applications in Mobility and Industry 493
Christoph Schlueter Langdon and Karsten Schweichhart

31 Competition, Security, and Transparency: Data in Connected Vehicles 513
Karsten Schulze

32 Data Space Functionality 521
Douwe Lycklama

33 The Energy Data Space: The Path to a European Approach for Energy 535
Martine Gouriet, Hervé Barancourt, Marianne Boust, Philippe Calvez, Michael Laskowski, Anne-Sophie Taillandier, Loïc Tilman, Mathias Uslar, and Oliver Warweg

Glossary 577

Abbreviation

AAI	Authentication and Authorization Infrastructure
AGV	Automatic guided vehicle
AI	Artificial intelligence
API	Application programming interface
B2B	Business to business
BI	Business intelligence
BSI	Bundesamt fuer Sicherheit in der Informationstechnik (Federal Office for Information Security)
CA	Certificate authority
CI/CD	Continuous integration/continuous deployment
CPO	Charge point operator
CRM	Customer relationship management
DCAT	Data Catalog Vocabulary
DER	Distributed energy resources
DES	Digital energy resources
DGA	Data Governance Act
DSO	Distribution system operators
DWH	Data warehouse
EHR	Electronic health record
ELT	Extract-Load-Transfer
ERP	Enterprise resource planning
ESN	Echo state network
ETL	Extract, Transform, Load
EV	Electric vehicle
G2B	Government to business
GIS	Geographic Information System
HRS	Hydrogen refueling station
IaaS	Infrastructure as a Service
ICT	Information and communication technology
IDM	Identity management
IDS	International Data Spaces

IDSA	International Data Spaces Association
IGN	National Geographical Institute
IIoT	Industrial Internet of Things
IoT	Internet of Things
IOWN	Innovative Optical and Wireless Network
IPCEI	Important Projects of Common European Interest
IPR	Intellectual property rights
ISO	International Organization for Standardization
ISV	Independent Software Vendor
KPI	Key performance indicator
MES	Manufacturing execution system
ML	Machine learning
NIST	National Institute of Standards and Technology (USA)
OCSP	Online Certificate Status Protocol
ODRL	Open Digital Rights Language
OEM	Original equipment manufacturer
OIS	Oncology information system
OPC	Open Platform Communication
OPC-UA	Open Platform Communications-Unified Architecture
PaaS	Platform as a Service
PKI	Public key infrastructure
PLM	Product lifecycle management
PV	Photovoltaic
RAM	Random-access memory
RAM	Reference architecture model
RCA	Root cause analysis
RDF	Resource Description Framework
ROS/ROS2	Robotic Operating System
SaaS	Software as a Service
SCORVoc	Supply-chain operation reference
SD-Exchange	Software-Defined Exchange Service
SDG	Sustainable Development Goal
SE	Secure Element
SEV	Secure Encrypted Virtualization
SGX	Software Guard Extensions
SLA	Service-level agreement
SME	Small and medium-sized enterprises
STREAM	Sovereign, trusted, reusable, exchangeable, actionable, and measurable data properties
TCB	Trusted computing base
TEE	Trusted execution environment
TPM	Trusted Platform Module
TPM	Technological protection measures
TSO	Transmission System Operators

TZ	TrustZone
UC	Use case
V2G	Vehicle-to-grid
VPC	Virtual Private Cloud
VPN	Virtual private network
VSE	Very small enterprise
W3C	World Wide Web Consortium
WG	Working group
ZVEI	Zentralverband Elektrotechnik- und Elektronikindustrie e.V./ German Electrical and Electronic Manufacturers' Association

Part I
Foundations and Context

Chapter 1
The Evolution of Data Spaces



Boris Otto

Abstract The role data plays in enterprises is changing as the digital transformation in many sectors gains speed. New business opportunities through data-driven innovation emerge from data sharing in ecosystems. In ecosystems, the interest of the individual must be brought into alignment with the interest of the ecosystem. Trust between participants, data interoperability, and data sovereignty are key requirements which can be met by data spaces. Data spaces are a distributed data integration concept which is taken up by consortia aiming at supporting ecosystem. GAIA-X and IDS specify reference architectures for distributed data infrastructures and data spaces, respectively. While the benefits of data spaces for a fair data economy are recognized by business and policy makers, a deeper understanding is required about the design and evolution of data spaces. This chapter introduces fundamental concepts, identifies design tasks and options, and, thus, provides guidance for the establishment of data spaces.

1.1 Data Sharing in Data Ecosystems

1.1.1 The Role of Data for Enterprises

The role of data for businesses has continuously changed over the last decades. The proliferation of the platform economy and the continuing consumerization of many areas of business, in particular industrial activities, are only two examples of developments which require enterprises to rethink the way they manage data—both internally and together with external partners.

When it comes to the role of data for enterprises today, four types of roles can be identified. First, data is still—as it has been over the last decades—an enabler of operational excellence within a company. Integration and automation of business processes requires an effective and efficient data resource management. Second, data

B. Otto (✉)
Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany
e-mail: boris.otto@isst.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_1

has become a product which is sold in the market. For example, mobile telecommunication providers are used to sell anonymized data on the mobility and movement of their customers. The information generated from this data can be used, for example, by transportation and traffic agencies for maintenance plans of highway infrastructures or for better traffic management. Third, data is a source of business innovation. Data-driven services in various domains require access and combination of data from different sources provided by different members in a business ecosystem. Original equipment manufacturers (OEM) of industrial machinery, for example, collaborate with their customers, with service operators, component suppliers, and big data analytics firms to be able to offer better end-to-end maintenance services to their customers. And fourth, data is considered a strategic resource for the long-term sustainability of the economy. The European Union, for example, estimates the value of the data economy at 550 billion euros at minimum [1].

The value of data can only unfold when data is used [2]. Therefore, both policy makers and the private sector have an interest in increasing the use, and the re-use, of data. In particular, industrial enterprises possess a "treasure of data" which stems from manufacturing processes, from condition monitoring of installed base products and from the customer experiences. However, holders of this data have also an interest in not getting exploited when it comes to data. Making data available—in particular data of good quality—is not for free but comes at a cost. Thus, data providers must have a stake in the question on how their data should be used by others.

1.1.2 Data Sharing and Data Sovereignty

Data sharing goes beyond exchanging data. Data exchange has been happening between companies for more than 40 years. Electronic Data Interchange (EDI) is a good example. Message standards such as ANSI X.12 or EDIFACT define data exchange messages for trade and commerce and standardize purchase orders, invoices, dispatch notifications, etc. However, collaboration between EDI partners does not go beyond the pure exchange of data. The use of the data is limited for internal purposes of the involved parties. Data sharing, in contrast, means a collaborative use of the data for a shared goal [3]. Examples are collaborative customer innovation or efficiency gains for an entire group of companies.

In this context, ecosystems are a multilateral form of organizing to achieve a shared goal. In contrast to networks, for example, ecosystems are more centered around a customer innovation, more dynamic when it comes to their composition and characterized by a balance of prosperity of the individual members on the one hand and the ecosystem as a whole on the other hand [4].

Customer innovation is often realized through end-to-end support of a customer process. Good examples can be found in the mobility domain. Intermodal mobility services which support mobile individuals to plan, orchestrate, and perform their trips from the start of their trip to their destination are only possible when different

members of the mobility ecosystem team up. Examples are local and regional public transport providers, railway companies and airlines, car rental and taxi services, hotel chains, etc.

To ensure fairness in the data sharing and, thus, sustainability of ecosystems, data sovereignty is a key prerequisite. Data sovereignty refers to the capability of a legal entity or natural person to determine and execute usage rights when it comes to their data.

Data spaces support data sharing and data sovereignty in ecosystems as they are based on a distributed software infrastructure which provides the required software functionality.

1.1.3 Example Mobility Data Space

Figure 1.1 shows the architecture model of a mobility data space which aims at enabling data-driven mobility services, for example, intermodal mobility as a service (MaaS) as mentioned above.

The architecture consists of three layers. The mobility ecosystem represents the first layer on which services are offered to the mobile citizen. These services improve the mobility experience of the individual, allow for better management of traffic flows, and increase the utilization of means of transportation, among others.

The services on the first layer require a mobility data space in the narrower sense, i.e., a "shared digital twin" of the various constituents of mobility ecosystems. The digital twin consists of the data of the individual digital twins and comprises, for example, timetables, charging statuses of electric scooters, utilization of buses and trams, the travel preferences and plans for individuals, etc. The shared digital twin represents the second architecture layer.

Data sovereignty [5] must always be ensured for all participants in this data space. Data providers must have control and transparency of what happens to their shared data, and data consumers must be able to trust both data providers and data sources. A federated software infrastructure is needed as the third architecture layer. Relevant software services ensure data interoperability, data sovereignty, and trust among participants.

1.1.4 Need for Action and Research Goal

The example of the mobility domain shows the important role data spaces play both for companies to innovate and also for governments and policy makers to ensure data sovereignty on an economic level.

Both the European Data Strategy [6] and the Data Strategy of the Federal Government of Germany emphasize this important role. The European Data Strategy

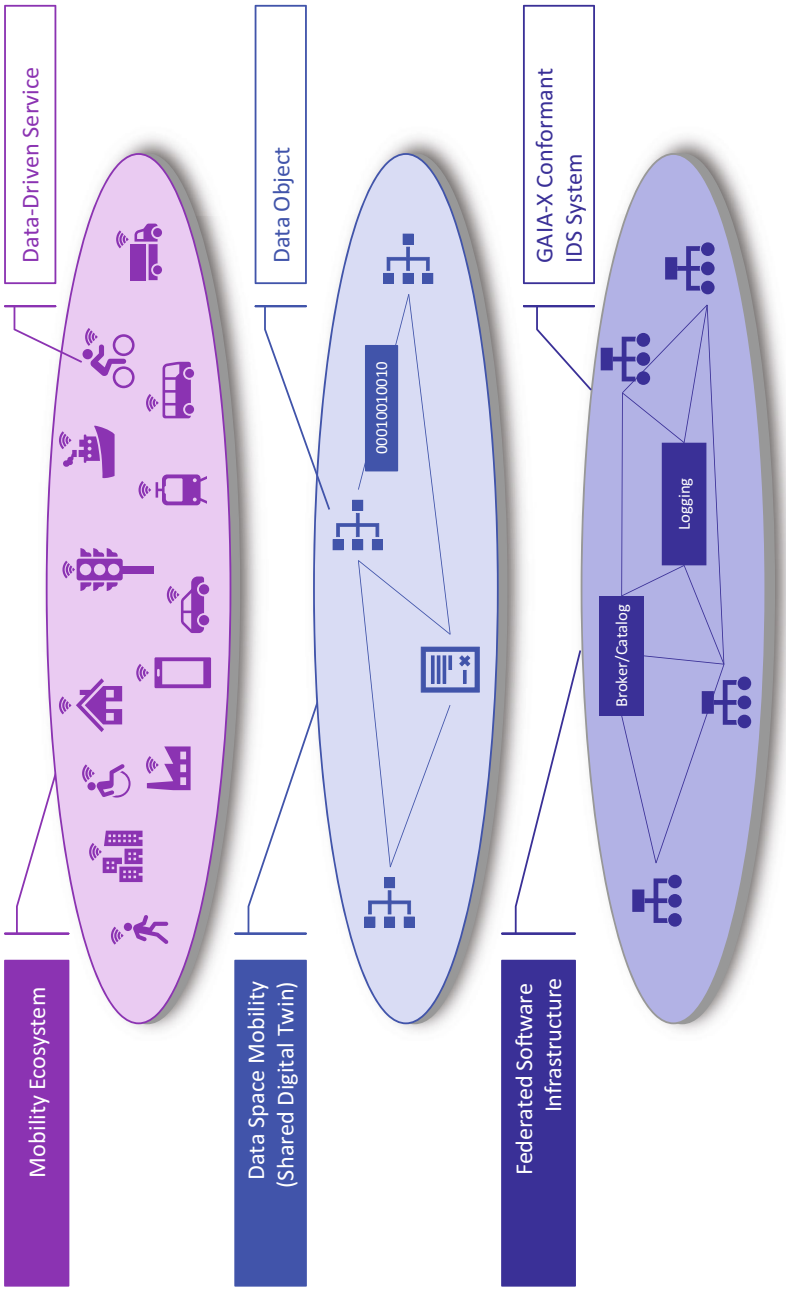


Fig. 1.1 Mobility data space (©2021, Fraunhofer ISST)

calls for the establishment of domain-specific data spaces in the Single European Market.

Data spaces are distributed by design and address not only individual companies but ecosystems or even entire domains. Their underlying software infrastructure must be in place before business innovation can be realized. It is an important cornerstone of the data economy mentioned above.

As with all infrastructures, many stakeholders are concerned with a set of questions when it comes to planning, designing, implementing, and maintaining this software infrastructures for data spaces. Typical questions are related to the functionality of the infrastructure services, to their openness, to their funding and financing, and to their governance. At present, many of these questions are still unanswered.

In this context, this chapter aims at laying some fundamental foundations for data spaces, to elaborate on the evolution of data spaces and to analyze the most important design tasks to be addressed. It helps both individual businesses to position themselves in existing and starting data space activities and it supports policy makers in their endeavor to pave the way for a fair data economy.

1.2 Conceptual and Technological Foundations

1.2.1 Data Spaces Defined

The notion of data spaces was coined some 15 years ago in computer science [7, 8]. Data spaces were introduced by that time as a data integration concept. In contrast to central data integration approaches (e.g., data consolidation hubs), data spaces do not require a physical integration of the data, but leave the data stored at the source. In addition to that, they do not require a common database schema that data from various sources must adhere to. Integration is rather achieved on semantic level using shared vocabularies, for example. Because of that, data spaces allow for data redundancies and "co-existence" of data. Furthermore, data spaces can be nested and overlapping so that individual participants can be part of multiple data spaces.

Besides this original technological definition of data spaces, the increased use of the term in the business community has led to an understanding of the data space notion as a form of collaboration on data. Practitioners in various industrial domains interpret data spaces as a business collaboration format driven by the desire to achieve shared goals. An example is Catena-X, the initiative launched by parts of the German automotive industry, which aims at a data space allowing for integrated trusted data chains in the automotive supply and production network. The business definition of data spaces refers to goals and decision-making rights and processes between the consortiums of participants.

In addition to that, the infrastructure nature of data spaces requires a common understanding of the concept from a legal point of view. Data spaces and their

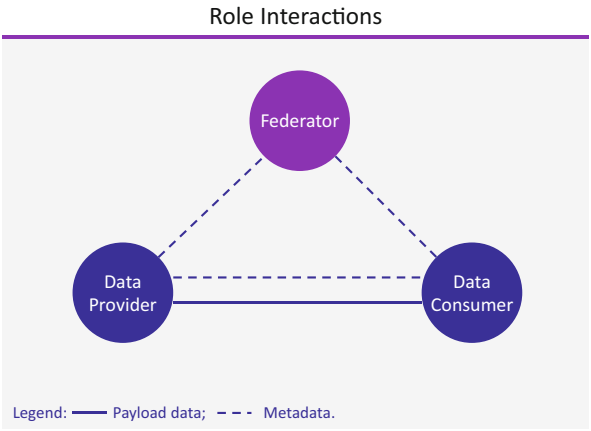


Fig. 1.2 Data space roles (©2021, Fraunhofer ISST)

underlying software infrastructures must support trust, interoperability, and portability of data and data sovereignty and must be nondiscriminatory. Thus, data spaces can be understood as intermediaries and data sharing service providers to which the EU Data Governance Act[1] applies which is currently under review.

1.2.2 Roles and Responsibilities in Data Spaces

Figure 1.2 shows the fundamental roles in a data space in their interactions between each other. As mentioned above, a data space is a distributed data integration concept. Thus, there is no central data store or data vault into which data providers deliver their data and from where it can be accessed and retrieved by data consumers. In contrast, the exchange of the data happens directly between the two participants.

However, to meet the fundamental requirements of data spaces in terms of trust among participants, data security, and interoperability, intermediary services are required. The role of the federator is to provide these intermediary services which include cataloging and brokering of data sources, ensuring trust between participants, and offering data sovereignty services.

Table 1.1 shows the responsibilities of the three roles when it comes to data sharing and exchange in data spaces.

[1] See Regulation of the European Parliament and of the Council on European data governance of 2020 [9].

Table 1.1 Data space roles and responsibilities

	Data provider	Data consumer	Federator
Publish data source	R		S
Search for data source		R	S
Register participant	R	R	S
Identify participant	R	R	S
Authorize participant	R	R	
Request data exchange	S	R	
Perform data exchange	R	S	
Log data exchange	S	S	R
Constrain data use	R		
Perform data use		R	

Legend: R, responsible; S, supportive

1.2.3 GAIA-X and IDS

The International Data Spaces (IDS) initiative was launched in 2015 with Fraunhofer research project funded by the German Federal Ministry for Education and Research. It aimed at the design and prototyping of a distributed software architecture for data sovereignty. In parallel, the IDS Association (IDSA)[2] formed as a not-for-profit industry association and took up the work of Fraunhofer research to further develop it into the IDS Reference Architecture Model (IDS RAM). The IDS RAM is a technology-agnostic architecture description for a data space software architecture. Central software components described in the IDS RAM (e.g., IDS Connector, IDS Broker, IDS Clearing House) found their way into DIN SPEC 27070 which provides a blueprint for a secure gateway for trusted data exchange.

Today, IDSA has more than 130 members from more than 20 countries. Major IDSA activities are the maintenance of the IDS RAM, the definition of the certification process and the role of the Certification Body, and the implementation of an open-source software (OSS) strategy.

The GAIA-X initiative[3] formed as a response to a call for a data infrastructure articulated in the German strategy on artificial intelligence. GAIA-X aims at data sovereignty in a broader context as IDSA does, because within the GAIA-X are not only data sharing and exchange but also the storing and handling of data on cloud platforms.

The Federation Services form the core of the GAIA-X architecture. They comprise a federated catalogue of distributed services, sovereign data exchange, identity and trust management, and compliance services.

The IDS and GAIA-X initiative are closely aligned in order to allow for seamless integration of the architectures and support processes [10].

[2] See https://internationaldataspaces.org/.

[3] See https://www.gaia-x.eu/.

1.3 Evolutionary Stages of Data Space Ecosystems

Data ecosystems are complex multilateral forms of organization which involve multiple different members. Therefore, ecosystems develop along evolutionary stages (see Fig. 1.3).

In a first stage, ecosystems are closed setups between a limited number of members. In this case, a separate entity of a federator is not necessarily needed, but the federator responsibilities are taken over by one of the members. An example of a closed ecosystem is the production and supply network of a single OEM in the automotive industry.

The second evolutionary stage is characterized by the openness of the ecosystem regarding their members. Participants are not always known but come and leave in a dynamic fashion. This leads to increased requirements when it comes to trust and interoperability, for example. The Catena-X initiative with its present scope is a good example for an open ecosystem because Catena-X is explicitly directed at the entire automotive supply chain with its multiple tens of thousands of companies.

A third evolutionary stage reflects the fact that individual members do not belong to one ecosystem only but are members of multiple ones. In this case, data sharing across the boundaries of an ecosystem must be possible which leads to additional requirements when it comes to trust among participants, interoperability of data and metadata, etc.

Table 1.2 shows different data space implementation options depending on the three evolutionary stages of ecosystems. Complexity in terms of interoperability, sovereignty, and trust and security increases along the evolutionary path.

To achieve interoperability of data space activities across ecosystem boundaries, for example, it must be made sure that federators agree on a unique system of identifiers and description schemes of data sources. Similarly, data usage control policies as a means enabling data sovereignty must be unambiguously understood across different ecosystems as well. The same holds true for digital identities which are used to identify and authenticate participants.

Because of that, an “ecosystem of federators” must be implemented—on top of the ecosystems of participants.

1.4 Designing Data Spaces

1.4.1 Ecosystem Perspective

As mentioned above, data spaces must be seen in the context of the ecosystem they support and the underlying software infrastructure. Taking an ecosystem view, design activities address the three architecture layers introduced in Fig. 1.1.

The business layer comprises incentive systems for participants and funding and financing models, for example. Even though data spaces follow a distributed design,

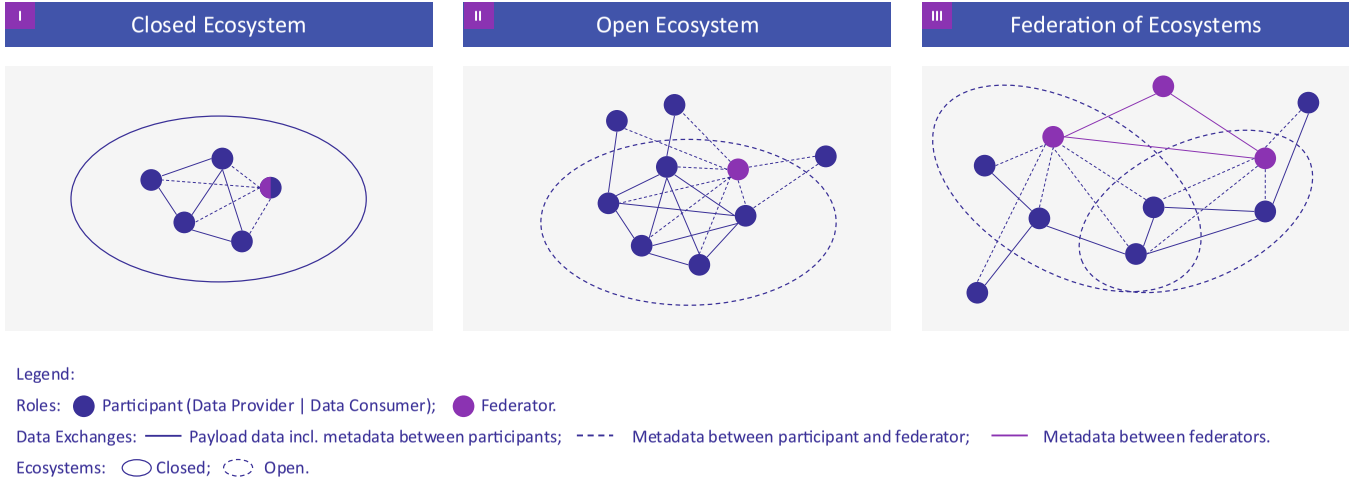


Fig. 1.3 Ecosystem evolutionary stages (©2021, Fraunhofer ISST)

Table 1.2 Data space implementations

	Closed ecosystem	Open ecosystem	Federation of ecosystems
Interoperability	• Proprietary schemas possible • Typical use of available domain-specific standards	• Domain-specific, open standards required • Common vocabularies advisable	• Cross-domain, open standards advisable • Mapping/translation between domain-specific standards • Uniqueness of identifiers (e.g. URIs) across domains needed
Sovereignty	• Traceability and transparency of data exchange not required if all participants are known	• Increasing demands for policy enforcement because of unknown participants • Policies to be unanimously understood within ecosystem • Policies to allow for automated negotiation	• Demand for policy enforcement because of unknown participants and cross-domain • Policies to be unanimously understood within ecosystem and to allow for automated negotiation
Trust and security	• Trust through the consortium • Digital certificates/ tokens not necessarily required	• Digital certificates/ tokens required because of unknown participants • Dynamic technologies required in case of sensitive data use cases	• Digital certificates/ tokens to support cross-ecosystem application • Uniqueness of identities required

they form platforms for ecosystem. Therefore, fundamental economic characteristics of the platform economy such as network effects do also apply for data spaces. The data space services must be attractive for data providers, and data consumers and incentive systems must be put into place to achieve a critical mass of data providers. The more data sources are available through the data space, the more attractive it will be for data consumers.

Furthermore, funding and financing models must consider the infrastructural nature of data spaces. Investments in data spaces must be made before data-driven services can flourish. As there is both a public (and community) and a private (or individual) interest in the existence of data spaces, data space funding should come both from public and private sources.

The governance layer comprises questions of ecosystem governance and collaborative data governance. The first is related to the institutionalization of data space consortia. At present, one can observe the creation of joint ventures with or often without profit-oriented business purposes. The Mobility Data Space initiative in Germany, for example, plans to establish a not-for-profit limited liability company.

For data ecosystems which use data from individuals (e.g., consumers), calls are articulated to create so-called data trusts in which many individual data providers pool their interest in order to form a counterweight to large data platform providers.

Organizational governance and data governance are closely interrelated in data space organizations. Collaborative data governance mechanisms need to be put into place to determine data visibility, transparency, and sovereignty in data spaces. For example, a data provider may allow any data space participant to find their data source or limit it to a dedicated group of users. Central element of collaborative data governance are rules which are commonly understood when it comes to articulating usage conditions on the data. These rules can be understood as "terms and conditions" in the data economy.

On the technology layer, shared vocabularies must be established on multiple levels. First, a vocabulary must be in place unambiguously describing the key concepts of a data space (e.g., roles, responsibilities, characteristics of data resources, usage policies, etc.). Second, in order to achieve interoperability not only with regard to the exchange of data but also with regard to the shared use of data, vocabularies are needed to harmonize the understanding of the meaning of "payload" data itself. A payload data example in the mobility domain are timetables which need to be consistently understood between the different data space participants.

Furthermore, the software infrastructure must be built. The GAIA-X architecture in combination with the IDS RAM forms a "blueprint" for data space implementation.

1.4.2 Federator Perspective

A key role in data spaces is the federator. When designing a data space, the instantiation of the federator role is critical.

Design tasks comprise among others:

- Portfolio of data space services: In most cases, the portfolio comprises services described in architecture proposals such as the GAIA-X Architecture Document and the IDS RAM. However, on top of that, further services may be needed and desired. Examples of such additional data services are data traceability, data quality assurance, data trustee services, as well as mapping services.
- Degree of decentralization: Services provided by the federator enable data exchange and sharing of participants. Thus, they ensure the functioning of distributed data space designs. However, federator services can be designed and implemented in a decentralized way as well. Federator services can be:
  - Distributed, i.e., focusing on interoperability
  - Federated, i.e., focusing on being perceived "as one"
  - Shared, i.e., being implemented once for all participants.
- Data Space Business Services: The software services of the federator are the core of the service portfolio of a data space organization. However, there may be a

need for additional services. Examples are on-boarding services for participants, integration services, and value-added business services such as billing, etc.

The individual design of the federator role depends on the purpose of the data space, the scope of the shared goal of the participants, and its legal and regulatory environment.

1.5 Summary and Outlook

This chapter motivates data spaces as a suitable instrument to establish data ecosystems based on fair data sharing, i.e., on trust between participants, data sovereignty, and data interoperability. It introduces fundamental concepts necessary to understand data spaces from a technological, business, and legal point of view and outlines stages of data space evolution. Furthermore, it identifies important design task which can serve both business consortia and policy makers in their endeavors to establish data spaces in a certain domain.

As the establishment of data space is still an emerging topic, the scientific body of knowledge is still in its infancy when it comes to understanding design options in detail and the factors that influence the design. Moreover, knowledge does not exist about the growth and adoption of data spaces.

Consequently, there is a strong need for research in different disciplines such as computer science, information systems, and management science but also for interdisciplinary studies.

References

1. European Commission. (2020). *The European data market monitoring tool*. European Commission.
2. Otto, B. (2015). Quality and value of the data resource in large enterprises. *Information Systems Management, 32*(3), 234–251. https://doi.org/10.1080/10580530.2015.1044344
3. Capiello, C., Gal, A., Jarke, M., & Rehof, J. (2020). *Data ecosystems: Sovereign Data Exchange among organizations (Dagstuhl Seminar 19391)*. https://doi.org/10.4230/DagRep.9.9.66.
4. Gelhaar, J., Groß, T., & Otto, B. (2021). A taxonomy for data ecosystems. In: *Hawaii International Conference on System Sciences 2021* (pp. 6113–6122). University of Hawai'i at Manoa, Hamilton Library.
5. Hummel, P., Braun, M., Tretter, M., & Dabrock, P. (2021). Data sovereignty: A review. *Big Data & Society, 8*(1). https://doi.org/10.1177/2053951720982012
6. European Commission. (2020). *A European strategy for data*. European Commission. Available online at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0066&from=EN. Accessed 4/25/2021.

7. Franklin, M., Halevy, A., & Maier, D. (2005). From databases to dataspaces. *ACM SIGMOD Record, 34*(4), 27–33. https://doi.org/10.1145/1107499.1107502
8. Halevy, A., Franklin, M., & Maier, D. (2006, June 26–28). Principles of dataspace systems. In: G. Gottlob & J. van den Bussche (Eds.), *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems—PODS '06, the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium, Chicago, IL, USA* (pp. 1–9). ACM Press.
9. Regulation of the European Parliament and of the Council on European data governance of 2020.
10. IDS Association & Fraunhofer. (2021). *GAIA-X and IDS*. International Data Spaces Association; Fraunhofer-Gesellschaft.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 2
How to Build, Run, and Govern Data Spaces



Lars Nagel and Douwe Lycklama

Abstract This chapter is a result of a collaborative effort between data space and industrial domain experts to define cross-sectoral and across initiatives fundamental design principles to build data spaces. A joint paper of this dimension is unique and a great step with regard to the convergence of the large initiatives on data sharing in Europe. The starting point is the implementation in five years of the vision defined by EU Strategy for Data. In that world multiple data spaces will have been widely adopted across Europe, organizations and individuals will have control over their data and a digital world with less dominance of, and dependency on, large, quasi-monopolistic players has been formed. In the following, the way forward is elaborated with fundamentals of data spaces as well as common building blocks describing how the design principles for all sectors are applied to reveal sector-specific benefit. Furthermore, a proposal for governance and business models for data spaces on a collaborative and individual level is presented. Finally, the roadmap for co-creating the soft infrastructure underlying European data spaces is drawn.

2.1 Data Space Design Principles

This chapter contains the essence from the position paper "Design principles for data spaces" (https://design-principles-for-data-spaces.org/) from taskforce 1 of the European-funded [1] coordinating and support action "OPEN DEI" (https://www.opendei.eu/). Here is the vision: Five years from now, the EU Strategy for Data will have been fully implemented, multiple data spaces will have been widely adopted across Europe, and European individuals and organizations will have regained the possibility of control over their data and, with that, their rightful and balanced place

L. Nagel (✉)
International Data Spaces e. V., Dortmund, Germany
e-mail: Lars.Nagel@internationaldataspaces.org

D. Lycklama
Innopay, Amsterdam, The Netherlands
e-mail: douwe@innopay.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_2

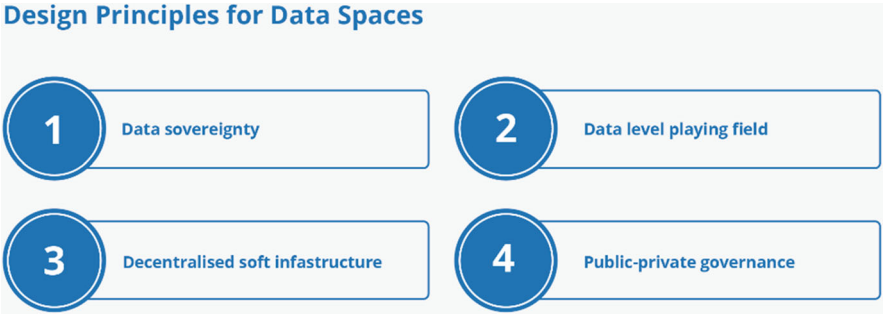


Fig. 2.1 Design principles for data spaces (©2021, International Data Spaces Association)

in the digital world. More initiatives will have been started and more value is captured in Europe. Ten years from now, this is mainstream, and the larger audience would not accept it any other way.

While the possibilities seem endless, European data spaces basically will bring about three new elements:

- Entirely new services for users, based on enhanced transparency and data sovereignty
- A level playing field for data sharing and exchange, leading to less dominance of, and dependency on, large, quasi-monopolistic players
- A new user behavior and digital culture, as users learn to play by the rules and use data (both their own and other users' data) in an ethical way

By sketching the vision and the approach to exploiting the potential of data spaces as specified above, these elements result in four design principles for European data spaces to be built on (see Fig. 2.1).

This will be the way for data spaces to become a solid and sustainable foundation for the next growth cycle within the digital economy of the EU.

2.1.1 Entirely New Services for Users Based on Enhanced Transparency and Data Sovereignty

While GDPR grants individuals the right to decide what data collectors are allowed to do with their personal data and what not, European data spaces will provide the tools to exert these rights and stay in control over that data. However, European data spaces will not do so for individuals only, but also for companies/organizations and their data. Driven by sector-specific needs, data spaces will promote the development of tools to share, exchange, and access all types of data, including data that is stored in smart objects and things. These tools will empower those entitled to the data to always demand transparency as to where their data is stored and what access

rights apply to it. They can use these tools to give or revoke their consent and to change access rights and specify new conditions of how their data can be accessed and used. Furthermore, they can choose to outsource data rights management to third parties (e.g., data intermediaries), just like users (individuals and organizations) today outsource the management of their financial balances to financial institutions (i.e., think investment profiles).

2.1.2 Level Playing Field for Data Sharing and Exchange

Ensuring a level playing field for all data space participants implies that new entrants face no insurmountable barriers (e.g., due to a quasi-monopolistic structure of the data ecosystem) when seeking admission to a data space. On a level playing field, players compete on the quality of their data and services, not on the amount of data they control. A level playing field for data sharing and exchange can emerge if such an ecosystem is ruled by the idea of cooperation instead of competition. This can be achieved by a sound design and thorough maintenance of the soft infrastructure underlying data spaces.

2.1.3 Need for Data Space Interoperability: The Soft Infrastructure

With its strategy for data, the Commission promotes the development of European data spaces for strategic economic sectors and public-interest domains.

While data spaces stimulate higher availability of data pools, technical tools, and infrastructure addressing domain-specific challenges and legislations, the EU Strategy for Data acknowledges that these data spaces should be interconnected and that this challenge requires specific attention. But Europe does not need to start from scratch—data sharing and exchange within specific domains and sectors is already happening in existing initiatives. However, each of these initiatives follows its own approach, and therefore they are not interoperable. So, part of the EU strategy should be to include and build upon existing data-sharing initiatives in the quest for interoperability and the specification of future "soft infrastructure" agreements.

Interoperability between domain-specific data spaces is crucial for two reasons. First, an individual or organization is never just part of one single space but operates in different spaces simultaneously. If data spaces are organized in silos, users have to adopt different solutions. This results in fragmentation, high integration costs, and monopolistic behavior of market participants. Second, use cases are not limited to a single data silo. Fragmentation of the data economy must be prevented to reap the maximum value for organizations and individuals in the EU.

The infrastructure for European data spaces will not be a monolithic, centralized IT infrastructure. Instead, it will be made of the totality of interoperable implementations of data spaces complying with a set of agreements in terms of functional, technical, operational, and legal aspects. Such a "soft infrastructure" will be invisible to data space participants. It will entail functional and nonfunctional requirements regarding interoperability, portability, findability, security, privacy, and trustworthiness.

Viewed from a technical standpoint, a soft infrastructure can be seen as a collection of interoperable, API-based IT platforms, where users control the flow of data through advanced mechanisms of identity and consent management. The design of the soft infrastructure will include mechanisms for economic exploitation of data sharing and exchange transactions (i.e., data monetization).

The soft infrastructure for data spaces will be technology-neutral, giving maximum freedom to all actors to make their own choices in accordance with their engineering capabilities.

2.1.4 Public-Private Governance: Europe Taking the Lead in Establishing the Soft Infrastructure in a Coordinated and Collaborative Manner

Europe is standing at a historic crossroads, demanding from us to decide about the next evolutionary step in the digital economy. This moment can be compared to the introduction of the GSM standard in the 1980s, which turned out to be the pivotal moment for the natural evolution of telecommunications, toward decentralization combined with innovation, competition, and accelerated adoption.

After 30 years of Internet infrastructure driven by private forces, it is time to balance private interests with public interest and create the next "GSM moment." Now we know better what we want and what we do not want in terms of our digital economy. The EU Strategy for Data and the Data Governance Act are essential cornerstones of this evolution, which will lead to a new organization of digital market forces. Public intervention means indicating the right direction, followed by activation of public and private energy in realizing this endeavor.

The recently proposed Data Governance Act confirms the notion of a governance structure constituted by multiple entities. For European data spaces, it is recommended to have a (domain) governance authority for each data space and a central governance authority overseeing all aspects in connection with interoperability of data spaces, i.e., the de facto "soft infrastructure." This central authority will interact with all data space-specific authorities. Therefore, $N \times$ data spaces plus one central—while relying on many shoulders and following harmonized approached—authority will need to be organized.

2.2 Building Blocks for Data Spaces

Now that we understand the fundamentals of data spaces and what is at stake, it is key to understand which elements together form data spaces in its archetypal nature. This chapter addresses a broad range of general building blocks that enable technical, business, operational, and organizational capabilities of data spaces from two perspectives: (1) the perspective of an essential soft infrastructure and (2) the perspective of the services that form data spaces within and across domains.

The design and implementation of a data space comprises a number of building blocks, which fall under two types: the technical building blocks and the governance building blocks (see Fig. 2.2).

2.2.1 Technical Building Blocks

The building blocks subsumed under this category enable the implementation of the technical architecture of a data space (see Fig. 2.3). They include network protocols, middleware components, (standardized) APIs, and more, facilitating the sharing of data between different parties in a secure and trustworthy fashion. A variety of technical components for building data spaces have been developed or adopted by different initiatives in Europe, such as FIWARE [2], Plattform Industrie 4.0 [3], CEF Digital [4], or the International Data Spaces Association [5].

Technical building blocks enable (plug and play) integration of different systems and platforms used by data space participants beyond the security limits of each participant. Additional technical building blocks may optionally be considered for facilitating creation of systems plugged into a data space (e.g., for implementing big-data analysis, supporting data visualization and analytics, or providing an

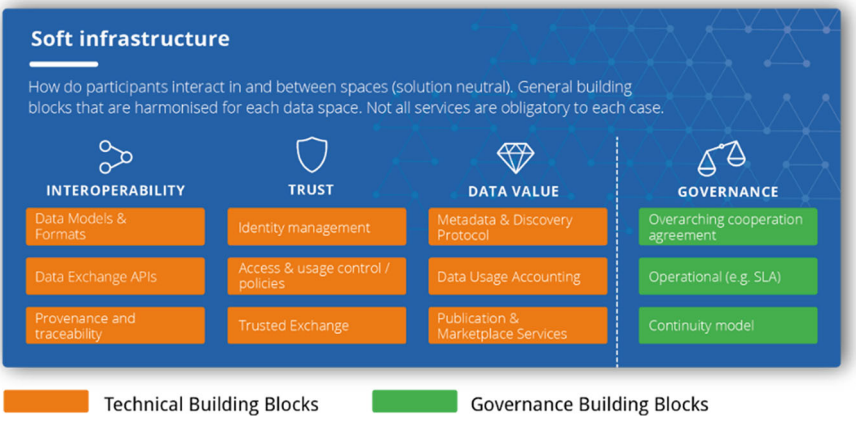


Fig. 2.2 Building blocks for data spaces (©2021, International Data Spaces Association)

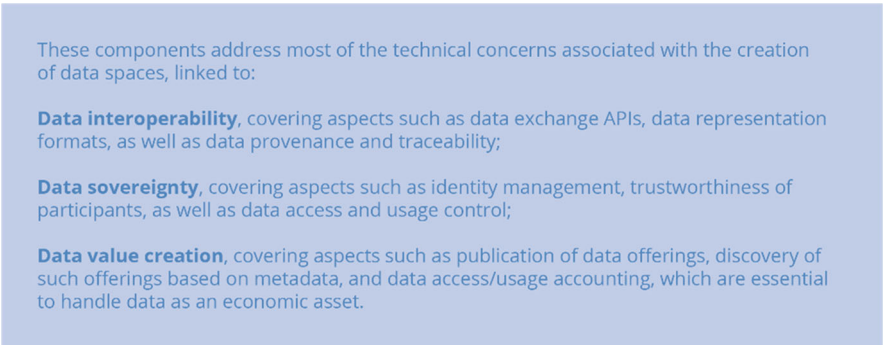


Fig. 2.3 Technical building blocks (©2021, International Data Spaces Association)

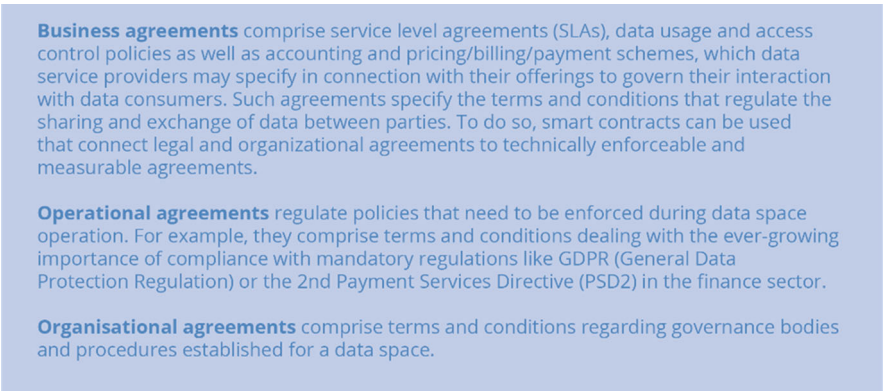


Fig. 2.4 Governance building blocks (©2021, International Data Spaces Association)

interface with IoT networks). These building blocks enable data usage in data spaces beyond current business capabilities of participants and lead to new business cases and data usage scenarios.

2.2.2 Governance Building Blocks

Governance building blocks refer to business, operational, and organizational agreements among data space participants (see Fig. 2.4). These agreements are enforced through legal frameworks participants have to adhere to, or via technical building blocks.

As data is a new type of asset that can be used and reused in different scenarios generating more or less business value (depending on context, availability, accuracy, etc.), common business models are not capable of adequately supporting the growing needs of business. It is therefore important to define and create a business

framework capable of supporting new business constellations. This would help the stakeholders of an ecosystem understand both the potential relationships between each other and the underlying business model(s). Jointly, with adequate rules and policies for sharing and using data in place, these data-driven business ecosystems will be able to create new business value more rapidly.

2.3 Synthesis of Building Blocks to Data Spaces

For integration of building blocks to data spaces, different sets of structuring principles can be applied to different architectures, depending on domain-specific requirements or technical requirements (e.g., streaming of data, high-frequency data, or event processing). Nevertheless, there are some guiding principles that need to be respected for all implementations, such as decentralization, scalability, collaboration support, federation, interoperability, compatibility, trust management, and auditability (see Fig. 2.5).

The different building blocks can be specified and developed independently of each other. When doing so, existing norms, standards, and best practices should be used to ensure cohesion of building blocks. Each data space solution can integrate multiple building blocks, as long as they are in line with data space reference architectures (e.g., the IDS Reference Architecture Model [6]). The building blocks are core elements of any data space. As such, they can be considered as sector-agnostic. Nevertheless, they can be used in sector-specific scenarios.

Data space stakeholders may also define additional building blocks to support innovative features and functions. For instance, data space architects may introduce building blocks that enable novel types of data space architectures combining centralized and decentralized approaches. Likewise, business stakeholders may introduce building blocks that enable novel forms of smart contracts to be agreed upon by participants of a data space, thereby facilitating business model innovations. Hence, the building blocks presented are not exhaustive, but rather indicative of the elements of a data space.

In general, each building block consists of reusable, generic components (i.e., which can be used across domains and industries) and more specific components (i.e., to meet requirements and regulations that are specific for certain industries, domains, or even concrete use cases). This allows individual participants to join different data spaces, use data in multiple contexts and scenarios, and be part of multiple data value chains.

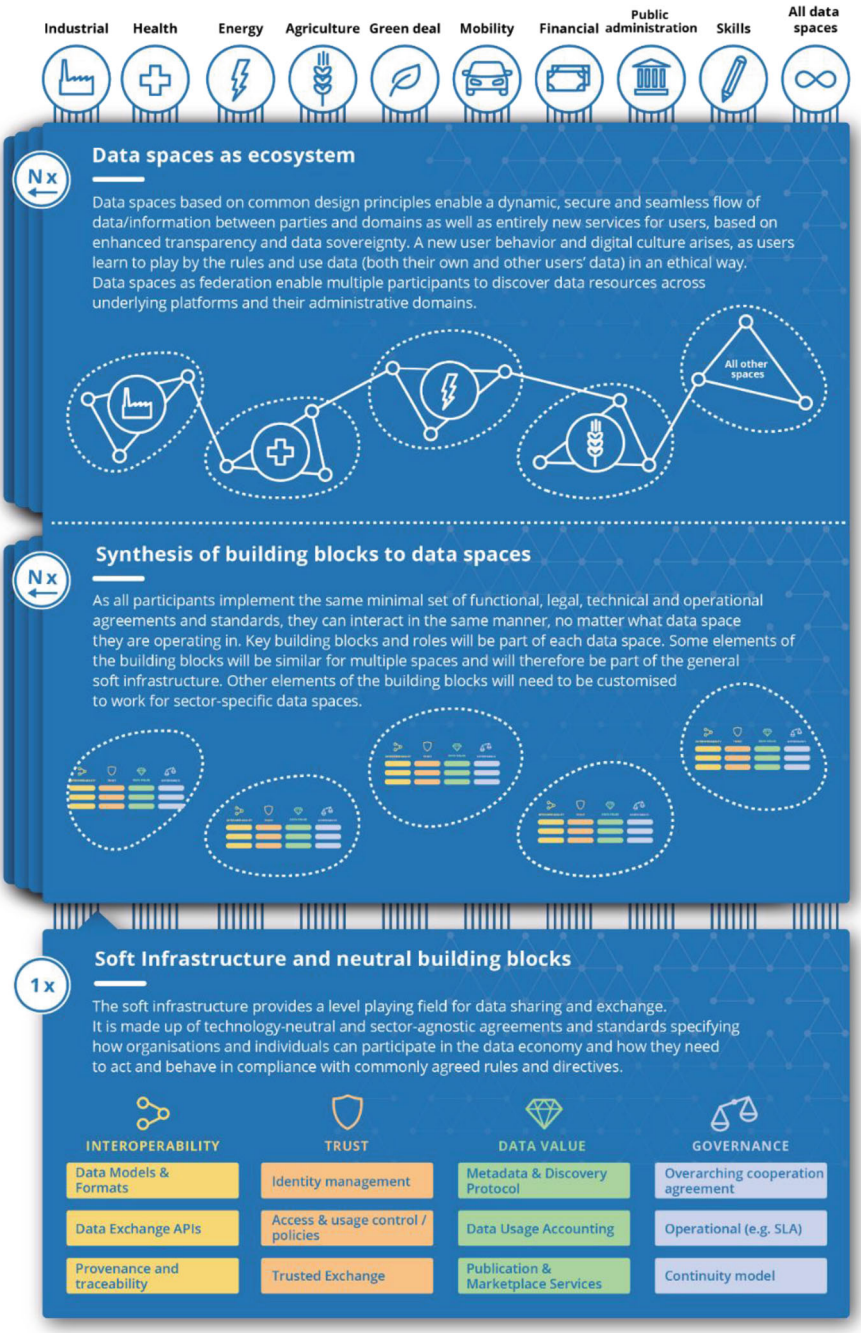


Fig. 2.5 Synthesis of building blocks to data spaces (©2021, International Data Spaces Association)

2.4 Harmonized Approach to Data Space Governance

Today's lack of a harmonized approach to establishing data spaces is more of a coordination and scaling problem than a technology problem. To set up data spaces that give users control over their data and interoperate with each other across sectors, adequate technology exists alongside with process knowledge to leverage it. What is required now is coordinated engineering and continuous maintenance, driven by sound European governance.

A data space is the total set of interoperable data-sharing applications by actors in a specific sector or domain, either by their own development or through a certified software vendor, data broker, or marketplace. It is adamant that from the onset the aim is that data spaces over time will systematically harmonize parts of their technical, operational, functional, and legal aspects, leading to the emergence of a uniform, de facto "soft infrastructure" ensuring cross-sectoral data space interoperability. This harmonization of common aspects in every data space into a soft infrastructure will enable users (citizens, businesses, governments) to stay in control of their data even across different sectors and applications (i.e., across different data spaces). This can be compared to the evolution of electronic payments in Europe (another special form of data sharing, unified by SEPA), which can be regarded as a soft infrastructure as well. It is a combination of rules and design decisions on top of an existing physical infrastructure of cables, services, and software stacks.

This soft infrastructure can only be achieved with good coordination—and good coordination comes with good governance. Good governance is about balancing the interest, input, and energy of private and public actors in order to ensure innovation and continuity in the long run. In this light, we must see the recently published Data Governance Act (DGA) as the enabling governance framework for European data spaces to be established.

2.5 The Way Forward and Convergence: Actions to Take in the Coming Digital Decade

The Commission is well positioned to take the lead in the coming decade in supporting the co-creation process of developing the data space soft infrastructure in a coordinated and collaborative manner, focusing primarily on governance (see Fig. 2.6). Continuous financing for a decade is crucial. Probably an IPCEI type of funding structure should be considered. The Data Governance Act has confirmed the importance of governance in such an endeavor. Each data space will have its own governance entity, while there will be an overall governance structure referring to all aspects that lead to interoperability of data spaces.

A lot of experience with data spaces is at hand, and a lot of research on the topic has been conducted in the past years. From a technology and process viewpoint, there is no doubt that data space interoperability and data sovereignty can be

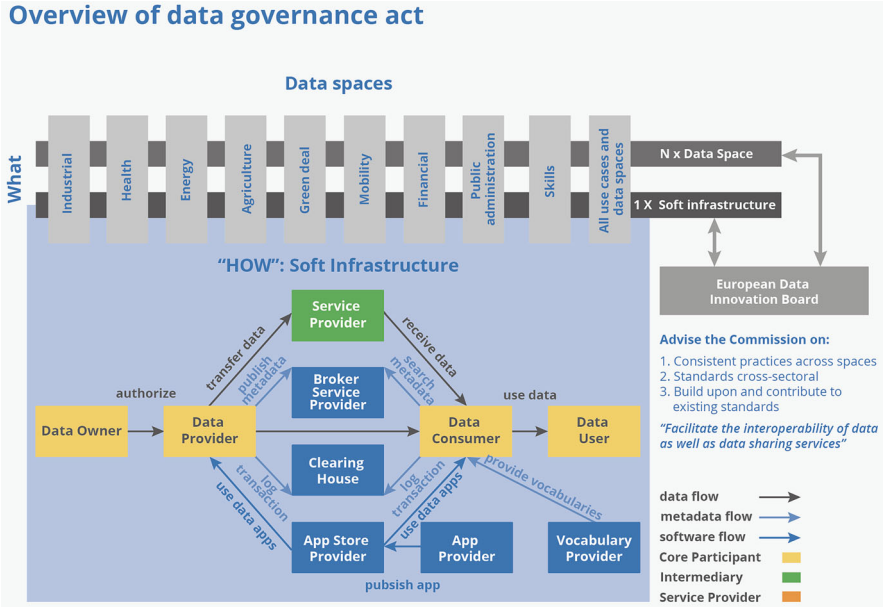


Fig. 2.6 Soft infrastructure for data spaces (©2021, International Data Spaces Association). Source: International Data Spaces Association reference architecture

achieved. This means that it is now a matter of coordination, collaboration, co-creation, agreement, and adoption. The regulatory direction set out by the DGA will certainly help achieve this goal.

The first phase for establishing data spaces is about converging current European initiatives (e.g., International Data Spaces Association [5], Data Sharing Coalition [7], MyData Global [8], BDVA [9], IHAN [10], FIWARE Foundation [2], or Gaia-X [11]) in order to co-create a single result, which will be well accepted for adoption by a critical mass of stakeholders: the first version of the soft infrastructure. This phase will take about 2–3 years. Three aspects will be mission-critical in this endeavor:

- **Create awareness:** Before the first version of the soft infrastructure is published, the concept, rationale, and functional range of the soft infrastructure need to be communicated and promoted on a large scale. Even though the current initiatives will represent the market as good as possible, not all potential stakeholders can be involved in the co-creation process. Therefore, they should have the option to raise their voice during and after the creation process. This is all the more important as after the convergence phase adoption will start immediately, and much more stakeholders than those directly involved today should be familiar with, and support, the agreements and standards.
- **Establish governance structure:** To do so, three steps are necessary: first, the governance structure proposed before must be shaped, and the right people must be appointed as members of the Data Innovation Board (DIB); second, under the

leadership of DIB operational processes must be defined (including communication, decision, and escalation lines); third, a coalition of the willing must gather with their use cases to populate the individual working groups on business/operational/legal and functional/technical matters. What is particularly important here is to include representatives from existing initiatives and ensure fair representation of all member states involved, and all industries affected.

- **Co-create a set of agreements for soft infrastructure:** Co-creation of the soft infrastructure mainly is about establishing coherent functional, operational, and legal agreements as well as agreeing technical standards, which together provide the foundation for interoperability across data spaces. These agreements and standards must be specified in a rule book.
- **Living form of standardization:** The digital soft infrastructure requests a living form of standardization and should be allowed to evolve over time; the common way of dealing with data must continuously respond to market needs and applications.
- **Initial implementation:** The organizations that have created the requirements should roll out and implement the first version of the digital soft infrastructure. This will provide referenceable integrations and, importantly, validate market adoption.
- **Rollout and adoption:** The digital soft infrastructure should then be extended across all sectors over the coming decade.

The soft infrastructure will lead to entirely new opportunities in the European data economy. These include opportunities in the AI field, where access to data is the key to success, and usage for manufacturers along industrial supply chain or in use cases in which the individual controls the data flows. But these are merely examples; this soft infrastructure will create additional security and business opportunities for all organizations and individuals across the EU, opportunities we cannot even dream of.

References

1. 1. Horizon (2020) https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en#:~:text=Horizon%202020%20was%20the%20EU's,the%20archived%20Horizon%202020%20website
2. FIWARE Foundation. https://www.fiware.org/foundation/
3. Plattform Industrie 4.0. https://www.plattform-i40.de
4. Connecting Europe Facility. https://ec.europa.eu/cefdigital
5. International Data Spaces Association. https://www.internationaldataspaces.org/
6. IDSA Reference Architecture Model, Version 3.0. https://internationaldataspaces.org/ids-ram-3-0/
7. Data Sharing Coalition. https://datasharingcoalition.eu
8. MyData Global. https://mydata.org/
9. Big Data Value Association. https://www.bdva.eu/
10. ihan, the European data economy testbed. https://ihan.fi/
11. GAIA-X AISBL. https://www.gaia-x.eu/

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 3
International Data Spaces in a Nutshell



Heinrich Pettenpohl, Markus Spiekermann, and Jan Ruben Both

Abstract International Data Spaces (IDS) enable the sovereign and self-determined exchange of data via a standardized connection across company boundaries. They address the many challenges in the overarching use of data in terms of interoperability, transparency, trust, security, and adaptation by a critical mass. For this purpose, the IDS develops a standardized architecture, which is described and continuously updated in the IDSA Reference Architecture Model (RAM) document. The core of the underlying concept is the linkage of data and usage conditions and their organizational and technical processing and enforcement. Organizational roles and responsibilities are considered on the one hand, and various technical components are defined on the other. The core of the architecture thereby, in contrast to existing solutions and approaches of completely centralized or decentralized architectures, is the paradigm of a federated structure. This enables the actual (raw) data to be exchanged exclusively between the participants without the need for a third party or a central data store. The chapter gives an overview of this approach and provides insights in terms of objectives, roles, and components that are used to enable sovereign data exchange. It enables a basic understanding of the core concept and assessing the impact of such an architecture approach.

3.1 International Data Spaces

Data sovereignty is a fundamental aspect of the International Data Spaces (IDS). It can be defined as a natural person's or corporate entity's capability of being entirely self-determined with regard to its data. This is the main reason why the International Data Spaces initiative got its name in 2015 [1]. A Data Space represents a data sharing concept without a central storage. Thus, data remains at its source and is only shared when needed. This enables data providers to be sovereign over the data as

H. Pettenpohl (✉) · M. Spiekermann · J. R. Both
Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany
e-mail: heinrich.pettenpohl@isst.fraunhofer.de; markus.spiekermann@isst.fraunhofer.de; JanRuben.Both@isst.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_3

long as data is needed, and when it is needed IDS have to ensure to attach usage policies to the data, which systems and users can follow. In this regard the IDS aims at meeting the following goals.

3.1.1 Goals of the International Data Spaces

Trust is always needed to enable data sharing in a data ecosystem. A participant has to trust in the systems itself, but also in the fact that other participants in the data ecosystems get valuable data, which only should be used with regard to the usage policies, defined by the data provider. Thus, trust is the basis of the International Data Spaces. Each participant and software in the IDS will be certified before being granted access to the ecosystem.

Security is strongly coupled with trust. All systems in the IDS have to fulfill state-of-the-art security, to also guarantee trust and data sovereignty. Thus, security requirements are also part of the certification criteria.

Data sovereignty is a fundamental aspect of the International Data Spaces (IDS). It can be defined as a natural person's or corporate entity's capability of being entirely self-determined with regard to its data. This means that a data owner can define usage restriction to their data, before sharing it with data consumers. Data consumers must accept the usage restrictions.

Data ecosystems enable new business models that individual actors cannot make possible themselves because they lack the entirety of data. No single actor has all the data it needs to offer innovative service itself. Therefore, a data ecosystem needs a data space to enable these new innovative services.

Standardized interoperability is needed to build up data space, because different data ecosystems will exchange different kinds of data in different formats and protocols. Only when the interoperability is standardized, every system can be interoperable in the IDS. Therefore, the IDS architecture is defined in a reference architecture model [2], data and endpoints are described semantically in the information model, and certification ensures that every system follows the architecture and uses the information model in the defined way. Also, there is the DIN Spec 27070 which defines the IDS Connector.

3.1.2 Reference Architecture Model

In the following, we provide an overview of the technological components inside the IDS Reference Architecture Model, starting with the most important technological building block of the IDS, the International Data Spaces Connector.

Figure 3.1 shows an overview of the IDS architecture. Alongside the Connector, additional key services are essential for a successful realization of the IDS. The following services are defined by the IDS.

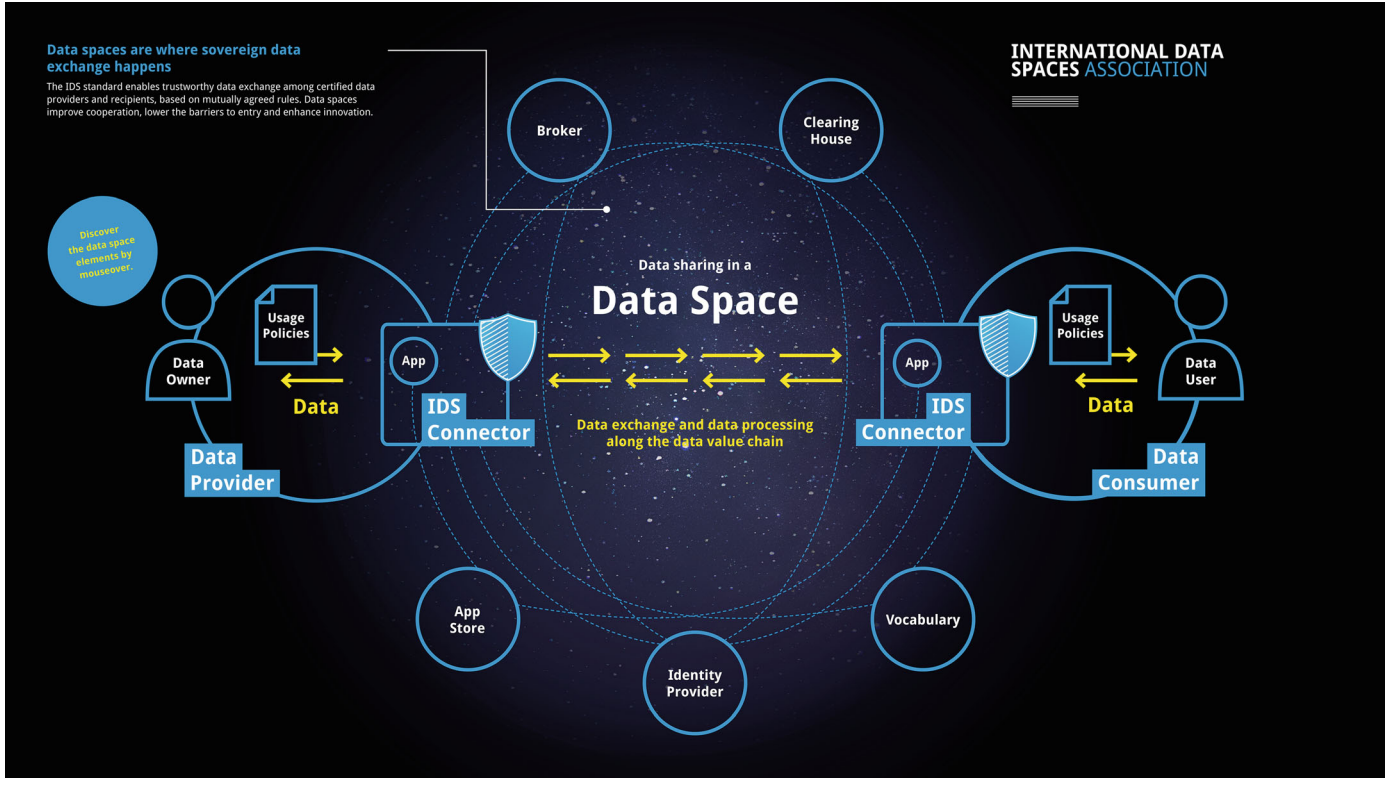


Fig. 3.1 Reference architecture model [3]. ©2021, International Data Spaces Association. Used under permission from International Data Spaces Association

3.1.2.1 The International Data Spaces Components

Connector

As a technological building block of the IDS, the Connector ensures that participants maintain sovereignty over the data. At the same time, it functions as an interface between the internal systems of the IDS participants and the IDS ecosystem itself.

Depending on the configuration, the tamper-proof Connector hosts a variety of system services ensuring, for example, a secure bidirectional communication, enforcement of usage policies upon exchanged content, system monitoring, and logging of content transactions for clearing purposes. The functionality of a generic Connector may further be extended by custom software (Data Apps) for data processing, visualization or persistence, etc.

As shown in Fig. 3.2, the IDS Connector can be viewed in the phases of configuration and execution. The structure of an IDS Connector in the execution phase consists of three container types:

1. The Core Container with basic functions for communication between IDS components (Connectors, Broker, and App Store). This is divided into the internal components Data Router, which manages the communication according to predefined configuration parameters, and Data Bus, which exchanges data with other components or stores data within the connector.
2. App Store Containers for Applications certified and downloaded from the IDS App Store.
3. Custom Containers which were not stored in the IDS App Store but deployed by the Connector operator itself.

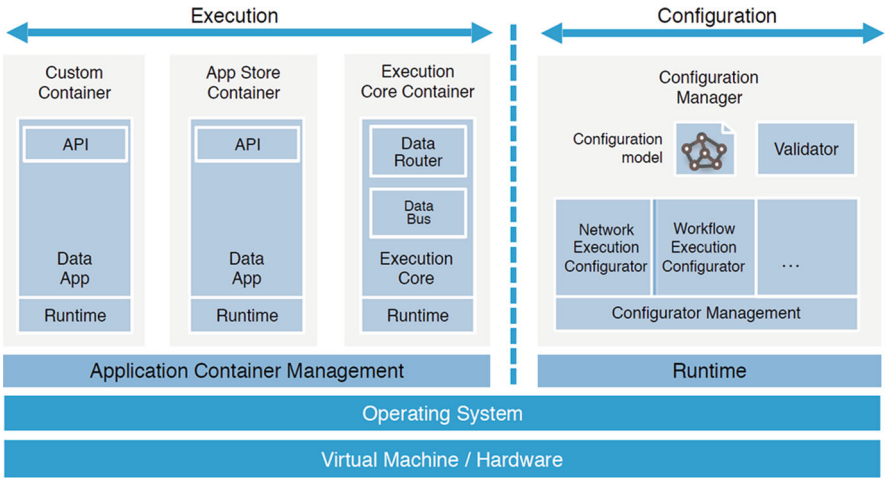


Fig. 3.2 International Data Spaces Connector Architecture [2]. ©2019, International Data Spaces Association. Used under permission from International Data Spaces Association

The application container management technology allows the different services to run in an isolated and secure environment. The Connector architecture follows the principle of processing data as close as possible to the data source.

Identity Provider
The Identity Provider should offer a service to create, maintain, manage, monitor, and validate identity information of and for participants in the IDS. This is imperative for secure operation of the IDS and to avoid unauthorized access to data. The Identity Provider administers self-descriptions and attested (certified) attributes of the connectors and issues tokens as needed for the required attributes of a connector.

- Each International Data Spaces Connector has a private key with a corresponding X509v3 certificate (device certificate).
- In contrast to conventional PKI-based enterprise IDM systems, these static certificates are however used for authentication only and not for the exchange of identity attributes.
- Instead, these are exchanged using dynamic tokens that the connectors obtain from an attribute server.

Metadata Broker
A Broker acts as a mediator between data providers offering data and data users requesting data. It also acts as a data source registry. In more detail, a Broker performs the following activities:

- Provides data providers with functions to publish their data sources
- Provides data users with functions to search through the data sources of data providers
- Provides data providers and data users with functions to make agreements on the provision and use of certain data

Clearing House
The Clearing House provides clearing and settlement services for all financial and data exchange transactions. The Clearing House logs all activities performed in the course of a data exchange. After a data exchange, or parts of it, has been completed, both the data provider and the data consumer confirm the data transfer by logging the details of the transaction at the Clearing House. Based on this logging information, the transaction can then be billed. The logging information can also be used to resolve conflicts (e.g., to clarify whether a data package has been received by the data consumer or not).

The Clearing House supervises the exchange of data (without infringing upon the data sovereignty of the data owners). In more detail, the Clearing House performs the following activities in its function as a clearing house:

- Supervises and records data exchange transactions
- Furnishes reports on the search for data sources and on data exchange transactions
- Supports the rollback of transactions in case of faulty or incomplete data exchange

App Store
The Industrial Data Space promotes the development of a business ecosystem in which participants may develop software (especially data services) and make this software available via the App Store. The App Store Operator performs the following activities:

- Provides functions by which software developers may describe data services and make these services available to other participants
- Provides functions by which participants may retrieve and download data services
- Provides functions for payment and rating of data services

Vocabulary Provider
In order to better define one's own data, domain-specific vocabularies can be created and made available in the vocabulary for all IDS participants. The Vocabulary Provider manages and offers vocabularies (i.e., ontologies, reference data models, or metadata elements) that can be used to annotate and describe datasets. In particular, the Vocabulary Provider provides the Information Model of the IDS, which is the basis for the description of data sources. In addition, other domain-specific vocabularies can be provided:

- Provides a central repository for schema and vocabulary information
- Provides a tool support for collaborative versioning (creating, maintaining, and archiving) of vocabularies and schemas
- Enables a linkage between the data transferred by the Connector and the vocabulary information

3.1.2.2 The International Data Spaces Roles

Participants can take on different roles, which are assigned to different categories depending on the level of interaction and organization, which are described in detail below (summarized in Table 3.1).

Table 3.1 International Data Space role categories

Category 1	Core participants	Data owner, data provider, data consumer, data user, app provider
Category 2	Intermediary participants	Metadata Broker Service Provider, Clearing House, Identity Provider, App Store, Vocabulary Provider
Category 3	Software and services	Software Provider, Service Provider
Category 4	Governance body	International Data Spaces Association, Certification Body and Evaluation Facility

Category 1: Core Participant
The core participants are involved and required every time data is exchanged in the IDS. Roles assigned to this category are data owner, data provider, data consumer, and data user.

Data Owner
The data owner is defined as a legal or natural person who creates and/or exercises control over the data. This control is enabled by defining usage policies and providing access to data. Data ownership includes at least these two major concepts:

- To have the (technical) capability and the responsibility to define the usage contracts incl. payment model and usage policies
- To provide access to the data

Data Provider
The data provider is responsible for providing the data for exchange between a data owner and a data consumer and uses software components that are compliant with the IDS Reference Architecture Model for this purpose. In most cases, but not necessarily, the data provider and data owner are identical. A connection can be established directly between a data provider and a data consumer. To facilitate a data request, the data provider can transmit appropriate metadata to the broker service. Further activities can be the logging of transactions at a clearing house as well as the enrichment or transformation of data by means of data apps.

Data Consumer
The data consumer receives data from a data provider. From a business process modeling perspective, the data consumer is the mirror entity of the data provider; the activities performed by the data consumer are therefore similar to the activities performed by the data provider.

Before connecting to a data provider, the data consumer can search for existing datasets by making an inquiry at a Broker Service Provider.

Data User
Similar to the data owner being the legal entity that has the legal control over its data, the data user is the legal entity that has the legal right to use the data of a data owner as specified by the usage policy. In most cases, the data user is identical with the data consumer. However, there may be scenarios in which these roles are assumed by different participants.

App Provider
App providers develop data apps to be used in the IDS. To be deployable, a data app has to be compliant with the system architecture of the IDS. In addition, data apps can be certified by a Certification Body in order to increase trust in these applications. App providers should describe each data app using metadata (in compliance with a metadata model) with regard to its semantics, functionality, interfaces, etc.).

Category 2: Intermediary
Intermediaries act as trusted entities. Roles assigned to this category are Metadata Broker Service Provider, Clearing House, App Store, Vocabulary Provider, and Identity Provider. Only trusted organizations should assume these roles.

The federated architecture of the IDS provides for the operation of (virtually) centralized components that map individual aspects of service delivery within the data space. Namely, these are the core components described above, with the exception of the Connector, which runs in a decentralized manner.

Each of these components must be integrated, operated, and maintained in a functioning data space. These activities are performed by the service provider in its role as intermediary. It should be mentioned here that there can be one service provider for all components or different service providers for individual components.

Category 3: Software and Services
This category comprises IT companies providing software and/or services (e.g., in a software-as-a-service model) to the participants of the IDS. Roles subsumed under this category are App Provider, Service Provider, and Software Provider.

Software Provider
A Software Provider provides software for implementing the functionality required by the IDS. Unlike data apps, software is not provided by the App Store, but delivered over the Software Providers' usual distribution channels, and used on the basis of individual agreements between the Software Provider and the user (e.g., a data consumer, a data provider, or a Broker Service Provider).

Service Provider
If a participant does not deploy the technical infrastructure required for participation in the IDS itself, it may transfer the data to be made available in the IDS to a Service Provider hosting the required infrastructure for other organizations.

This role includes also providers offering additional data services (e.g., for data analysis, data integration, data cleansing, or semantic enrichment) to improve the quality of the data exchanged in the IDS.

Category 4: Governance Body
The IDS is governed by the Certification Body and the International Data Spaces Association.

International Data Spaces Association
The International Data Spaces Association is a nonprofit organization promoting the continuous development of Data Spaces. It supports and governs the development of the Reference Architecture Model. The International Data Spaces Association is currently organized across several working groups, each one addressing a specific topic (e.g., architecture, use cases and requirements, or certification). Members of the Association are primarily large industrial enterprises, IT companies, SMEs, research institutions, and industry associations.

Certification Body and Evaluation Facility
The Certification Body and the Evaluation Facility are in charge of the certification of the participants and the technical core components in the IDS.

3.1.2.3 Usage Control

In addition to classic access control, which controls access to certain resources, the IDS reference architecture focuses on data-centric usage control [4]. This aims at granting usage restrictions for data even after access. This is achieved by binding rules to the exchanged data, which can be continuously controlled, e.g., how messages are processed, aggregated, or forwarded to further endpoints. On the one hand, the data-centric view allows users to continuously control data flows, not just access. On the other hand, usage control through the IDS connectors ensures that data is not processed in an undesirable way, e.g., by forwarding personal data to public endpoints.

To illustrate the relevance of usage control, examples can be given that cannot be achieved using access control only. In the area of secrecy, it can be achieved that classified data cannot be forwarded by data consumers to third parties who have not been authorized. Separation of duties can be achieved by ensuring that datasets of, e.g., two competing companies are not aggregated or processed by the same third party in a service. This enables a company to control that their own data is not used by a third party to benefit their direct competitors.

Usage control is enforced by monitoring data flows by control points. Within these checkpoints, decision-making engines decide on permission, denial, or necessary modification of the data.

The required restrictions must be formally defined by the data owners. User-friendly graphical interfaces are available for this purpose, which transform the specifications into machine-readable output.

The usage restrictions can be attached to the data in two different ways. On the one hand, usage restrictions can be attached directly to the data. For example, decryption can only take place if the usage restrictions are guaranteed to be respected. On the other hand, usage restrictions can be stored in a central instance independently of the data. In this case, the usage restrictions must be exchanged between the systems.

A policy editor (Policy Administration Point, PAP) integrated in the connectors can be used to specify the usage restrictions.

3.1.3 Certification

Certification is an important element in the IDS to establish a trustworthy data space. A distinction is made between the certification of components and Data Space participants or organizations [5].

3.1.3.1 Security Profiles

The IDS RAM defines four different security profiles: Base Free, Base, Trust, and Trust+ (Managed Trust). New profiles may be added in the future. The Base Free profile supports the operation of IDS concepts and technologies outside the public trusted Data Space, e.g., for research projects for the operation inside of one security domain, e.g., inside a company. The Base profile defines the minimal level of trust mechanisms, including the certification process. The Trust profile defines extended security features. The Trust+ or Managed Trust profile relies on Trusted Hardware based on TPM (Table 3.2).

3.1.3.2 Participant Certification

The certification of participants enables trustworthy interaction with these participants. On the one hand, entire data spaces can become trustworthy, and on the other hand, it is possible to filter data spaces in which everyone can participate according to trustworthiness.

The trust to be established in the participants of the data space is often essential, especially in an industrial context.

The certification itself focuses on the achievement of defined security levels, which include the infrastructure and the compliance with processes.

3.1.3.3 Component Certification

The certification of the components focuses on compliance with the required functionalities and security levels that correspond to the International Data Spaces. In

Table 3.2 IDS security profiles and the related dimensions [2]

	Base free	Base	Trust	Trust+
Development	Open source	IDS community	IDS community	Bound to strong SLAs
Roles	Own infrastructure	All IDS roles supported, billing and clearing optional	All IDS roles supported	All IDS roles supported
Communication abilities	Only private IDS with self-signed certificates	Full interoperable, reduced trust	Full interoperable, free decision of communication	Full interoperable, free decision of communication, hardware anchor
Higher security classes	Standard security level required	Standard security level required	High security level	Higher security level

©2019, International Data Spaces Association. Used under permission from International Data Spaces Association

addition, interoperability is ensured. A particular challenge is to ensure that correct information is given about which participants can access the data and which software components will be able to access the data.

In summary, certification must be ensured by an approved inspection body and the central certification body of the IDS, which act as developer-independent entities. These instances therefore provide an official certificate, i.e., a signature of the individually relevant manifests that enable reliable technical verification, as well as a digital X.509 certificate for the connector.

3.1.4 Open Source

The IDS ecosystem's implementations are by their very nature collaborative, as the results must be verified by stakeholders from different industries before being success stories. This open environment will not only speed up the validation process, but it will also help IDS component implementations achieve the highest level of quality possible through collaboration. As a result, technical components of IDS (Sect. 3.1.2) can be found in the IDSA GitHub repository (https://github.com/International-Data-Spaces-Association), where they are built using best-practice OSS development processes through continuous implementations and feedback from communities and stakeholders.

References

1. Otto, B., & Jarke, M. (2019). Designing a multi-sided data platform: findings from the International Data Spaces case. *Electronic Markets, 29*(4), 561–580.
2. Otto, B., Steinbuß, S., Teuscher, A., & Lohmann, S. (2019). *Reference architecture model* [online]. Version 3.0 | Apr 2019 [viewed 30 Apr 2021]. Available from: https://internationaldataspaces.org/download/16630/
3. IDSA Head Office. (2021). *IDSA infographic* [online] [viewed 18 March 2022]. Available from: https://internationaldataspaces.org/download/20861/
4. Eitel, A., Jung, C., Brandstädter, R., Bader, S., Kühnle, C., Birnstill, P., Brost, G., Gall, M., Bruckner, F., Weißenberg, N., & Korth, B. (2021). *Usage control in the international data space* [online]. Position paper. Version 3.0 [viewed 30 Apr 2021]. Available from: https://internationaldataspaces.org/download/21053/
5. Menz, N., Resetko, A., & Winkel, J. (2019). *IDS certification explained* [online] [viewed 3 May 2021]. Available from: https://internationaldataspaces.org/download/16456/

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 4
Role of Gaia-X in the European Data Space Ecosystem



Hubert Tardieu

Abstract The Gaia-X project was initiated in 2019 by the German and French Ministers of Economy to ensure that companies would not lose control of their industrial data when it is hosted by non-EU cloud service providers.

Since then, Gaia-X holds an international association presence in Belgium with more than 334 members, representing both users and providers across 20 countries and 16 national hubs and 5 candidate countries.

The Association aims *to increase the adoption of cloud services and accelerate data exchanges* by European businesses through the facilitation of business data sovereignty with jointly approved (user and provider) policy rules on data portability and interoperability.

Although for many enterprises, data sovereignty is seen as a prerequisite for using the cloud, a significant driver to boost the digital economy in business is incentivizing business data sharing. Two decades of cost optimization have constrained business value creation, driving many companies to neglect the opportunity to create shared value within a wider industry ecosystem.

Now, thanks to the participation of large numbers of cloud users in the domains of Finance, Health, Energy, Automotive, Travel Aeronautics, Manufacturing, Agriculture, and Mobility, among others, Gaia-X is ideally positioned to help industries define appropriate data spaces and identify/develop compelling use cases, which can then be jointly deployed to a compliant-by-design platform architecture under the Gaia-X specifications, trust, and labeling frameworks.

The creation of national Gaia-X hubs that act as independent think tanks, ambassadors, or influencers of the Association further facilitates the emergence of new data spaces and use/enabler cases at a country level, before these are subsequently extended to a European scope and beyond. Gaia-X partners share the view that data spaces will play a similar role in digital business as the web played 40 years ago to help the Internet take off.

H. Tardieu (✉)
Gaia-X European Association for Data and Cloud AISBL, Brussels, Belgium
e-mail: hubert.tardieu@board.gaia-x.eu

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_4

The Gaia-X Working Groups are at the core of the Gaia-X discussions and deliverables. There are three committees: the Technical, the Policies and Rules, and the Data Spaces and Business.

The Technical Committee focus on key architectural elements and their evolution, such as and not limited to:

- **Identity and Access Management:** bridge the traditional X509 realm and new SSI realm, creating a decentralized network of identity federations
- **Service Composition:** how to assemble services in order to create new services with higher added value
- **Self-Description**: how to build digital trust at scale with measurable and comparable criteria

The Policy and Rules Committee creates the deliverables required to develop the Gaia-X framework (compliance requirements, labels and qualification processes, credentials matrix, contractual agreements, etc.):

- The Labels and Qualification working group defines the E2E process for labels and qualification, from defining and evolving the levels of label, the process for defining new labels, and identifying and certifying existing CABS.
- The Credentials and Trust Anchors working group will develop and maintain a matrix of credentials and their verification methods to enable the implementation of compliance through automation, contractual clauses, certifications, or other methods.
- The Compliance working group collects compliance requirements from all sources to build a unique compliance requirements pool.

The Data Spaces Business Committee helps the Association expanding and accelerating the creation of new Gaia-X service in the market:

- The Finance working group focuses on business modeling and supports the project office of the Association.
- The Technical working group analyzes the technical requirements from a business perspective.
- The Operational Requirements working group is the business requirements unit.
- The Hub working groups hold close contact with all Gaia-X Hubs and support the collection and creation of the Gaia-X use and business cases. These working groups maintain the international list of all use cases and data spaces and coordinate the Hubs.

4.1 A Quick Introduction to Gaia-X

The Gaia-X Association was created by its 22 founding members [1] on September 15, 2020. It was officially authorized by royal Belgium decree on December 21, 2021. Gaia-X is open to all cloud service providers and cloud users which

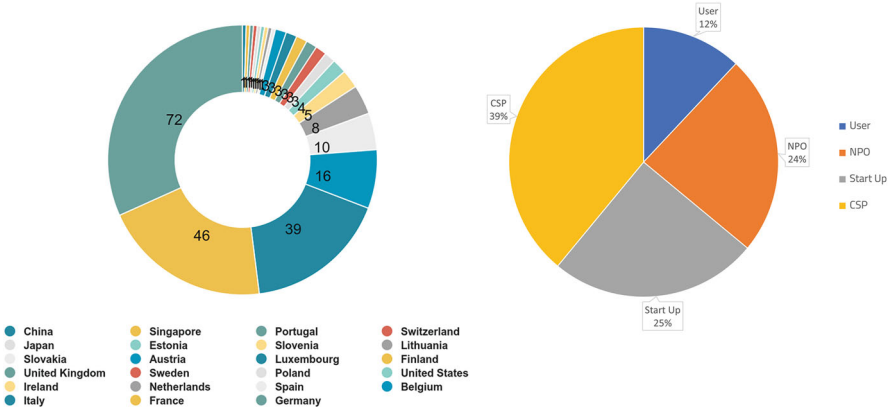


Fig. 4.1 Gaia-X Association members distribution (29 March 2021)

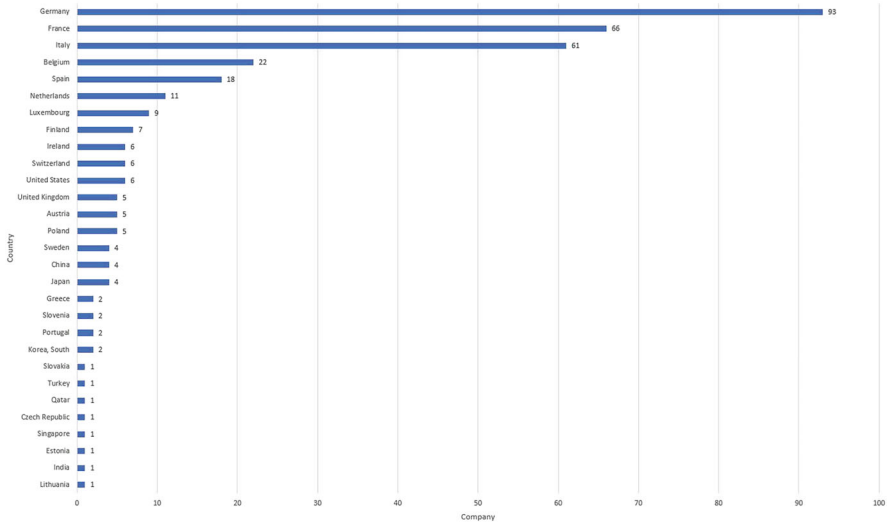


Fig. 4.2 Gaia-X Member distribution by country (end of April 2022)

endorse its Article of Associations and the objectives of the Association as defined in the original Franco-German Position Paper [2]. At the Gaia-X Board on March 29, 2021, the Association added a further 212 members on top of its initial founding members, bringing together all the major cloud service providers (CSPs) and a significant number of user companies in the selected data spaces (see Fig. 4.1).

At this time, the Association has 343 members, the distribution of which may be seen in Fig. 4.2.

The inclusion of the main US and Chinese CSPs has been agreed by the 22 founding members since the top 3 US CSPs already provide 70% of cloud

services to European cloud users. However, cloud usage by European businesses (26% of processing in 2020) will only tend to increase if CSPs commit to respect policy rules on portability and interoperability of infrastructure, data, and applications. Such policy rules will be defined in joint Gaia-X working groups that are made up of users, European CSPs, international CSP, academia, and start-ups. Jointly designed policy rules will be published by the Policy Rule Committee that has been nominated by the Board of the Association.

Gaia-X expects cloud usage in Europe to double over the next 4–5 years, a growth which will increase market opportunities both for European CSPs and International CSPs whose offerings respect policy rules.

The openness, non-compete, and non-discrimination guidelines of Gaia-X will facilitate ex-ante compliance with relevant policy rules because they will be designed jointly with all cloud service providers. However, the 22 founding members have decided that the Board will be restricted to members of companies that have their worldwide headquarters in Europe.

The Association will define and prototype federated services for topics, such as identity management, data sharing, and compliance checking. These will be made available as open source to all members. It will also facilitate the creation of data spaces in markets that have been prioritized by the European Commission in its digital strategy [3]—the data spaces will be restricted to deployment cases, where ecosystem partners include the major European players in the respective markets. Today, the Association is active in the following markets: aerospace, agriculture, circular economy, education, energy, finance, geoninformation, health, manufacturing/industry 4.0, media, mobility, public sector, smart cities, smart living, and space. Further data spaces will be facilitated through national Gaia-X hubs (see Chap. 5).

4.2 The Business World with Gaia-X

The objectives of Gaia-X are ambitious but are already reflected by the number of partners (343 by the end of April 22) and the impressive participation of the first two Gaia-X Summits in November 2020 and November 2021, bringing in more than 4500 participants each. Gaia-X is conscious of the challenges it faces and has decided (at the start of 2021) to create a Gaia-X Institute, a body which will ensure that the conceptual foundations of the Association are rooted in proper academic research and thinking in the areas of the economy of data, legal principles of compliance, and objective measurement of success.

4.2.1 Economy of Data

The economy of data has been studied for 20 years in several European universities including Toulouse, Mannheim, and Bologna.

The most significant result has been the definition of multisided markets by Jean Tirole, who was awarded the Nobel Prize of Economy in 2014 for his work on this topic.

The extension of multisided markets concepts to the business world has been widely developed at the Toulouse School of Economy by Jacques Cremer who wrote the foreword of a book summarizing these advances [4].

The value that can be created by sharing data can be generated in two ways:

1. Several partners agree to share similar types of data to increase the *volume of available data.* This requires a common definition of data to ensure that any partner can readily use data created by the others. Increasing the *volume of data sharing* is important in areas such as the effective training of AI algorithms. A good example of volume data sharing is given by "Here," a former Nokia company bought by Audi, BMW, and Daimler, which has decided to share all data related to road conditions that is collected by each of their connected cars. Their objective is to get a comprehensive set of road images, whatever the day of the year or the hour of the day. Such data is essential to feed AI algorithms for vehicle autonomy at level 5 (fully autonomous driving on normal roads). Interestingly, "Here" has recently welcomed the Intel subsidiary Mobileeye to its ecosystem—this new partner brings specialisms in LiDAR sensors capturing images on the road.
2. Several partners decide to share different, but complementary types of data in order to create services that none of them can offer with their own data alone. *Complementary data sharing* brings tremendous possibilities for added value in use cases as diverse as preventive maintenance and fire detection. One example is Skywise, an ecosystem launched by Airbus, which uses an industrial data platform to combine production data collected on Airbus aircrafts with operational and flight data collected by airline companies like EasyJet or United Airlines. The benefit for Airbus is to gain access to operational data so it can improve the design of its future aircraft. For EasyJet, the value is to have access to detailed data collected at the time of aircraft production in order to better anticipate potential future failure. Airbus is claiming 30% productivity improvement thanks to such industrial data sharing. Another example is the use of satellite data generated by the Copernicus system—by combining satellite broad range image data with data collected by surveillance stations on the ground, local authorities can detect forest fires earlier.

While the principle of business data sharing is intuitively understood, the mechanism of value creation is less well defined, although its principles were described by Jean Tirole. Initially formulated to reflect the market of payment cards, the principles of value creation describe a two-sided market with, on one side, cardholders and, on the other, merchants. The cardholders expect many shops to accept their cards, but before the merchants will invest in payment terminals, they want to be sure that a sufficient number of customers are willing to use them. A two-sided market needs to reach a critical mass of adoption to be sustainable. On its own, each side has little

chance of reaching a viable scale, but if an operator takes the risk in committing to each side that they will reach critical size, then the market can take off.

In the extreme, the operator will become unbeatable, requiring authorities to intervene to avoid a monopoly situation. While Tirole's work focused on the consumer market, it has recently been extended to business markets, especially in the context of reflecting the incentives for complementary data sharing—with the most attractive opportunities residing within ecosystems of partners belonging to the same value chain.

While competition in conventional business models has forced companies to concentrate on specific parts of a value chain (e.g., production of aircrafts for Airbus, operation of aircrafts, and customer relationships for EasyJet), complementary data sharing provides an opportunity for partners within a given value chain to reap the incremental benefits of sharing business data, while still maintaining their business autonomy.

4.2.2 Compliance

What are the incentives which can be proposed and which arguments can be used to facilitate the creation of data spaces?

The first approach is to impose *new regulations*, as has been done in the finance sector with the 2019 Payment Services Directive (PSD2). This requires that any financial institution, which is providing personal bank accounts has the obligation, if the owner of the account agrees, to provide API-enabled access to certain account data to stimulate the creation of additional value services. PSD2 has clearly been a catalyst for the explosion of Fintech activity over recent years. The Finance sector is ahead of other industries in respect of data sharing and is currently the only one to have a directly related regulation. However, the approach is limited because APIs are sector specific and not suitable for cross-domain data sharing. A promising approach is the one taken under the Connecting Europe Facility program run by the European Commission where FIWARE open API is recommended for cross-domain and cross-border data sharing.

The second approach is to agree on *a common data model*, as it has been done in the management of physical building infrastructures (plan, design, construct, operate, and maintain). The Building Information Modeling (BIM) concept has existed since the 1970s, but it has become generalized across the construction industry since the early 2000s. ISO defines BIM as "Use of shared digital representation of a built asset to facilitate design, construction and operation processes to form a reliable basis for decision." For the professionals involved in a project, BIM enables a virtual information model to be shared by the design team, the main contractor, subcontractors, and the owner/operator.

Unfortunately, the timeframe described in the two previous sections is not compatible with a rapid deployment of data spaces. How can quicker progress be made?

The first option is to get rid of *obsolete and overly constraining regulations,* which were originally created for the pre-digital world. Typically, in payment markets, merchants heavily subsidized the acquisition of customers by selling their services below cost and could, according to existing regulations, be accused of dumping. This is similar to the story of "free" services which have allowed GAFAM to take off in their respective markets. An attractive proposition is to build *regulatory sandboxes*, sponsored by governmental authority, that allow a controlled relaxation of regulations which would otherwise block complementary data sharing, thereby allowing alternative models to be tested.

An example of such a regulatory sandbox is the one proposed by the French Commission de Regulation de l'Energie: it concerns the creation of an energy data space to make the best possible use of data collected by the 30 million smart meters in France—of which only 5% is currently usable because of regulatory issues. After a period of a few years, services to be offered will have been clarified and the business cases proven. The results will help industry and regulators to decide how to organize the energy data space for the longer term.

Among the main risks and concerns that prevent companies from sharing business data, the most often cited relate to "security" and "data usage control." *Data usage control* defines the policies under which data will be shared in a transparent manner. In this way, data is associated with well-defined services that are agreed between the partners—it cannot be considered as open data, even between the partners. Through the use of industry data platforms, it is technically possible to control and check data usage according to agreed policies by linking data and services and enforcing that these compliance checks with these policies are performed by the platform operator.

Security is a risk that is not specific to data sharing and can be addressed through proper identity and access management and appropriate use of data encryption.

The primary technical challenge is typically the lack of *interoperability*. Today a large share of data is stored in public clouds, 70% of which are operated by US providers. While cloud customers gain flexibility in terms of their own access to compute and storage resources, they can face challenges when sharing data with ecosystem partners that use different cloud service providers. Interoperability challenges exist at an infrastructure level because of incompatible communication protocols and also at an application level because of a lack of common API's (a challenge that PSD2 has helped to address in financial services). One way forward is for cloud service providers to demonstrably comply with agreed common policies for infrastructure and applications.

While the current focus for digital and business data sharing is policy and regulatory compliance, it is important to remember that compliance was initially created to combat money laundering. It then extended into corporate social responsibility and further into personal data protection.

Corporate compliance is not always well accepted, largely because enterprises consider it to be an additional process and control mechanism linked to objectives, which are beyond the immediate control of their own company. In addition, unlike

traditional regulations which trigger ex-post penalties if not respected, compliance brings ex-ante constraints which have to be checked by the company itself.

Corporate compliance was initially instigated in the USA but must be seen as mandatory in a world of globalized industry. Two attitudes are possible in Europe: either *defensive* to complain about additional bureaucracy and costs, or *offensive* to include appropriate objectives for Europe, which will become a self-validated prerequisite for all companies operating in Europe.

We believe that business data sharing and associated data sovereignty are objectives which can be translated to specific policies for interoperability, security, and data usage control.

In our view, compliance is the only way for Europe to make business data sharing possible in a timely and transparent manner that guarantees data sovereignty for participating companies.

Further details can be found in [5].

4.2.3 Measuring Success

Creating data spaces will be a collective task involving several players who recognize that data sharing with other partners will be of collective benefit for their companies when endorsing the compliant-by-design portability and interoperability architecture proposed by Gaia-X. Data space success will be achieved through investment from initial partners supported by public funding. In contrast to traditional investments in software which can be protected by Intellectual Property Rights (IPR), data spaces creation will be funded by initial participants to the eventual benefit of all players in the resulting ecosystem. In the past, standardization tasks have been funded by public money with the active participation of industry users and providers, typically resulting in long lead times for design, implementation, and adoption of new standards. The stated intent of Gaia-X to facilitating the doubling of cloud usage in Europe within the next 4–5 years doesn't allow for such a lengthy process.

Moreover, governments and the European Commission, which have announced that they will financially support the creation of data spaces (EC has announced a contribution of 1.2 B€ for the creation of nine data spaces) want to be able to justify to their citizens that they have properly invested public money.

Tokenomics, which is a new branch of economics, might be an appropriate tool for co-operatives to reflect capital investment and gain voting rights. The design of rules and policies for achieving the desired goals of data spaces creation could be based on the game theory using tokens for incentive alignment and rules enforcement. The related mechanism is often referred to as *tokenomics.*

"Generally speaking a token is a thing which serves as a visible, tangible or intangible representation of a fact or a right; for example, a driving license card is a token which represents the fact that you are trained and allowed to drive a car.

A cryptographic token is a cryptographically secure, provable representation of a fact or right, which can additionally be processed in digital systems like decentralized networks. Tokens are digitized multipurpose instruments, ranging from simple-single to multi-complex designs. It could be value, stake, voting right, or anything. A token is not limited to one specific role or utility, it can fulfill a lot of roles in its ecosystem.

Tokenomics encompasses the concept of economic system design and implementation to incentivize specific behaviors in a community, using tokens to create a self-sustaining ad hoc economy. It includes game theory, mechanism design, and monetary economics" [6].

An example of Tokenomics was proposed during the Gaia-X summit where the Bosch Team used the case of the design and implementation of a collective transportation system to be installed in a given territory; a funding mechanism can be created by selling tickets which are tokens of no value before the transportation system operates but which will progressively gain value. Cities or regions can participate through the acquisition of tokens which can be later transformed into real tickets for delivering social policy (e.g., for elderly, unemployed, and students).

For the Association, Tokenomics is being considered for two purposes: (1) increase its financial resources in order to accelerate the creation of data spaces. In addition to regular fees paid as a member, companies belonging to an ecosystem can put additional funding represented by a token. This additional funding will be used to support the efforts of active companies that contribute their expertise in the creation of data spaces; companies that want future benefits from data spaces can buy tokens, which will support the costs of the active companies. (2) Public institutions can support the creation of data spaces by acquiring tokens which give them voting rights in the definition of the data spaces as well as opportunities to facilitate the participation of start-ups in the future ecosystem.

Bringing together the economics of data, compliance and measuring success through Tokenomics will pave the way for the new world of Gaia-X.

4.3 The Gaia-X Principles

Gaia-X is an initiative to provide a regulatory and technical framework that supports the generation of data and infrastructure ecosystems. It will provide a set of "Federation Services," which allow interactions between all participants without any specific company taking on a dominant role in controlling the flow of information (see Fig. 4.3).

A common set of policy rules and functional and technical specifications will provide interoperability and portability of data and applications, avoiding the lock-in effect that many users experience with existing providers. This will also provide economies of scale for the many small- to medium-sized cloud service providers and specialized data centers. The distributed and federated infrastructure further provides the basis to integrate EDGE and HPC/QLM "as a Service" capabilities.

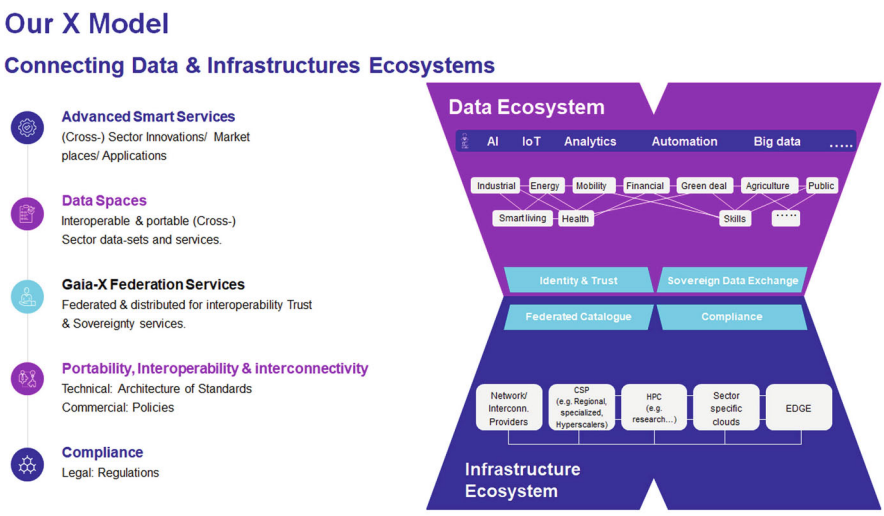


Fig. 4.3 The Gaia-X architecture (©2020 German Federal Ministry for Economic Affairs and Energy)

The infrastructure is the basis for the generation of data spaces (as they are also described in the European Data Strategy), which will form the basis for "advanced smart services." These will be built around "collaborative use cases" where multiple parties agree to share data with a view to improve supply-chain process efficiency (smart connected supply chains generate about 20% savings), as well as bringing new value in competitive scenarios (e.g., mobility data spaces where each provider publishes their itineraries and real-time traffic information and costs, allowing the creation of new mobility services).

One critical enabler for these services is that each data provider maintains sovereignty of their data usage. Gaia-X uses the "International Data Spaces" Reference Architecture to ensure that data usage controls are provided, and compliance is assured.

4.3.1 Objectives

Objectives of Gaia-X have been described in the Franco-German Position on Gaia-X [2] published on February 18, 2020; they are a prerequisite for new members of the Association. This document has been agreed between the 22 founding members regrouping Cloud Users and Cloud Service Providers; it has been endorsed by the 27 European Countries and the European Commission in its document "Towards a next generation cloud for Europe":

> The European cloud federation initiative will aim at creating synergies between national and cross border initiatives, to enhance and broaden their scale and coverage. The Gaia-X

initiative for a European federated data infrastructure is a leading example of a public-private initiative aiming at a broad European scope.

Gaia-X is a strong private partnership to implement the European Strategy for Data announced in February 19, 2020, in its entire dimensions of data space creation and next generation of cloud compliant with portability and interoperability requirements.

4.3.2 Policy Rules and Specifications for Infrastructure Application and Data

Policy rules will be jointly created by cloud users and cloud service providers both European and international. The aforementioned Franco-German paper which has been agreed by all Gaia-X partners described them in the following terms:

For the area of infrastructure:

- ***Reversibility***: Changing the cloud provider with portability for data and services, in the frame of Art. 6 "Porting of Data" of the European Free Flow of Non-personal Data Regulation. First codes of conduct have been handed over to the European Commission.
- ***European CSP certification scheme***: The scheme was developed by a European Working Group, including German BSI (C5) and French ANSSI (SecNumCloud), and handed over to ENISA in June 2019. Based on the European Cybersecurity Act or within the framework of the New Legislative Framework, certification schemes for an efficient onboarding of cloud infrastructure services providers into Gaia-X should be developed. To this end, Gaia-X has to develop the appropriate self-description and discovery schemes.
- ***Security of data***: Security policy is associated with the data about its usage and shall be controlled irrespective of the providers. Beyond existing cybersecurity approaches, the need and requirements for a trusted execution within the edge environment should be considered.
- ***Identity and access management (IAM)***: Resources and devices shall be identified in a way which is common regardless of the provider. All agents and devices, like all assets, are identified regardless of which provider is involved in a Gaia-X service instantiation. This is covered by interoperability, common trust requirements (international technical standards and harmonized legal framework), and an approach covering the Gaia-X IAM requirements. Both Gaia-X core components and provider offer a sufficiently high degree of security regarding the integrity, confidentiality, traceability, and availability of Gaia-X identities. The solution will be selected based on industry best practices and accepted international standards. It forms the Gaia-X baseline for flexibility regarding different technical and national/legal requirements. On top, distributed access and (decentralized) identity management schemes, including verifiable credentials, should be considered enabling a robust ecosystem.

- ***Energy efficiency***: Transparency of energy consumption and its comparability regarding equivalent workloads should be encouraged. Users should have better visibility of the energy consumed by processing their data, including in the context of self-description. Criteria may be applied to all type of cloud facilities, including edge computing facilities.
- ***Protection against non-European extra-territorial regulations:*** Protection against abuse of national regulations that allow to access data stored in cloud infrastructures or services is an essential part of the European federated data infrastructure.

For the portability and interoperability of applications among cloud providers:

- ***Avoiding "lock in"*** by agreeing on open API describing the use of technical facilities provided between SaaS offerings and third parties, including individual bricks internally (e.g., changing the data storage service or other individual services).
- ***Common data standards*** enabling data sharing through file exchange or functional API (as in PSD2 for finance).
- ***Common definition of data security policy*** defining data security policy in logical and legal terms.
- ***Encryption*** to be used for stored data where relevant, with a portability of the keys used to encrypt and interoperability of key management systems.
- ***Virtualization of distributed data*** across multiple service providers (e.g., data cashing, data prefetching) to enable distributed cooperating application and services.
- ***Edge computing*** as a possible processing paradigm to create possibilities for real-time processing and distributed algorithms, cloud-native apps vs edge apps.
- ***Data portability and interoperability*:** Data interoperability has to go further than data portability, i.e., domain-specific data semantic harmonization is the next degree of data interoperability and should include inter alia through (domain-specific) standardized management dishes for digital twins.
- ***Service and contractual interoperability*** to enable on-demand CSP collaboration.

4.3.3 Federated Services in Business Ecosystems

Theoretically, in closed few-to-few ecosystems, a consortium of partners could just implement federation services on their own (see Fig. 4.4). When they need to open up to a more dynamic ecosystem or want to achieve interoperability and data sovereignty beyond their own ecosystem borders, they will need to rely on federation services owned/provided by the Gaia-X Association.

Federated services provide basic digital services that are required in closed few-to-few communities to create an ecosystem (data or infrastructure) without any party

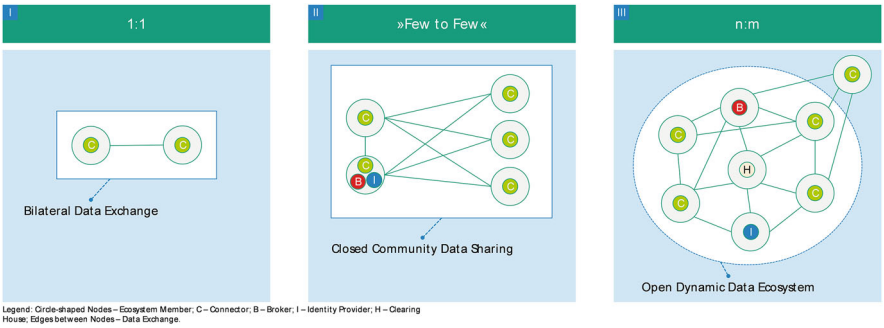


Fig. 4.4 Evolvement stages of business ecosystems (©2020 Fraunhofer ISST)

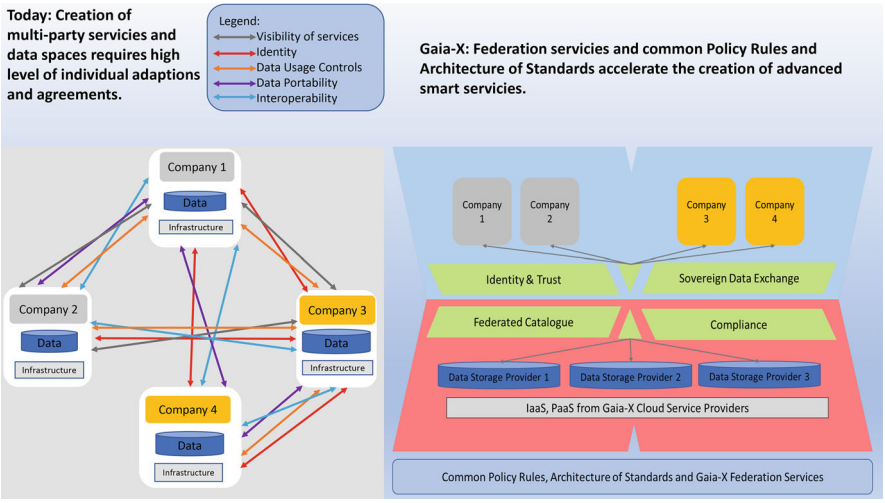


Fig. 4.5 Gaia-X Federation Services

being in a position to take advantage of having access to the key controls. This is very similar to the setup in the physical world where governments control commercial registers and trade rules which in turn enable (for example) a group of independent auditors or payment providers to provide such services. A key element in this setup is that the provider of such services is not part of the individual business relationships, working instead on commonly defined standards.

Federation services in Gaia-X (see Fig. 4.5):

- **"Identity and Trust"** provides a common mechanism to identify individual companies and data providers together with associated trust mechanisms. In a way, this is like what Amazon/Google/Microsoft do with their authenticators—only that the identity and trust provider can and must not use these services for any external commercial purpose.

- **"Federated Catalogue"** this is the listing of all available services and data endpoints. It provides the necessary attributes to search for services that are compliant with specific regulations, and provides an independent assertion through the Compliance Services.
- **"Data Exchange services"** ensures compliance to the data usage policies.

Federation Services are governed by the Gaia-X—a membership-based organization where no single company or organization can define the rules, but where, building on policy rules and architectures of standards, the members jointly establish the definition and implementation of Federation Services. (This model is similar to one of the cornerstones of the WWW. "Domain Name Services" provide a federated registry of domain names; it is governed by the multi-stakeholder group ICANN.) Availability of Federation Services for Identity, Trust, Catalogues, and Compliance greatly simplifies the setup of multiparty ecosystems as it is not necessary to provide individual solutions and agreements, since the Federation Services already provide a base for legal and technological compliance:

Federation Services enable efficient creation of ecosystems:

- **Commercially:** No single party has control over the ecosystem.
- **Legally and technically:** They provide a basis for legal compliance and provide interoperability across different Identity, Trust, Catalogue, and Data Usage Controls.

4.4 The Gaia-X Data Spaces

Currently, Gaia-X has committed to cover six data spaces, which will be gradually augmented when Gaia-X has onboarded representatives of additional domains in the fields of agriculture, education, and green. Gaia-X national hubs will help to identify these new partners, initially on a national geography basis, but eventually to be extended to a European dimension.

4.4.1 Finance and Insurance

The Finance and Insurance Data Space is driven by Caisse des Depots et Consignation in France. It already includes Arkea, BNPP, BPCE, CDC, Credit Agricole, Commerzbank, Intesa San Paolo, and MAIF.

Four main benefits are expected by Gaia-X partners:

- Easily build, assemble, and use trusted and value-creating data-based cloud services

- Gaia-X as a European-wide accelerator of innovation and co-construction to provide secured data sharing and artificial intelligence services at scale and in a compliant and secure way
- Create new financial product/services and foster new business models that are "compliant by design" with European regulations and values
- Foster Europe's competitiveness and financial market' stability

The next steps are to develop a "compliant by design" framework aligning Gaia-X policy rules and financial sector regulations and deliver a first data space demonstrator based upon the "Financial Big Data Cluster" initiative launched in Germany.

4.4.2 Energy

The energy domain is driven by EDF and includes also global companies as Engie and Enel, as well as distributors like Enedis.

The key focus of the domain will be smart grids, with its four dimensions of customer/services, industrial data, environment data, and financial data. Electricity data is not yet sufficiently shared, and this prevents the creation of new business models based upon multiparty cooperation between technology providers and data owners. The energy partners want to enable interoperability with proper data governance and security to facilitate data analytics and edge computing.

4.4.3 Automotive

In December 2020, German automotive manufacturers have created and announced the German Auto Data Alliance with BMW, SAP, Siemens, Robert Bosch, and ZF Friedrichshafen with the support of Deutsche Telekom. In its press release announcing the Automotive Alliance, BMW stated:

> How much CO_2 is produced in the production of an SUV? And in which models are defective parts installed? These are pressing questions for automakers. The BMW Group is looking for ways to make its supply chains faster, more transparent and safer.

Bosch and BMW explained during the summit that because of a lack of interoperability between the systems operated by OEM and Part Suppliers, there is a 6-month delay between failure occurrence and failure notification to the part supplier. In addition, 80% of parts investigation at the supplier are redundant as failure and root cause are already understood.

The German Automotive Alliance is currently extending its reach to French and Italian car manufacturers.

4.4.4 Health

Health data space is currently driven by Philips with the participation of Healthineers (formerly Siemens Medical), Sanofi, APHP (Parisian Hospitals), and Dassault Systems which has recently bought Medidata (specialized in clinical test management).

Philips want to establish a Longitudinal, 360-degree health profile of citizens enabled by ecosystems at the national and EU level. Establishing secure, open, federated health data spaces and cloud services is a key enabler for the transformation of healthcare to continuous health tracking, image-guided therapy, computational pathology, and genomics as the basis of personalized medicine.

As for the other regulated business of Health Philips wants to develop a "compliant by design" framework aligning Gaia-X Policy rules and healthcare sector regulations.

4.4.5 Aeronautics

Three years ago, the Aeronautics sector launched the Skywise system regrouping Airbus with 130 Airlines companies to reduce aircraft maintenance cost by 30% (See Supra).

During the Summit, Gaia-X invited the other industry participants: Dassault Aviation, Safran, Thalès, Ariane Group, MDBA Systems, Leonardo but also EASA, Eurocontrol as well as GIFAS, BDLi, and Academics (TÜM, ISAE, etc.) to cooperate for the creation of an Aerospace Data Space.

The initial common objective is to leverage common specific use cases (e.g., export control management, high-performance computing for simulation and modeling, pseudo-real time data collaboration for air traffic optimization, etc.) that will support the design of the aerospace policy rules, data ontology, architecture of standards, and service federation that will ensure the European aerospace ecosystem.

4.4.6 Travel

The travel data space is driven by Amadeus and includes Air France KLM and Aéroport de Paris. The group wishes to extend to other French companies like SNCF and other European travel companies.

The objective is to offer to the traveler in Europe a multi-modal ecosystem encompassing all aspects of travel. Two use cases have been identified: **seamless travel** allowing the sharing of digital identity, passenger name record, travel record, and social media interactions, and **health pass** to provide safe corridors and avoid repeated health checks. The compliance with GDPR and Health Authorities regulation is a prerequisite for these two cases.

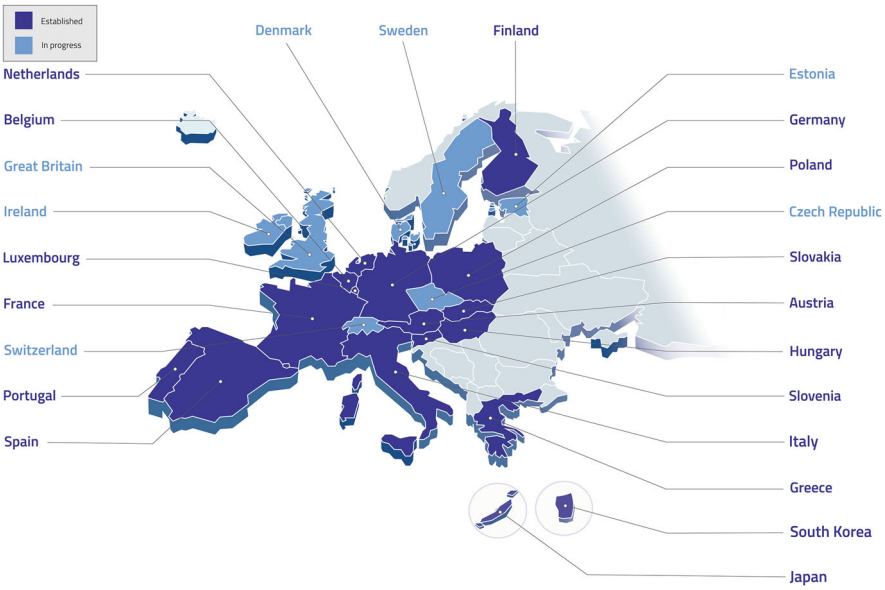


Fig. 4.6 Hub Map

4.5 The National Hub Organization and the Launching of Additional Data Spaces

Gaia-X National Hubs already exist in 16 countries and 5 candidate countries as per Fig. 4.6. Japan and Hungary are also in the process of joining.

Additional Gaia-X Hubs will be created to cover all 27 European countries. Their objective is to collect cloud users' priorities and to channel the various national members to the appropriate European Data Spaces. In addition, national hubs will help addressing new domains such as education, farming, green, and public sector—starting with local companies, possibly within two or three countries, before launching them at a European level.

4.6 Conclusion: Data Spaces—The Enabler of Digital in Business

Data spaces have so far been seen as a nice-to-have feature like standards. They are very useful if they exist, but virtually impossible to justify in economic terms. Therefore, only a political decision can trigger their creation, with consequential long lead times (5–10 years) between the initial intention and the agreed standards.

Cloud adoption is so critical for European industries that Gaia-X has decided to take another route. Initiated by 22 founding members, Gaia-X has demonstrated

during the Summer of 2020 that 159 new members from all over the world were in line with the Gaia-X value proposition. Currently, there are more than 334 members.

During the Gaia-X summit attended by more than 5000 participants, two additional proofs were given: (1) the data space concept is now understood at board level (three of the six data spaces were presented by board members of Philips, Airbus, and CDC), and (2) the hyperscalers (representing more than 70% of cloud provision in Europe) are ready to adhere to this approach of the business.

Many similarities can be seen with the situation of the Internet 40 years ago. The 70s saw the emergence of the internet protocol (IP) as a new way for data transmission through packets; it was seen as an efficient data transmission protocol that allowed better use of transmission circuits by sharing resources thanks to software. Before the invention of the web, it remained a technical mechanism for telecom operators. But the web could not have existed without IP, and IP only took off with the generalized use of the web.

Gaia-X partners believe that data spaces will have a similar role in the business for cloud; there will be no take off of cloud in business without a strong adoption of data spaces by companies belonging to the same ecosystem, but conversely data spaces will not take off if cloud providers do not offer portability and interoperability of their services.

References

1. List of Gaia-X Founding Members: The founding members on the German side include Beckhoff Automation, BMW, Bosch, DE-CIX, Deutsche Telekom, German Edge Cloud, PlusServer, SAP, and Siemens. In addition, the Fraunhofer-Gesellschaft, the International Data Spaces Association, and the European cloud provider association Cispe are cofounders of the Gaia-X Association. On the French side, Amadeus, Atos, Docaposte, EDF—Électricité de France, IMT—Institut Mines-Télécom, Orange, Outscale, OVHcloud, Safran, and Scaleway are among the participants.
2. Gaia-X Franco-German Position Paper. https://www.bmwi.de/Redaktion/DE/Downloads/F/franco-german-position-on-gaia-x.pdf?__blob=publicationFile&v=10
3. European Commission Digital Strategy. https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf
4. Tardieu, H., et al. (2020, March). *Deliberately digital: Rewriting enterprise DNA for enduring success*. Springer.
5. Tardieu, H. (2021, June). Data sovereignty and compliance. *The Journal of Regulation & Compliance (JoRC)*, to be published. https://mafr.fr/fr/article/data-sovereignty-and-compliance/
6. Lamberty, R., de Waard, D., & Poddey, A. (2020). *Tokenomics: Primer*. https://arxiv.org/abs/2008.02538

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 5
Legal Aspects of IDS: Data Sovereignty—What Does It Imply?



Alexander Duisberg

Abstract The claim of data sovereignty is inherently linked to putting the legal instruments and tools in the hands of each participant in the ecosystem, allowing freedom of contract as well as ensuring that exercising data exchange and consorted data usage in the data economy is in compliance with general and specific regulations, ranging from anti-trust to GDPR and cyber-security regulations as well as sector specific regulations. The IDS provides a framework and a technology to allow the parties to limit their transaction costs and to ensure effective enforcement through the concept of usage control. In a future world, this will include increased automation of contract execution (conclusion, performance, and enforcement), whereas the steps to reach that goal are plentiful and, as of now, still require to "set the scene" with the means of the traditional contractual agreements. This article provides an overview and orientation on the key legal areas and aspects to consider for stakeholders, participants, and the business more generally and in the application of the IDS architecture.

5.1 Data Sovereignty: Freedom of Contract and Regulation

The claim to success for the IDS is driven by the combination of a common framework of values and reliability, including the reference architecture, connector technology, certification, and an ecosystem that any and all contributors and participants can trust, in order to share and exploit personal and non-personal data within a multitude of sectors and appliances.

At its heart, IDS intends to facilitate and enable the freedom of research, development, and business by sharing and using data between different players *and* giving any contributor of data the opportunity to manage and maintain control over the data that it puts at the disposal of others.

A. Duisberg (✉)
Bird & Bird LLP, Munich, Germany
e-mail: Alexander.Duisberg@twobirds.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_5

The claim of data sovereignty is inherently linked to putting the legal instruments and tools in the hands of each participant in the ecosystem, allowing freedom of contract as well as ensuring that exercising data exchange and consorted data usage in the data economy is in compliance with general and specific regulations, ranging from anti-trust to GDPR and cybersecurity regulations as well as sector-specific regulations. The following intends to provide an overview and orientation on the key areas to consider for stakeholders, participants, and the business more generally.

5.1.1 No Ownership or Exclusivity Rights in Data

Whereas the initial discussion about "who owns the data" dominated the initial phase of the data economy, policy makers, practitioners, and academics have now achieved an—close to unanimous—understanding that the defining and allotting "ownership" or other forms of exclusivity rights in data per se will not facilitate the development of the data economy. As a result of extensive consultation and debate, the focus has shifted to considering the issue of access and re-utilization of data as key to foster sharing and exchange of data, in order to unleash the potential of innovation through data, both in the private and the public sector.[1] While the outcome might appear "counter-intuitive" to some participants ("How can I not own 'my data'?"), it is clear that it cannot be up to legislation or government to take decisions which could favor one side of the market and ecosystem, for example, the "data producer" or the "data holder" or the "data processor" or the "data aggregator," etc. It is a grown consensus that while any unilateral determination of "ownership rights" in data would be premature while entering into and exploring the potential of largely unknown territories of the data economy, it appears quite likely to prevent rather than facilitate innovation through data.

As a result, the means and tools of enabling the data economy are based in contract law, i.e., regulation of data rights *inter partes* rather than *erga omnes*. The success of data sharing in the data economy depends on proper mechanisms of enforcing the contractual rights throughout the ecosystem.

By putting sample contracts at the disposal of all participants of the ecosystem, IDS gives orientation as well as the freedom to create contracts of their own that data providers, data brokers, and data consumers and any other participants can construe and implement their models for sharing data through licensing agreements of all sorts and kinds, without being prescriptive as to the kind and nature of the contractual relations.[2]

[1] EU Commission (19 February 2020) A European strategy for data. https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf. Accessed 28 January 2021; and Recitals of Regulation (EU) 2018/1807 on a framework for the free flow of non-personal data in the European Union (Free Flow of Non-Personal Data Regulation)

[2] See further details on the different kinds of data licensing agreements under Sect. 5.2.2.1.

5.1.2 Usage Control: Legally and Technically

Based on the concept of data licensing contracts that stipulate data usage rights under respective contract terms, the concept of IDS is to provide a technical solution which enforces such usage rights. The objective of IDS is to add to every dataset the respective protocols that define which data users ("data consumers") are authorized and shall be technically enabled to access the data that a data provider and/or data broker wishes to share with them, under which purposes and for which duration. That may include whether a user may extract, develop, combine, and further enhance such data, as well as onward-share the data and/or any derivatives of such data and related database, next to the dealing with what shall happen after such usage rights have ended.

In other words, IDS has the overall objective of usage control by combining organizational elements under the auspices of freedom of contract with a technical solution. That said, it does not (yet) have the aim of providing a fully automated technical implementation of all contractual parameters as an executable in binary form (comparable to the concept of smart contracts in the blockchain/distributed ledger technology).[3] In fact, the path toward parameterizing different types of contracts and ascertaining related contractual remedies under a governing law (to be selected) bears a multitude of complexities, which need to be further explored. The current effort of the "Legal TestBed" initiated by the "Plattform Industrie 4.0"[4] is a first important step in that direction. By nature of how the formation and interpretation of contracts work, it is not a trivial task and, hence, important to manage expectations what semi-automated contracting and contract enforcement can achieve.[5] Yet, the vision of IDS is right and the implementation requires a legal framework that includes and supports the implementation of technical usage control. With that, usage control will have a stronger effect than the traditional licensing models.

[3] For legal implications of the blockchain in Industrie 4.0 context, reference is made to the publication of the Working Group 4 of the Plattform Industrie 4.0 Blockchain and the Law in the Context of Industrie 4.0, available at https://www.plattform-i40.de/PI40/Redaktion/EN/Downloads/Publikation/blockchain-and-the-law-industrie4.0.pdf?__blob=publicationFile&v=6. Accessed 28 January 2021.

[4] The initiative "Legal TestBed" works on developing a digital testbed to investigate automated business processes and aims to provide policy makers and companies with recommendations for action concerning new legal standards. See further information on the initiative at https://legaltestbed.org/en/start/. Accessed 28 January 2021.

[5] Reference is made to the project and activities of the Working Group "Legal TestBed," which is working on a limited, automated contract negotiation and formation project, and further information available at https://www.plattform-i40.de/PI40/Redaktion/DE/Kurzmeldungen/2020/2020-10-06-rethinking-law.html (only available in German) and https://legaltestbed.org/en/the-testbed/. Accessed 28 January 2021.

5.1.3 Database Rights

Under current law, the most notable (yet largely unknown and underestimated) legal instrument available and applied to collections of data is the right of the database maker, as provided under the EU Database Directive 96/9/EC and implemented in each EU Member State.[6] The database right protects the investment into the systematical and methodical order of a collection of data in order to prevent extraction and/or re-utilization of the whole or a substantial part of the contents of the database,[7] but not the data as such.[8] That said, the database rights provide an important tool in managing data collections, which needs to be considered much more thoroughly in the context of the data economy. The database rights are limited to rightholders (including enterprises) who are nationals or have their habitual residence in the EU, and are construed as a sui generis right, giving the investor exclusivity rights for a duration of 15 years. It includes that the rightholder may transfer, assign, or grant usage rights under a contractual license; it applies independently of whether the actual content or the database as such is (also) eligible to copyright protection.[9,10] It is important to note that the holder of database rights does not enjoy absolute protection, but may only claim a breach of his rights where he has "made available to the public" if a lawful user extracts or re-utilizes other than insubstantial parts of its contents.[11]

Also, any substantial change to the existing database (e.g., made by a lawful user) may result in the creation of a fresh sui generis right in such new database.[12] EU Member States may also define certain usage rights, as permitted exceptions to the sui generis right, such as data extraction for private purposes, teaching, and scientific research, as well as extraction and/or re-utilization for public security or administrative or court procedures.

In essence, any provider of structured data should examine whether he/she can claim the database rights. If so, the rightholder should consider all options for granting licenses in the database as well as safeguarding his/her legal position against unwanted modifications and alterations, which could result in the creation of new sui generis rights.[13]

---

[6] E.g., Sections 87 lit. a–e German Copyright Act

[7] See Art. 7 para. 1 EU Database Directive.

[8] See Rec. (48) EU Database Directive, cf. "whereas the provisions of this Directive are without prejudice to data protection legislation" and *Rezlauf*, Holistic Approach to Handling Big Datasets Including Personal Data for EU-Companies, CRi 2020, 69.

[9] See Art. 7 para. 4 EU Database Directive.

[10] For example, a digital music file is a data collection (of bits and bytes) that enjoys copyright protection, whereas a data collection of sensor data retrieved from a machine does not.

[11] See Art. 8 para. 1 EU Database Directive.

[12] See Art. 10 para. 3 EU Database Directive.

[13] The sample contracts of IDS cater for this situation.

Notably, the EU Commission has envisioned as part of its EU Data Strategy to possibly revisit the EU Database Directive, in order to further enhance data access and use (see Sect. 5.1.6).[14]

5.1.4 Trade Secrets

Confidentiality agreements are generally a viable tool to protect confidential information. But how and what do you keep secret if you share data? What seems an oxymoron by nature needs to be carefully considered in any kind of data transactions.

When sharing data, it is not in first place the information that a data provider shares. It is rather the data provider who intentionally enables the data consumer (or a variety of them) and other participants in the ecosystem to derive and/or generate, each individually, new and different information and value from using the data. In other words, it is important to understand that the data provider cannot necessarily maintain control over what a data consumer makes out of the data that he/she provides.

When further looking into the information models of IDS,[15] the data provider does have certain control over the metadata that he/she shares and, thus, can define or limit the scope of possible conclusions that a data consumer can draw on the data provider's sensitive business information.

To give a practical example, the owner of a steel mill that shares runtime data of his/her machines in real time is providing considerable transparency about his/her current level of bandwidth and manufacturing capacity at any given point in time. He/she will want to avoid that this information is disclosed to his/her competitors and/or intermediaries who might use the information to influence market pricing. The factory owner who is interested in sharing data with a supplier of predictive maintenance services for his/her steel mill may want to (1) enter into confidentiality agreements as well as (2) select and limit the type of metadata that he/she shares with the service provider and certain intermediaries. When going further in sharing data with a business innovator working on an AI-based optimization of the manufacturing process, he/she may want to limit the data he/she shares to other parts of the metadata, in respect of the same manufacturing process.

[14] See p. 13 EU Commission (19 February 2020) A European strategy for data.

[15] The primary purpose of the information model as part of the IDS Reference Architecture is to enable (semi-)automated exchange of digital resources within a trusted ecosystem of distributed parties, while preserving data sovereignty of data owners. Once the relevant resources are identified, they can be exchanged and consumed via semantically annotated, easily discoverable services (see Section 3.4 of the IDS Reference Architecture Model 3.0, available at https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-Reference-Architecture-Model-3.0.pdf. Accessed 28 January 2021).

In addition to the tools of usage control and confidentiality agreements, however, it is important to consider the scope and inherent limitations under the EU Trade Secrets Directive (EU) 2016/943 and its varying implementation into national law of the EU Member States.

In essence, a data provider can claim trade secret protection for "*information*... which...is secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to personal with the circles that normally deal with the kind of information in question...and *has commercial value because it is secret*...and has been *subject to reasonable steps*...*to keep it secret*."[16] While this test can exclude trade secret protection for data with no information value, it is clear that the commercial value is likely to lie in the related metadata. A data provider must therefore determine (and document internally) *before* sharing it whether certain data has commercial value *because* it is secret and which protective measures the data provider has taken to keep it secret (such as limited access internally on a need to know basis, technical security measures, etc.). Subsequently, the data provider will be in a position to define access rights for the data, subject to confidentiality agreements with data consumers, data brokers, and others, and which must be combined with a reliable technical and organizational framework in order to prevent unauthorized third-party access so that the trade secret protection extends when sharing the data.

The benefits of trade secret protection in shared data are obvious, as the trade secret owner has actionable rights to request cease and desist against unlawful data usage (i.e., without the trade secret owner's consent), claim damages against misappropriation of trade secrets, etc.[17]

5.1.5 Competition Law

Competition law in the digital world raises the most complex and currently uncertain issues to be solved.[18] The traditional tools of merger control have been expanded over the last few years, in order to capture and regulate scenarios where the market

[16] Art. 2 para. 1 lit. a–c EU Trade Secrets Directive; italics by the author

[17] See Art. 4 and 6 EU Trade Secrets Directive and its national implementations, e.g., Sections 6–8 German Trade Secrets Act.

[18] On December 15, 2020, the EU Commission has proposed the regulation on contestable and fair markets in the digital sector (Digital Markets Act) (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0842&from=de). Article 12 Digital Markets Act sets out an obligation to provide information on concentrations for the undertakings within the meaning of the Digital Markets Act. Also, the German legislator reformed the German Act Against Restraints of Competition (ARC). Largely this amendment entered into force on January 19, 2021. In future, mergers will only be subject to control if, among other things, one of the companies involved has an annual turnover in Germany of at least EUR 50 million, instead of the previous EUR 25 million, and in addition, another company involved has an annual turnover in Germany of at least EUR 17.5 million, instead of the previous EUR 5 million.

impact is less influenced by the mere size of the merging companies, but rather to also address situations where the focus of the contemplated transaction is on a data-rich target company with limited economic size considered by the traditional parameter.[19] Further, the investigations conducted by the German Federal Cartel Office against Facebook, in which it looked at the terms of use to capture a scenario of (allegedly) excessive data collection, give an indication on how authorities are trying to expand their reach in regard to the control of market dominant positions that rely on the accumulation of vast amounts of data and related data-driven business models.[20] Ultimately, the EU Commission has addressed in its EU Data Strategy the issue of imbalance in market power in relation to data-rich businesses ("data advantage"), putting on the agenda the need to define new ways of preventing market distortion and lack of competition,[21] leading in particular to the new EU Digital Markets Act.[22]

In its decision of February 6, 2019, based on Sec. 19 para. 1 ARC, the German Federal Cartel Office prohibited Facebook from, inter alia, using conditions that make the use of the eponymous social network Facebook by private users resident in Germany dependent on Facebook being able to link and use user- and device-related data collected during the use of the group's own services such as WhatsApp with the user accounts maintained for Facebook.com without the consent of the users. However, with decision of August 26, 2019 (VI-Kart 1/19 (V)), as a provisional order, the Higher Regional Court of Cologne found that even in the case of an assumed data protection infringement, Facebook had not abused its market-dominating position within the meaning of Sec. 19 para. 1 ARC. Hence, the Higher Regional Court Düsseldorf had ordered the suspensive effect of the appeal at Facebook's request. At the request of the German Federal Cartel Office, the Cartel Senate of the Federal Court of Justice reversed this order in a decision of June 23, 2020 (BGH, 23 June 2020—KVR 69/19), and rejected Facebook's application for an order of suspensive effect of the appeal.

[19] According to the explanatory memorandum to the 9th amendment of the ARC 2017 regarding Sec. 18 para. 3a ARC, the market position of an undertaking can also be significantly influenced by its access to data; see BT-Drs. 18/10207.

[20] It would exceed the scope of this contribution to discuss in detail the implications of this decision and the pending appeal procedures. But among the various questions to debate, one also needs to consider to which extent anti-trust authorities have the authority to assess the validity of GDPR compliance in such context (see the decision of the German Federal Cartel Office (https://www.bundeskartellamt.de/SharedDocs/Entscheidung/DE/Entscheidungen/Missbrauchsaufsicht/2019/B6-22-16.pdf?__blob=publicationFile&v=8. Accessed 28 January 2021), and the decisions of the OLG Düsseldorf, August 26, 2019—VI-Kart 1/19 (V), and of the German Federal High Court of Justice, June 23, 2020 – KVR 69/19. Also see the review by *Haus* and *Cesarano*, Mehr-Daten für Facebook, NZKart 2019, 637 and the remark of the last-named decision by *Mohr*, LMK 2020, 432972.

[21] "Progress will need to be made together on the following issues: A case in point comes from large online platforms, where a small number of players may accumulate large amounts of data, gathering important insights and competitive advantages from the richness and variety of the data they hold. This can affect, in turn, the contestability of markets in specific cases—not only the market for such platform services, but also the various specific markets for goods and services served by the platform, in particular if the platform is itself active on such related markets." See p. 8 EU Commission (19 February 2020) A European strategy for data.

[22] See Sect. 5.3.2 with some further explanations on the EU Digital Markets Act.

Where companies exchange sensitive or confidential information relating to their market position (e.g., price information, but also other information that might allow coordinated behavior), each party must ensure—and assumes proper responsibility—not to restrict competition and/or distort markets through agreements restraining competition or resulting in coordinated behavior.[23] While exchanging, for example, sensor data through IDS is per se unlikely to result in direct coordination over pricing between competitors, or other restraints and anti-competitive practices (both at a horizontal level between direct competitors, and vertically between the different levels in a chain of distributing products and services in a given market), it is clear that each participant in a bilateral or consorted exchange of *data* assumes responsibility and must ensure to comply with applicable competition law. Accordingly, data providers, data brokers, and data consumers may need to limit the exchange of *market relevant information*, which is contained in or could be directly derived from exchanging "raw data" together with relevant metadata.

As a further consequence, the operator of a data space will want to safeguard in its information notices and terms of use that each participant to a data exchange is properly aware and undertakes to avoid exchanging market-sensitive information that could be abused and/or that could result in coordinated behavior.

5.1.6 EU Strategy on Data: The Relevance of Data Spaces

The EU Commission has set important milestones for transforming the single market into a digitally enabled market and making the EU "leading in a data-driven society" and empowering "people, businesses and organisations . . . to make better decisions based on insights from non-personal data, which should be available to all"[24] and creating a "data-agile economy."[25] Together with this ambitious claim, the EU Commission has presented its EU Data Strategy as a cornerstone to a wider framework of existing and upcoming regulation.[26]

The EU Data Strategy envisions three fundamental objectives, namely, (1) the free flow of data within the EU and across sectors; (2) full respect of European rules and values, including in particular personal data protection, consumer protection, and competition law; and (3) fair, practical, and clear rules for fair access and use of data, based on trustworthy data governance mechanisms.[27]

The EU Commission envisions new legislative measures to support those objectives, including in particular: (1) a cross-sectoral governance framework for data

[23] See Art 101–109 Treat of the Functioning of the European Union TFEU.

[24] See p.1 and 4 EU Commission (February 19, 2020) A European strategy for data.

[25] Ibid p.8

[26] EU Commission (19 February 2020) A European strategy for data

[27] Ibid. p. 5

access and use, (2) making available more high-quality public sector data for re-use, (3) a data act for horizontal data sharing across sectors, which may include revisiting the EU Database Directive and the EU Trade Secrets Directive, in order to facilitate and enhance access and (re-)use of data.[28]

Most notably, the EU Commission recognizes and endorses the fundamental importance of data spaces as part of the first pillar (1), aiming to "enable a legislative framework of the governance of common European data spaces,"[29] as well as providing significant investment and funding in High Impact Projects on European data spaces and federated cloud infrastructures.[30] It is clear by the wording and further considerations of the EU Commission that IDS and its key elements (i.e., the connector technology, reference architecture, usage control, certification scheme, and model contracts) provide the lead image and stand at the heart of this particular part of the EU Data Strategy. The EU Commission emphasizes that such data spaces shall "overcome legal and technical barriers to data sharing across organisations, by combining the necessary tools and infrastructures and addressing issues of trust, for example by way of common rules developed for the space. The spaces will include; (i) the deployment of data-sharing tools and platforms; (ii) the creation of data governance frameworks; (iii) improving the availability, quality and interoperability of data—both in domain-specific settings and across sectors."[31] The EU Commission has defined as a key action point to that end combined investments in the range of EUR 4–6 billion including direct investments of up to EUR 2 billion.[32]

As part of its data space strategy, the EU Commission has identified the following sectors where it intends to create "Common European data spaces": industrial/manufacturing, Green Deal, mobility, health, financial, energy, agriculture, public administration, and skills.[33] In other words, IDS represents a role model, if not a blueprint, for these sectors to prepare and develop—in an active dialogue with the relevant stakeholders—the related data space implementations in accordance with the EU Data Strategy.

[28] Ibid. p. 11 et seq.

[29] Ibid. p. 12 and 16

[30] To this Gaia-X, which is a project with representatives from politics, business, and science creating secure, federated European system that meets the highest standards of digital sovereignty while promoting innovation (see https://www.bmwi.de/Redaktion/EN/Dossier/gaia-x.html. Accessed 28 January 2021), is a constituent element that has further materialized since the EU Commission set out its Data Strategy in February 2020 (see p. 18 EU Commission (19 February 2020) A European strategy for data).

[31] See p. 16 et seq. EU Commission (19 February 2020) A European strategy for data.

[32] Ibid. p. 19

[33] Ibid. p. 22 et. seq. and its Appendix with further details

5.1.7 Data Governance Act: First Comments

The Data Governance Act of May 30, 2022,[34] is a pillar in the EU Data Strategy and will be complemented by the European Data Act to foster data sharing among businesses, and between business and governments,[35] and stands next to the Digital Markets Act[36] and the Digital Services Act.[37] It contains key elements of regulation for operators of data spaces.

Its principal areas of regulation cover the following objectives: (1) making public sector data available for re-use in situations where such data is subject to rights of others (such as privacy rights, IP rights, trade secrets, or other commercially sensitive information); (2) sharing data among businesses, against remuneration in any form; (3) allowing personal data to be used with the help of a personal data sharing intermediary that safeguards data subjects' rights under the GDPR; and (4) allowing data use on altruistic grounds.[38] As regards public sector data, the Data Governance Act complements and stands in addition to the Open Data Directive.[39] The overall objective is to "facilitate data sharing by reinforcing trust in data intermediaries,"[40] which are expected to play a significant role in data spaces, whereas the rules on access and use of data shall be covered by the Data Act.[41] The Data Governance Act defines as overall requirements that data should be "findeable, accessible, interoperable and re-usable."[42]

The Data Governance Act accentuates the role and provides notification obligations for providers of data sharing services ("data intermediaries"), in an approach to create a European model for data sharing of personal and non-personal data through "neutral data intermediaries," as an alternative to the current prevalence and market

[34] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance (Data Governance Act), OJ L 152/1 of June 3, 2022. Accessed 13 June 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN. Accessed 28 January 2021

[35] See EU Commission press release: Commission proposes measures to boost data sharing and support European data spaces https://ec.europa.eu/commission/presscorner/api/files/document/print/en/ip_20_2102/IP_20_2102_EN.pdf. Accessed 10 February 2021.

[36] Proposal Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0842&from=de. Accessed 10 February 2021

[37] Proposal for a Regulation on a Single Market For Digital Services (Digital Services Act) https://eur-lex.europa.eu/legal-content/de/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN. Accessed 10 February 2021

[38] Ibid. p.1

[39] EU Directive (EU) 2019/1024 on open data and the re-use of public sector information (Open Data Directive); Rec. (5)-(14) of the Open Data Directive and p.1 EU Commission (25 November 2020) Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act)

[40] p.1 EU Commission (25 November 2020) Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act)

[41] Ibid.

[42] Ibid.

power of integrated tech platforms that are commonly run by corporate businesses.[43] While this approach emphasizes the importance of data held in the public sector and ensuring data sharing across Member States,[44] it is by no means limited to the same.

A key requirement to safeguard trust and control over the data sharing between data holders and data consumers is ensuring the neutrality of the data intermediary. That implies that the data intermediary only acts as an intermediary in data transactions and does not use the data for other purposes.[45] The data intermediary shall have an establishment in the EU, or an appointed representative if offering the intermediary services from outside, and must follow a notification procedure (yet to be developed in the Member States) such that it notifies the competent registry/authority of its intention to provide intermediary services. In essence, the focus on data intermediaries accentuates the role and responsibility of operators of data spaces.

With regard to scientific research, for example, in relation to the health sector or environmental issues under the Green Deal, the Data Governance Act defines the role of "Data Altruism Organisations recognized in the Union," which are subject to a voluntary registration regime.[46]

From an institutional perspective, the Data Governance Act will create a "European Data Innovation Board," consisting of representatives of the Member States, the EU Commission, and representatives of relevant data spaces and specific sectors (e.g., health, agriculture, transport, and statistics).[47] It shall coordinate national processes and policies and support cross-sector data use within the "European Interoperability Framework" (EIF).[48]

As for data held by the public sector, the Data Governance Act establishes a few key principles: generally, public sector bodies shall not enter into exclusive agreements for the re-use of data they hold, nor may they restrict the availability of the data for re-use, unless (as an exception to the rule) where a data consumer receives exclusive rights for a maximum of 3 years, in order to provide a service or product in the general interest and under a national concession issued in accordance with general transparency principles.[49] In all other cases, public sector bodies shall grant rights to re-use public sector data based on the rules of transparency, equal treatment, and non-discrimination on grounds of nationality. The actual conditions for re-use must be proportionate and objectively justified with regard to categories of

[43] Ibid. p.6

[44] See Art. 3–8 Data Governance Act.

[45] Rec. (26) and Art. 11 para. 2 Data Governance Act

[46] Art. 1 para. 1 lit. c and Art. 17 Data Governance Act

[47] Rec. (40) Data Governance Act

[48] Rec. (41) Data Governance Act, see https://joinup.ec.europa.eu/collection/connecting-europe-facility-cef/about and https://joinup.ec.europa.eu/site/core_vocabularies/Core_Vocabularies_user_handbook/ISA%20Hanbook%20for%20using%20Core%20Vocabularies.pdf for examples of standards and specifications used by the European Data Innovation Board.

[49] Art. 4 para. 1–3 Data Governance Act

data and purposes of re-use and may define (among others) obligations in regard to secure processing environments provided and controlled by the public sector.[50] Transfers of highly sensitive non-personal data to third countries may be restricted by national Member States laws.[51]

The Data Governance Act requires data intermediaries to follow a notification procedure in regard to the following types of intermediation services: (1) between data holders (as legal persons) and data consumers both in bilateral or multilateral data exchanges, or the creation of platforms or databases that enable the exchange or joint exploitation of data, as well as the establishment of a specific infrastructure for the interconnection of data holders and data consumers; (2) between data subjects that want to make their personal data available and potential data consumers, thereby facilitating the data subjects to exercise their rights under the GDPR; and (3) services of data cooperatives particularly in the areas of micro, small, and mid-size enterprises.[52] The concept of notification does not imply an approval by the authorities,[53] but rather provides a mechanism to determine certain conditions for data intermediary services (before commencing their activity)[54] and to establish a supervisory control over an intermediary's compliance with such conditions,[55] which can impose "dissuasive financial penalties" (to be further defined by the Member States) if need be.[56] The EU Commission stressed that additional measures regarding rights on access and use of data are envisaged for the EU Data Act, as in discussion since February 2022.[57,58]

The conditions under Art. 11 Data Governance Act are of particular interest in the given context of IDS: the intermediary may not use the data it receives for other purposes than putting them at the disposal of data consumers; he/she shall not use the metadata collected from the data sharing service for other purposes other than developing that actual service, ensuring fair, transparent, and non-discriminatory access to the service for data holders and data consumers, including as regards prices; facilitating data exchange in the formats that the intermediary receives the data and converts data into other formats only to ensure interoperability within and

[50] Art. 5 paras. 2 and 4 Data Governance Act

[51] Art. 5 para. 11 Data Governance Act

[52] Art. 9 para. 1 lit. a–c Data Governance Act

[53] Art. 10 Data Governance Act

[54] Art. 11 Data Governance Act

[55] Art. 13 Data Governance Act

[56] *Spindler*: Schritte zur europaweiten Datenwirtschaft—der Vorschlag einer Verordnung zur europäischen Data Governance, CR 2021 p. 98–108 concludes that this results in a "general prohibition subject to a prior notification obligation." That appears rather drastic, whereas a notification obligation can indeed make sense, in order to create visibility for regulatory supervision (without, however, stipulating or accentuating the concept of a general prohibition).

[57] See https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-data-act. Accessed 10 February 2021.

[58] Proposal for a Regulation on harmonised rules on fair access to and use of data (Data Act), COM (2022) 68 final.

across sectors or if specifically requested by the data consumer, or if required under law, or to ensure harmonization with international or European data standards; preventing fraudulent or abusive practices; ensuring continuity of services and adequate technical, legal, and organization measures to prevent unlawful transfer or access to non-personal data; providing a high level of security for the storage and transmission of non-personal data; maintaining procedures to ensure compliance with competition law rules at the EU and Member State level; advising data subjects on potential data uses and standard terms and conditions attached to such uses; and advising on the relevant jurisdiction(s) of processing where an intermediary provides tools for obtaining consent from data subjects or permissions to process data made available by legal person.

From an IDS perspective, the Data Governance Act endorses the approach and key elements that IDS is promoting, including in particular the approach toward neutral intermediaries that rely on the reference architecture and the connector technology, in order to enable data sharing between data holders and data consumers in bilateral and multilateral data sharing ecosystems. The reference architecture and the information model of IDS are coherent with the requirements regarding data formats and interoperability. Operators of data spaces will need to pay particular attention, however, as to their monetization model. Under the current draft Data Governance Act, it appears excluded that an operator of a data space could generate innovative services through further-going metadata analytics, other than in order to "further develop" those data sharing services that the intermediary is actually providing to the specific data holders and data consumers concerned.[59] There are certainly good reasons to argue for such a limited remit, in respect to the definition of "metadata" that relates to the data holders and data consumers.[60] However, it appears important to discuss further clarity on whether an intermediary should not be able to generate ("anonymized") metadata aggregations and reports containing findings and learnings, as well as whether to use such data possibly as training data for machine learning, as long as the information contained in the actual metadata itself is properly protected and not shared with third parties for commercial gains or other unauthorized purposes.

Obviously, the Data Governance Act is subject to further debate, including the relatively generic requirements on security (currently not referencing state-of-the-art security but only referring to "high level of security,"[61] whereas, e.g., Art. 32 para. 1 GDPR shows a way to make sure a controller at least "takes into account the state of the art"). Further, it is important to bear in mind that the Data Governance Act represents only one part of legislation in regard to the European Data Strategy.[62]

[59] Art. 11 para. 2 Data Governance Act

[60] Art. 2 para. 4 Data Governance Act

[61] Art. 11 para. 8 Data Governance Act

[62] *Spindler*: Schritte zur europaweiten Datenwirtschaft—der Vorschlag einer Verordnung zur europäischen Data Governance, CR 2021 p. 98–108, noting, however, that Prof. Spindler does not touch upon the concept of data spaces as such in his wider considerations.

5.1.8 Personal and Non-personal Data

With its definition of personal data, the GDPR has determined an ample scope: "personal data means any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, on online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."[63] Anonymization of personal data can enable a data controller to further process the data without considering the data protection requirements, provided that the data is no longer related to an identified or identifiable person.[64] Please note that pseudonymization of personal data (replacing identifiable personal data by a pseudonym) is reversible and can therefore only be used as an additional security safeguard for the processing. Examples such as IP addresses show that "personal data" is far more than might be obvious on first sight.[65] In addition, the rise of Big Data and AI has clearly shown that what might appear, at a given point in time, as anonymous data or other data with no connection or relevance to natural persons may actually turn out to be an element of personal data, once it is combined with other identifiers.[66]

However, the EU Commission has recognized the distinction by referring to "non-personal data" within areas of (limited) regulation, such as the Free Flow of Non-Personal Data Regulation.

Any concept regarding data spaces, such as IDS, must therefore be prepared to cater for compliance requirements regarding both personal and non-personal data.

---

[63] Art. 4 para. 1 EU Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR)

[64] See also Federal Commissioner for Data Protection and Freedom of Information: Position paper on anonymization under the GDPR with special consideration of the telecommunications sector. https://www.bfdi.bund.de/DE/Infothek/Transparenz/Konsultationsverfahren/01_Konsulation-Anonymisierung-TK/Positionspapier-Anonymisierung.pdf;jsessionid=47F123BF62DE633F32BB671CD74BAF74.2_cid329?__blob=publicationFile&v=2 (only available in German). Accessed 15 February 2021.

[65] Rec. (30) GDPR; p. 16 et seq. WP 29 WP136—01248/07/EN—Opinion 4/2007 on concept of personal data. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf. Accessed 28 January 2021; and ECJ C-582/14 Breyer v. Federal Republic of Germany

[66] "In most real-life situations, a dataset is very likely to be composed of both personal and non-personal data. This is often referred to as a "mixed dataset". Mixed datasets represent the majority of datasets used in the data economy and commonly gathered thanks to technological developments such as the Internet of Things (i.e. digitally connecting objects), artificial intelligence and technologies enabling big data analytics" (p. 8 EU Commission Guidance on the Regulation on a framework for the free flow of non-personal data in the EU https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52019DC0250&from=EN. Accessed 28 January 2021; and p. 13 WP 29 WP136—01248/07/EN—Opinion 4/2007 on concept of personal data).

5.1.8.1 GDPR

From an IDS and data sharing perspective, the main question to be considered is which actors take which roles (controllers, processors, and joint controllers), in order to assess the related obligations. Additional considerations relate to the operators of data spaces in regard to the technical and organizational measures that they provide, in order to enable controllers and processors to share and process (personal) data in a data space.

The roles of data providers (acting as the [original] data controllers)[67] and data consumers (acting as [subsequent] data controllers) appear clear in a bilateral data sharing scenario, i.e., resulting typically in a controller-to-controller transfer and where the operator of the data space provides the infrastructure and takes the role of a data processor.[68] The data provider will need to assess the legal basis for making the data available to the data consumer under Art. 6 GDPR (e.g., consent of the data subject, performance of a contract with the data subject, legitimate interest) and ensure compliance with further obligations of a data controller (e.g., privacy notice under Art. 13 and 14 GDPR; safeguarding data subjects' rights under Art. 15 et. seq. GDPR, documentation obligations under Art. 30 GDPR, etc.). When it comes to scientific research—in particular in regard to special categories of data (e.g., health data, Art. 9 GDPR)—it may well be that data providers can claim specific legal justifications that the Member State legislators may have enacted, under the opening for national derogations.[69]

When it comes to multilateral data sharing, the data provider will need to consider the legal basis for providing the personal data concerned in regard to the data consumer separately—which may result in different legal basis applying, depending on the nature of processing, the various data consumers envision. Notably, the scope for "purpose variation" is limited under Art. 6 para. 4 GDPR.[70] Where the data provider relies on consent, he/she will need to provide appropriate consent management tools (including the option to withdraw consent); where he/she relies on legitimate interest, he/she will need to safeguard the right of objection.[71]

More complexity comes about where various data controllers jointly determine the purposes and means of processing personal data. This can occur in multilateral or consorted data sharing scenarios, either between the data provider and various data consumers or also (only) between various data consumers. In each of those

[67] Defined as the natural or legal person that "determines the purposes and means of the processing of personal data" (Art. 4 para. 7 GDPR)

[68] Art. 28 GDPR

[69] Art. 6 para. 2, Art. 9 para. 2 lit. g–j GDPR

[70] The [subsequent] data controller's processing purposes would need to maintain an inherent connection to the original processing purposes under which the [original] data controller had collected and transferred the personal data in the first place; in that context, notably, pseudonymization may provide a suitable safeguard (Art. 6 para. 4 lit. e GDPR).

[71] Art. 21 GDPR

scenarios, the parties involved will need to enter into a joint controllership agreement[72] and will assume joint and several liabilities for data protection compliance of their jointly controlled processing activities.[73]

The operator of a data space should put appropriate tools in the hands of data controllers, in order to ease the data providers' implementation of GDPR compliance. That can work by providing standardized documentation which the data controller(s) and processor(s) concerned can easily adapt and conclude as required—yet recognizing that a fully automated compilation of relevant documentation is very likely still a long way to go.

5.1.8.2 Free Flow of Non-Personal Data Regulation

The EU Regulation on the free flow of non-personal data[74] has two core objectives: ensuring the free movement of non-personal data across Member State borders, i.e., removing data localization requirements between Member States (and preserving availability and access to data for regulatory control purposes)[75] and easing the portability of data (in particular with regard to professional users switching cloud providers).[76] The regulatory approach on data portability is "soft-handed" and intentionally not interventionist, but self-regulatory, requiring further development through a "code of conduct" at the EU level.[77] From an IDS perspective, the relevant claim of data portability essentially regards the relation between the data provider and the operator of a data space. The data provider must have the option to move his/her account to another data space. IDS' reference architecture, the information model, and the data connector technology that allows to process and connect interoperable data formats are suitable measures to meet these requirements, which should therefore be instrumental in preparing related "code of conduct" under Art. 6 of the Free Flow of Non-Personal Data Regulation, if and when required.

5.1.9 Cybersecurity

The rise of cybersecurity threats is inherent to the growth of digital ecosystems and data sharing within data spaces. Clearly, robust cyber resilience and related organizational measures are a pre-condition for data providers and data consumers to share personal and non-personal data. However, it is also a question of regulation. The NIS

[72] Art. 26 GDPR

[73] Art. 82 para. 4 GDPR

[74] (EU) 2018/1807 of 28 May 2019

[75] Rec. (13) and (18) and Art. 4 para. 1, Art. 5 Free Flow of Non-Personal Data Regulation

[76] Art. 6 Free Flow of Non-Personal Data Regulation

[77] Art. 6 para. 1 Free Flow of Non-Personal Data Regulation

Directive and the Cybersecurity Act are key pillars of a European cybersecurity framework and are complemented by the requirements on technical and organizational measures in regard to personal data[78] as well as security measures required by data intermediaries.[79]

5.1.9.1 NIS Directive

The NIS Directive and its implementation into national security laws set the framework for adequate security of providers of essential services as well as digital service providers.[80] Besides the providers of essential services, the NIS Directive requires EU Member States to impose security requirements also on providers of digital services, which are defined as online marketplaces, online search engines, and cloud computing services.[81] Cloud computing services are defined as "a digital service that enables access to a scalable and elastic pool of shareable computing resources." The EU Commission has issued an Implementing Regulation[82] on the basis of Art. 16 para. 8 NIS Directive that specifies the security obligations of the providers of digital services. Accordingly, providers of essential services and/or digital services that fall within the scope of the NIS Directive will be able to use IDS, if and where they can define, configure, and rely on the security settings for data exchange.

5.1.9.2 Cybersecurity Act

The EU Cybersecurity Act (CSA) complements the provisions of the NIS Directive with additional regulations on the tasks and powers of the EU Agency for Network and Information Security (ENISA) and with the baselines of a new cybersecurity certification scheme for Information and Communications Technology (ICT) products services and processes. This cybersecurity certification scheme under Art. 51 CSA is still under development by the ENISA. On May 27, 2021, the Federal Government adopted and published its revised German IT security Act 2.0 (ITSiG 2.0). With the new ITSiG 2.0, the German Federal Offices for Information Security's powers is largely expanded and provisions inter alia regarding the storage of log

[78] Art. 32 GDPR

[79] Art. 11 para. 7 and 8 Data Governance Act

[80] In the future, subject to the updated version of the NIS 2.0 Directive ("NIS2"), as per the Council's and EU Parliament's adopted version of 13 May 2022. Notably, Member State implementations vary under the Directive, including the related level of sanctions and enforcement. Germany has implemented an enhanced "IT Security Act 2.0" on 27 May 2021 (anticipating some of the changes discussed at the NIS2 level subsequently), raising the bar of sanctions up to EUR 2 million (Section 14 para. 5 ITSiG 2.0).

[81] Art. 4 no. 5 and Annex III NIS-Directive

[82] Commission Implementing Regulation (EU) 2018/151 of 30 January 2018

data, inventory data disclosure, and implementation of detection measures for network and IT security are implemented.[83] Where regulated entities need to follow these requirements, IDS can potentially offer the architecture and framework for related compliance, noting, however, that it remains within the responsibility of the regulated entity/ies to define and implement their security requirements for related data exchange.

5.2 Preparing Contractual Ecosystems

The IDS sets out the landscape of participants in the ecosystems, and deriving from that, which participants need to bound through contractual agreements with other participants, in order to support the data exchange between the data providers and data consumers (Fig. 5.1).

The entire concept of IDS and data sovereignty is based on the principles of contract law, in order to ensure that data providers can determine and enforce the rules and conditions under which they share with and enable data consumers to use their data, be it in bilateral ("1:1") or multilateral usage scenarios ("1:1" and "1:*n*"). In that context, the fundamental principles of freedom of contract, including the freedom to choose the governing law, must always be at the disposal of the contracting parties.

IDS provides the framework and technology to allow the parties to limit their transaction costs and to ensure effective enforcement through the concept of usage control. In a future world, this will include increased automation of contract execution (conclusion, performance, and enforcement), whereas the steps to reach that goal are plentiful and, as of now, still require to "set the scene" with the means of traditional contractual agreements.

The following considerations explain some of the fundamental concepts and approaches for data licensing agreements, i.e., those agreements that data providers and data consumers will conclude, and how that can work on the basis of platforms that enable such data exchange. While the following explanations stand against the background of German law (and hence need to bear in mind that underlying statutory law can impact the formation and interpretation of contracts), they are to a considerable degree generic in nature and can be applied to other jurisdictions (even if adaptations under local law remain indispensable).

[83] Draft ITSiG 2.0 (9 December 2020) https://intrapol.org/wp-content/uploads/2020/12/IT-SiG-2.0-RefE-Stand-9.12.2020.pdf (only available in German). Accessed 28 January 2021

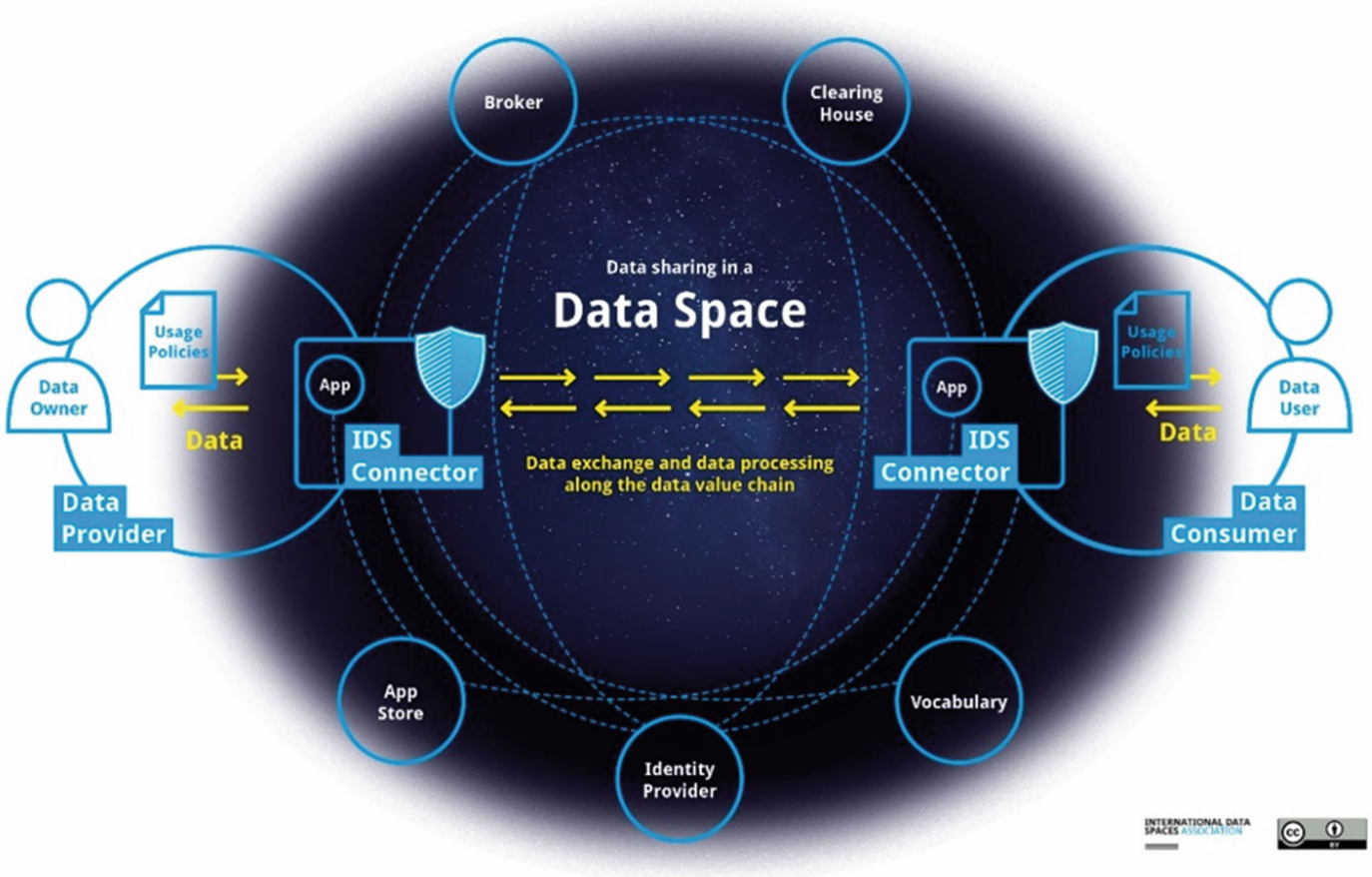


Fig. 5.1 Data sharing in a data space. © 2021, International Data Spaces Association [CC-BY]. Used under permission from International Data Spaces Association

5.2.1 Platform Contracts

As mentioned above, the role of intermediaries and, in particular, operators of platforms is key for the success of facilitating data exchange and accelerating the growth of the data economy.[84] Accordingly, it is important to consider the contractual setting that a provider of data connector services has to offer, possibly in first place and by which the data providers and data consumers can transact. That said, it is equally possible that data providers can do without the services of a platform provider, if and where they simply draw on the IDS connector technology and organize the data exchange (1:1 or 1:*n*) by themselves.

5.2.1.1 Key Principles

A platform operator that offers services to facilitate a data exchange will typically define its contractual relationship with data providers and data consumers by general terms and conditions, similar to those of other digital marketplaces. The platform operator will want to consider the following key constituent elements in this context: (1) the platform that provides the technical infrastructure and related services (including service levels) for a reliable and secure data exchange, but is not prescriptive as to the actual commercial and legal terms under which data providers and data consumers perform data exchanges; (2) the platform operator that may put at the disposal of data providers and data consumers template contracts to facilitate transactions and reduce transaction costs for the various types of data licensing transactions, including related compliance documentation (such as data processing agreements as required under the GDPR) and may provide technical mechanisms to facilitate automated or semi-automated contracting as well as facilities for contract negotiation; (3) the platform operator that will typically set out a registration process for data providers and consumers, as well as define an acceptance process for the platform operator's terms of use; (4) the platform operator that will set certain requirements for accepting policy frameworks (such as the IDS reference architecture, security settings, permitted usage and exclusions, requirements on IPR and GDPR compliance), codes of conduct, etc. that the data providers and data consumers have to follow; (5) remuneration and usage fees; (6) provisions on warranty and liability in regard to the functioning of the platform; (7) indemnities resulting from possible third-party claims raised against the platform operator resulting from the data providers' transactions conducted with data consumers; (8) provisions and limitations regarding the platform operator processing, using and retaining data exchanged through the platform by data providers and data consumers; (9) term and termination; (10) confidentiality; (11) GDPR compliance in the relation between the platform operator and the data provider/data consumer and between the data

[84] See also Sects. 5.1.4 and 5.1.7.

provider and data consumers; (12) IDS certification of the platform operator; and (13) choice of governing law and dispute resolution.

These many aspects to consider show that contractual frameworks of platform operators cannot be easily transformed into binary executables or simple "smart contracts," as well as that IDS-based platform operators need to resort to a predefined governing law and rules on dispute resolution (which can include online arbitration) in a particular jurisdictions. However, IDS can provide a template for platform operators. The sample "Terms and conditions of participation in an Industrie 4.0 platform" certainly gives a very good starting point and is available under a Creative Commons license.[85]

5.2.1.2 Legal TestBed: A Lead Example

As part of its activities, the legal working group of the "Plattform Industrie 4.0" has initiated a widely remarked "Legal TestBed" that is designed as a sand-box exercise for simulating a legal contract execution process (conclusion, performance, and enforcement) in an Industrie 4.0 context.[86] As part of the exercise, the working group has created the "Terms of use for an Industrie 4.0 platform" (Terms of Use), by way of an extensive drafting and consultation process among legal practitioners of academia, industry, and private practice. These Terms of Use are designed to ensure a reasonable balance between the interest of the platform operator and the users (data providers and data consumers), in order to facilitate data transactions and/or operational processes (such as performance of a logistics order and performance process) on the platform.

The Terms of Use cover the principles set out above (Sect. 5.2.2.1). By virtue and subject to the terms of the Creative Commons license, any third party is free to use and adapt these Terms of Use for its own platform operations.

5.2.2 Data Licensing Agreements

While building the data economy is a process that has only started, it appears that the commercial practice of data licensing is advanced in certain specific areas, whereas in other areas and sectors it is still largely "unknown territory." Accordingly, data licensing agreements do not yet follow general common standards that practitioners can "pull off the shelf," such as in the world of software licensing. In any event, therefore, it is helpful to be aware of the fundamentally different types of contractual

[85] https://www.plattform-i40.de/PI40/Redaktion/EN/Downloads/Publikation/RTB_contract_template.html. See the following chapter.

[86] For further information on the initiative, see https://legaltestbed.org/en/start/. Accessed 9 February 2021.

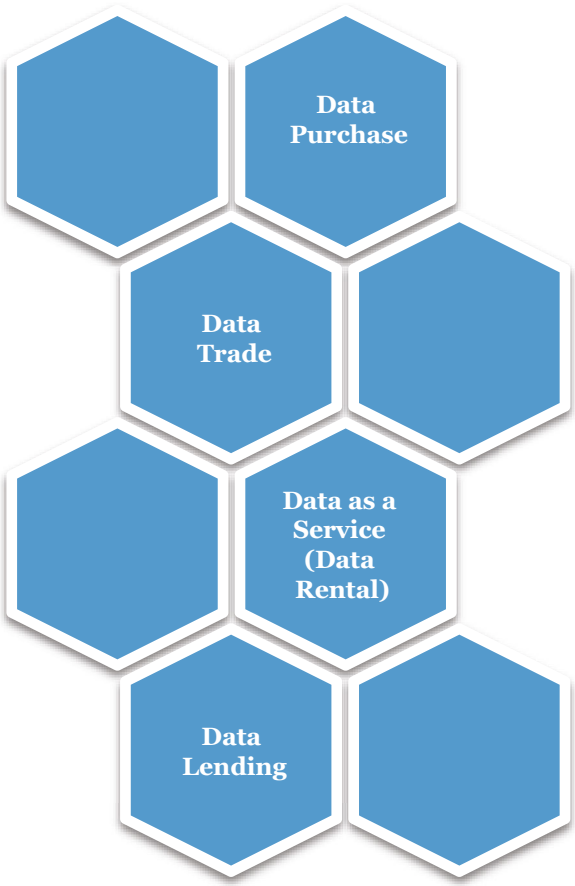


Fig. 5.2 Categorization of contract types. © 2021, Dr. Alexander Duisberg, Bird & Bird LLP

arrangements that data licensing transactions can follow. In simplified terms, the key distinctions that any data provider must consider are around the nature of the transaction, i.e., whether he/she intends to grant (1) perpetual or temporary, (2) exclusive or non-exclusive, and (3) royalty-free or paid-up usage rights—and in which combination of each of these aspects. From a German legal perspective (which can at least be helpful also for other civil law jurisdictions), the categorization of contract types can help in this regard (Fig. 5.2).

A "data purchase" implies perpetual (exclusive or non-exclusive) usage rights against a one-time remuneration. "Data as a Service" implies temporary (exclusive or non-exclusive) usage rights in data (comparable to a rental model), whereas "data lending" would imply that the lender asks for no compensation, and in a "data trade" the data provider would receive data as a non-monetary compensation. From a German law perspective, each of these different transactions falls in the category of a different contract type, to which the German Civil Code attaches different requirements, as well as contractual remedies in case of a breach. As a result, the

data provider needs to consider the legal consequences and risks involved when setting his/her contract terms against that background.

5.2.2.1 The Contract Matrix

The following matrix provides initial guidance (under German law) on the key elements to consider in relation to the various parameters applied to the different contract types. Some of these parameters are suited for a binary implementation (e.g., exclusive or non-exclusive usage rights, etc.) which therefore facilitates automated contracting (Fig. 5.3).

Obviously, any exclusive grant of usage rights is limited to 1:1 data licensing transactions—and would (potentially) even exclude further usage by the data provider, unless he/she explicitly reserves such rights. An important element to consider is the sui generis database right which is based on the (unique) EU Database Directive and its implementation to EU Member States laws.[87] A data provider that licenses structured data will need to consider the implications, i.e., whether and to which extent he/she defines limitations on the data consumer in creating new database rights by investing into substantively different methods of making datasets searchable.[88]

5.2.2.2 The IDS Sample Contracts

In addition to the "Terms and conditions of participation in an Industrie 4.0 platform," the IDS itself has developed two basic templates to cover data purchases and data as a service type of licensing transactions, designed against the background of German law and considering the particular implications of German rules governing standard terms and conditions. Again and as stated above, the intention of providing template agreements is not about being prescriptive, but rather to endorse the overarching principle of freedom of contract, whereas trying to reduce the transactional costs of setting up and negotiating suitable contracts for data licensing.[89]

[87] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases and Sections 87a et seq. German Copyright Act

[88] Art. 10 para. 3 EU Database Directive: "Any substantial change, evaluated qualitatively or quantitatively, to the contents of a database, including any substantial change resulting from the accumulation of successive additions, deletions or alterations, which would result in the database being considered to be a substantial new investment, evaluated qualitatively or quantitatively, shall qualify the database resulting from that investment for its own term of protection." Section 87a German Copyright Act: "A database whose content has been changed in a qualitatively or quantitatively substantial manner shall be deemed to be a new database insofar as the change requires a substantial qualitative or quantitative investment."

[89] See Sect. 5.1.

	Data Sale (Verkauf)	Data Trade (Tausch)	Data Donation (Schenkung)	Data as a Service (Data Rental) (Miete)	Data Lending (Leihe)
Contractual model	Perpetual transfer of data against one-off payment	Perpetual transfer of data without monetary consideration (e.g. data as consideration)	Perpetual transfer of data without consideration	Temporary transfer of data in return for a monthly fee; indefinite until notice of termination or specific contract term	Temporary transfer of data without any monetary or other form of consideration
Duration	Indefinite contract term	Indefinite contract term	Indefinite contract term	Specific contract term OR indefinite until notice of termination	Specific contract term OR indefinite until notice of termination
Usage rights	• Exclusive or non-exclusive • Geographically limited or unlimited • Restriction to agreed usage purposes	• Exclusive or non-exclusive • Geographically limited or unlimited • Restriction to agreed usage purposes	• Exclusive or non-exclusive • Geographically limited or unlimited • Restriction to agreed usage purposes	• Exclusive or non-exclusive • Geographically limited or unlimited • Restriction to agreed usage purposes	• Exclusive or non-exclusive • Geographically limited or unlimited • Restriction to agreed usage purposes
Licensing term	Perpetual/ unlimited in time	Perpetual/ unlimited in time	Perpetual/ unlimited in time	Temporary (specific contract term OR indefinite, but subject to notice of termination)	Temporary (specific contract term OR indefinite, but subject to notice of termination)
Sublicensing	• Sublicensable subject to contractual qualification, OR • Prohibition of sublicensing	• Sublicensable subject to contractual qualification, OR • Prohibition of sublicensing	• Sublicensable subject to contractual qualification, OR • Prohibition of sublicensing	• Sublicensable subject to contractual qualification, OR • Prohibition of sublicensing	• Sublicensable subject to contractual qualification, OR • Prohibition of sublicensing
Copying and distribution	• Copying, distribution and publishing of data or parts of data is prohibited, OR • Contractually permitted	• Copying, distribution and publishing of data or parts of data is prohibited, OR • Contractually permitted	• Copying, distribution and publishing of data or parts of data is prohibited, OR • Contractually permitted	• Copying, distribution and publishing of data or parts of data is prohibited, OR • Contractually permitted	• Copying, distribution and publishing of data or parts of data is prohibited, OR • Contractually permitted
Sui generis right of database maker	• Data Consumer's general prohibition to copy, distribute and publish significant parts of database • The same applies to repeated and systematic actions with regard to insignifcant parts of database • Parties are free to agree otherwise	• Data Consumer's general prohibition to copy, distribute and publish significant parts of database • The same applies to repeated and systematic actions with regard to insignifcant parts of database • Parties are free to agree otherwise	• Data Consumer's general prohibition to copy, distribute and publish significant parts of database • The same applies to repeated and systematic actions with regard to insignifcant parts of database • Parties are free to agree otherwise	• Data Consumer's general prohibition to copy, distribute and publish significant parts of database • The same applies to repeated and systematic actions with regard to insignifcant parts of database • Parties are free to agree otherwise	• Data Consumer's general prohibition to copy, distribute and publish significant parts of database • The same applies to repeated and systematic actions with regard to insignifcant parts of database • Parties are free to agree otherwise
Usage types	Known and unknown types of use depending on contractual restriction	Known and unknown types of use depending on contractual restriction	Known and unknown types of use depending on contractual restriction	Known and unknown types of use depending on contractual restriction	Known and unknown types of use depending on contractual restriction

Fig. 5.3 Contract matrix. © 2021, Dr. Alexander Duisberg, Bird & Bird LLP

That said, any parties wishing to use the connector technology and transact under the framework and reference architecture of IDS will need to include the reference to IDS and, in particular, recognize the requirements on certification.[90]

5.3 Implementing Compliance

Obviously, any participant in an IDS-based data exchange must be aware and ensure to take the appropriate measures to act in compliance with applicable laws, in particular with mandatory rules of data protection law (GDPR) and rules of competition law.

5.3.1 GDPR

The GDPR applies at any time where data providers share personal data, i.e., "any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person,"[91] triggering all the relevant provisions regarding a legal basis[92] and its limitations on purpose variation[93] and transparency through privacy notices,[94] ensuring data subjects rights,[95] documentation requirements,[96] and breach notifications.[97]

5.3.1.1 Controllers, Joint Controllers, and Processors

Data providers and data consumers will need to assess which type of relationship they will have, i.e., (1) whether the data consumer processes personal data as a new controller for its own purposes or (2) whether it is processing personal data together with the data controller for jointly defined purposes (i.e., acting as joint controllers). Where the data provider and the data consumer effect a data transfer to enable a new

[90] See Sect. 5.4 and Chap. 3 "Certification Process."
[91] See Sect. 5.1.8.
[92] Art. 6 GDPR
[93] Art. 6 para. 4 GDPR
[94] Art. 13, 14 GDPR
[95] Art. 15–21 GDPR
[96] Art. 30 GDPR
[97] Art. 33, 34 GDPR

processing purpose, the data provider must assess the legal basis for the transfer in accordance with Art. 6 or Art. 9 GDPR (the latter for special categories of personal data). While legitimate interest[98] is very often the suitable basis for transferring data for processing purposes that the data provider has predefined, this approach will not work where it is foreseeable that the data consumer wishes to process that personal data for arbitrary purposes which are unclear at the time of the transfer. Equally, it may be a significant challenge to rely on legitimate interest where a data provider shares personal data in a 1:*n* relation with a (possibly unknown) multitude of data consumers. Further, it is important to recognize that legitimate interest cannot serve as a legal basis when it comes to special categories of personal data (e.g., health data).[99] Accordingly, data providers and data consumers may also need to consider if and where they need to base data transfers and subsequent processing on the data subject's consent, and manage related consent management tools.

5.3.1.2 Documentation

Where the data provider and the data consumer(s) or various data consumers among each other pursue common purposes of data processing, they will need to enter into joint controller agreements under Art. 26 GDPR. Again, IDS can provide template documents that allow the parties involved to reduce their transaction costs in setting up and negotiating such agreements. Yet, the parties involved will need to at least define the substantive content (data categories concerned, recipients, processing purposes), whereas certain standard elements including the required technical and organizational measures[100] can (possibly) be incorporated by way of reference to the security standards provided by the platform operator.

The platform operator, by contrast, will normally act and position itself as the data processor who provides the technical facilities and, hence, effectively processes personal data on behalf of the various data controllers. Accordingly, the platform operator will provide (and the various data controllers can draw on) standard data processing agreements as required under Art. 28 GDPR, in which the parties involved (controllers and processors) will need to determine the data categories, data recipients, and processing purposes, as well as information on sub-processors and various other details, including the technical and organizational measures.

Further documentation requirements include that data controllers display the related privacy notices[101] and maintain proper records of processing activities.[102]

---

[98] Art. 6 para. 1 lit f GDPR

[99] Art. 9 GDPR does not provide such general legal basis, but only allows processing—without the data subject's consent—in limited circumstances, such as for research purposes in particular scenarios set forth under Art. 9 para. 2 lit. (j) GDPR in accordance with the national derogations set forth by the Member States.

[100] Art. 32 GDPR

[101] Art. 13, 14 GDPR

[102] Art. 30 GDPR

5.3.1.3 Breach Notifications

Where personal data breaches occur, the data controller(s) will need to assess the risk for the rights and freedoms of the data subjects and, if so, notify within 72 h the competent data protection authority and, in case of significant risks, also the data subjects.[103] In an environment of sharing personal data, this requires a clear allocation of responsibility and reporting back to the data controller. With the means of data processing agreements[104] and the usage control under IDS, each data controller should be well equipped to follow through with its notification obligations, i.e., preparing a report on which data has been affected by which incident, what consequences might arise from the breach, and which measures the controller has taken to mitigate the impact.

5.3.1.4 Enforcement and Sanctions

All stakeholders need to be aware of the significant level of fines that the GDPR attaches to non-compliance[105] and the increased enforcement actions taken by the data protection authorities.[106]

Obviously, the primary responsibility falls with the data controllers, but also data processors can be held liable, or even both, controllers and processors, can be held liable to pay damages to data subjects, jointly and severally.[107]

---

[103] Art. 33, 34 GDPR

[104] Art. 28 GDPR

[105] Up to 4% of the annual aggregate turnover of a data controller for certain breaches, such as failures regarding establishing the proper legal basis, ensuring data subjects' rights (Art. 83 para. 5 GDPR), and up to 2% for failures such as lacking proper documentation (Art. 83 para. 4 GDPR)

[106] The French data protection authority CNIL imposed a fine of EUR 50 million against Google Inc. for lack of transparency, inadequate information, and lack of valid consent regarding the ads personalization (https://edpb.europa.eu/news/national-news/2019/cnils-restricted-committee-imposes-financial-penalty-50-million-euros_en). The Italian data protection authority Garante imposed a fine of EUR 12,250,000 against Vodafone for unlawful processing of personal data of millions of users for aggressive telemarketing purposes (https://edpb.europa.eu/news/national-news/2020/aggressive-telemarketing-practices-vodafone-fined-over-12-million-euro_en). The Hamburg Commissioner for Data Protection and Freedom of Information imposed a EUR 35.3 million fine for data protection violations in H&M's Service Center (https://edpb.europa.eu/news/national-news/2020/hamburg-commissioner-fines-hm-353-million-euro-data-protection-violations_de). The Berlin Commissioner for Data Protection and Freedom of Information issued a fine of around EUR 14.5 million against Deutsche Wohnen SE for non-compliance with general data processing principles (https://edpb.europa.eu/news/national-news/2019/berlin-commissioner-data-protection-imposes-fine-real-estate-company_de). The State Commissioner for Data Protection in Lower Saxony has imposed a fine of EUR 10.4 million against notebooksbilliger.de AG. The company had been using video surveillance to monitor its employees for at least 2 years with no legal justification (https://edpb.europa.eu/news/national-news/2021/state-commissioner-data-protection-lower-saxony-imposes-eu-104-million-fine_de). Accessed 9 February 2021.

[107] Art. 82 paras. 4 and 5 GDPR

IDS itself does not take the role of a data controller or data processor. Accordingly each participant in the IDS ecosystem must be aware and take responsibility for its compliance with the GDPR—respectively assess in the first place if and to which extent it is willing and capable to process personal data in light of those requirements, or determine that its data contributions and data exchange shall exclude personal data from the outset.

5.3.2 Competition Law

One of the significant challenges that all participants to data sharing ecosystems need to be aware of and observe are the requirements of competition law, in regard to horizontal cooperations between competitors, "vertically" in downstream distribution models for data, as well as wherever the market position of a data provider (or data consumer) and the nature of the information could result in a distortion of markets and/or an abuse of a market dominant position.[108] Arguably, European competition law is only picking up with the challenges of the digital economy. The future EU Digital Market Act ("DMA") sets an important milestone in regulating platform operators that have the role of a "gatekeeper". The DMA will likely enter into force at the begin of 2023. While it is premature to assess the actual impact of this regulation, the sanctions (of up to 10% of a gatekeeper's aggregate annual revenues) give a strong message aiming at a fair data economy and preventing "data oligopolies". Platform operators exceeding a certain size (in terms of market valuation, numbers of users, etc.) will fall under the DMA.[109]

[108] See Art 101–109 Treat of the Functioning of the European Union TFEU; Art. 101 No. 1 TFEU: "...All agreements between undertakings, decisions by associations of undertakings and concerted practices which may affect trade between Member States and which have as their object or effect the prevention, restriction or distortion of competition within the internal market, and in particular those which: (a) directly or indirectly fix purchase or selling prices or any other trading conditions; (b) limit or control production, markets, technical development, or investment; (c) share markets or sources of supply; (d) apply dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage; (e) make the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts." Art. 102 TFEU: "Such abuse may, in particular, consist in: (a) directly or indirectly imposing unfair purchase or selling prices or other unfair trading conditions; (b) limiting production, markets or technical development to the prejudice of consumers; (c) applying dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage; (d) making the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts."

[109] A provider of core platform services shall be designated a gatekeeper if (a) it has a significant impact on the internal market (annual EEA turnover equal to or above EUR 7.5 billion in the last 3 financial years, or where the average market capitalization or the equivalent fair market value of the undertaking to which it belongs amounted to at least EUR 75 billion in the last financial year,

Beyond those generic principles it is clear that each participant in a data space must apply particular care in regard to the nature of information that it discloses and shares by way of a data exchange.[110] Accordingly, data providers and data consumers must take the necessary precautions to avoid that sensitive industrial information is disclosed which could allow entry into price ties, creating oligopolies, or induce coordinated behavior in breach of the applicable EU and national competition laws.[111]

5.4 Certifications from a Legal Perspective

While certifications are by nature a technical issue, they represent an important pillar for building trust in IDS. In that context, a few legal aspects play a significant role.

5.4.1 Role of Procedural Rules

IDS has not only created a certification standard,[112] but is presenting the same in conjunction with procedural rules of certification.[113] These procedural rules of certification are built on the procedural rules of the "Trusted Cloud" initiative of the Federal German government, which have been developed and published in a joint initiative of various stakeholders.[114] As such, they represent a well-developed,

---

and it provides a core platform service in at least three Member States); (b) it operates a core platform service which serves as an important gateway for business users to reach end users (more than 45 million monthly active end users established or located in the Union and more than 10 000 yearly active business users established in the Union in the last financial year); and (c) it enjoys an entrenched and durable position in its operations or it is foreseeable that it will enjoy such a position in the near future (Art. 3 Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act Proposal)).

[110] Consequently, unfair competition or anti-trust issues should also be assessed, such as the market relevance of the platform and whether certain mechanisms could, e.g., create a market barrier due to entry thresholds where certain organizations do not qualify or may be excluded.

[111] See Plattform Industrie 4.0 Result Paper Industrie 4.0—Implications for competition law: https://www.plattform-i40.de/PI40/Redaktion/EN/Downloads/Publikation/competition-law.pdf?__blob=publicationFile&v=8. Accessed 10 February 2021. As an addition regarding the EU Commission's guidance on private sector data sharing and the undertaking's access claim under competition law to a platform, see *Frenz*: EU-Digitalisierungsrecht. Datennutzung—Wettbewerb—Klimaschutz, EuR 2020, 210.

[112] See Chap. 3 "Certification Process" and https://internationaldataspaces.org/use/certification/. Accessed 10 February 2021.

[113] See https://internationaldataspaces.org/download/19008/. Accessed 15 February 2021.

[114] With courtesy and permission of the authors, see Rules of Procedure for Certification According to the Trusted Cloud Data Protection Profile for Cloud Services (TCDP) https://tcdp.de/data/pdf/15_Rules_of_Procedure_v1.0_EN.pdf. Accessed 10 February 2021.

reasonably balanced set of rules following common market standards to conduct certifications.

5.4.2 Additional Aspects

The issuance and management of the IDS certification standard can raise competition law aspects, if and when it develops to have a market-relevant impact. Accordingly, the IDS intends to enable further (commercially oriented) certification bodies to certify against the IDS standard in going forward at such point, whereas for the time being, the International Data Spaces Association (IDSA), acting as a non-profit organization on a mere cost basis) will currently act as the sole certification body.

Any entity seeking a certification ("applicant") will enter into a related agreement with the certification body (i.e., IDSA or in the future other certification bodies) to conduct the certification assessment in accordance with the procedural rules. The actual examination may be assigned to a separate examination body ("audit body"), which will act either as a sub-contractor of the certification body or, preferably in order to maintain organizational independence, through a separate contract with the applicant. As regards contractual liability, the certification body and the audit body will seek to exclude liability to the extent possible under German law (and related rules on standard contract terms, i.e., limiting liability for ordinary negligence to the "typically foreseeable damage"). In addition, such bodies will want and need to maintain a suitable general liability insurance.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 6
Tokenomics: Decentralized Incentivization in the Context of Data Spaces



Jan Jürjens, Simon Scheider, Furkan Yildirim, and Michael Henke

Abstract A significant challenge in bootstrapping a jointly used infrastructure such as Data Spaces is to incentivize the participants to invest in setting up the infrastructure. In this chapter, we investigate this challenge and possible solutions, focusing on an approach called "Tokenomics."

The incentivization scheme should be utilized by governance frameworks, in which the participants of Data Spaces remain capable of action and independent through automated, effective, and fair decision-making processes. Also, potential participants should be motivated to participate in the establishment and further development of the system, while on the other hand, undesirable behavior should be penalized. In combination with distributed ledger technology (DLT) and machine-readable, legally compliant smart contracts, participant behavior can be affected in such a way that both data quality and quantity are improved for the whole Data Space.

To derive possible design options for Tokenomics approaches, we examine different token frameworks and their impact on participants. The investigation of the frameworks is carried out taking into account five significant domains: technical, behavior, inherent value, coordination, and pseudo-archetypes. Furthermore, we investigate which token designs provide smaller or larger incentives in order to join or maintain a DLT-based ecosystem.

J. Jürjens (✉) · S. Scheider
Fraunhofer Institute for Software and Systems Engineering ISST, Dortmund, Germany
e-mail: jan.juerjens@isst.fraunhofer.de; simon.scheider@isst.fraunhofer.de

F. Yildirim
TU Dortmund University, Dortmund, Germany
e-mail: furkan.yildirim@tu-dortmund.de

M. Henke
Fraunhofer Institute for Material Flow and Logistics IML, Dortmund, Germany
e-mail: michael.henke@iml.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_6

6.1 Tokenomics in the Context of Data Spaces

The last years have been characterized by data to become a significant value driver. In this context, the creation of secure data sharing ecosystems that ensure both data governance and traceability is of particular importance. By providing such ecosystems, potentials for new applications and services are unleashed [1]. Industry 4.0 and the related business process automation require ever-increasing amounts of data. Organizations can therefore no longer rely solely on internal and publicly available data sources to remain competitive. Instead, they also need information from external individuals and organizations. So far, most ecosystems are controlled and driven by central actors [2]. Trust is one of the most important prerequisites for successful cooperation in networks, since the extensive exchange of data between partners also means that sensitive internal information is passed on. In practice, there is a lack of trust between the actors. This lack of trust can lead to information being withheld [3, 4]. Network effects also cause them to evolve into monopolistic or at least oligopolistic structures. This dilemma has triggered a debate on data sovereignty, especially in Europe. The concept of data sovereignty is intended to give rise to alliance-driven data ecosystems with associated platforms. For this reason, the global alliance, International Data Spaces Association, which has been driven by companies and research institutions such as Fraunhofer since 2015, set its goal to develop standardized architectures for such platforms and test them in practice [5, 6].

A significant challenge in bootstrapping a jointly used infrastructure such as Data Spaces is to incentivize the future participants of this infrastructure to invest funds or effort in setting up the infrastructure in the first place. This challenge is closely related to the well-known "tragedy of the commons" [7] and its historic remedy to create cooperatives [8]. It is also closely related to the topic of open-source economics [9].

In this chapter, we investigate this challenge and possible solutions, focusing on an approach called "Tokenomics" which has the goal to address this challenge using a token-based incentivization scheme [10, 11].

Distributed ledger technology (DLT) in general and blockchain technology (BCT) in particular can be utilized and are highly relevant to consider in situations where data sharing is pursued in a decentralized manner. BCT received attention in 2008 when an individual or group under the pseudonym *Satoshi Nakamoto* introduced the technology in a white paper that also introduced the decentralized cryptocurrency Bitcoin [12, 13]. An important feature of BCT is the possibility of using tokens that can act as incentives in a decentralized network. Incentive systems can thus be created for actors to join the network and act in the interests of the ecosystem [14].

Therefore, in the following two sections, we present two use cases where the use of blockchain-empowered tokens makes it possible to address the described challenges. For this purpose, we investigate different token classifications depending on the use cases. In the second section, we turn to supply chain management and describe two projects that use token deployment to create incentive systems. The

two projects are then compared in terms of similarities and differences. In the third section, we use a similar approach to examine the Personal Data Market (PDM) use case. For this reason, we present five different Tokenomics approaches and derive a suggestion for token archetypes of PDM token systems.

6.2 Token-Based Supply Chain Management

The use of distributed ledger technologies (DLTs) promises improvements of supply chain processes, especially with regard to the traceability of products along the entire value chain. DLT offers an ideal solution to reliably connect and manage IoT devices. In this context, supply chain management and logistics are two of the main application domains for linking blockchain technology (BCT) with new technologies such as Internet of Things (IoT) or artificial intelligence (AI) [15, 16]. DLT-based tokens play a key role in this context as they can be used to incentivize various actors in participating in such a value chain and establish a truly decentralized network. Consequently, we address the topics of DLT-based tokens for supply chain management in this chapter and present various token classifications. We then compare two blockchain projects in terms of token incentivization.

6.2.1 Supply Chain Traceability

Due to advancing globalization, traceability of goods in supply chains has gained importance. Also, consumers have a greater interest in consuming goods that meet certain sustainable and ethical standards [17, 18]. In this context, supply chains have become increasingly complex, complicating manufacturers' sourcing efforts [19]. In addition, regulations and international standardization are placing new demands on the supply chain management functions of companies. The European Parliament calls for food traceability and requires food suppliers as well as market participants to provide information on the origin of goods [20]. In order to trace the origin of a product, supply chain management ideally needs to be operated by multiple interconnected actors in the supply chain. However, traditional systems often miss cross-company interconnections and are operated in isolation—so-called data silos. Therefore, supply chain actors often are not able to trace relevant goods and receive information on their location or origin [21, 22]. In recent years, various methods have been developed to monitor processes and activities in the networked industry. The Industry 4.0 emerged and led to the digitization of processes, where supply chains provide real-time access to relevant products and production information for all stakeholders involved [23, 24]. The fourth industrial revolution is characterized by the emergence of various technologies, such as IoT and BCT, which conquer the boundaries between the physical and digital world and have a big potential in their

combination. Such technologies can transform modern supply chain networks into complete digital ecosystems. The digitization of supply chains offers outstanding business speed, agility, and the development of traceability mechanisms that enable near-complete identification and recording of products and processes. It is worth noting that blockchain-enabled supply chain approaches coupled with IoT applications could improve communication and traceability data delivery, enabling additional data management and analytics benefits for the logistics industry [25]. In supply chain management, payment processes can also be automated using smart contracts and, in combination with cyber-physical systems, increasingly autonomized. By eliminating manual activities, the processes involved are accelerated considerably [16, 26].

6.2.2 *Distributed Ledger Technology and Tokenomics*

Before blockchain technology was developed, many attempts were made to fulfill the desire of passing an asset digitally from one party to another via a secure, peer-to-peer Internet connection [27]. To address this need, the Bitcoin white paper "A Peer-to-Peer Electronic Cash System," published by *Satoshi Nakamoto* in 2008, proposed a new way for transferring money within a digital peer-to-peer network [13]. Until then, peer-to-peer currencies faced the problem that it was impossible to avoid a single unit being spent twice without a central intermediary [28]. When executing a transaction digitally without a central intermediary, one party could use a digital asset that has already been used in another transaction and, hence, spend it twice [29, 30]. If a central intermediary should be avoided in a decentralized network, the goal is to overcome those risks with individuals not relying on only one central authority but several ones. BCT is a specific approach for storing and ordering information of transactions without a central intermediary [31]. It serves as a decentralized data backup system that enables all members of a transaction to manage the uncovered information carefully and, at the same time, ensure its validity. Based on BCT, several peer-to-peer transfer technologies were developed and later combined under the term distributed ledger technology. Although DLT and BCT are often used interchangeably, BCT represents a specific type of DLT [27].

Apart from the public ledger envisioned by *Nakamoto*, blockchain solutions today can generally be classified into one of three categories: public, private, and consortium blockchains [32]. One feature within blockchain technology is of particular importance: tokens. Tokens can provide incentive design optimization to induce honest behavior among actors in a competitive environment [14]. For each type of goods managed in the supply chain, a smart contract with tokens is set up. As a type of blockchain recordkeeping, these tokens digitally replicate physical assets in distributed ledgers [33]. The process of tokenization transforms tangible and intangible assets into digitally encoded tokens [34]. The owner's digital record can be transferred between parties through the DLT network without a central authority [35]. There are two significant types of blockchain recordkeeping mechanisms:

currencies and tokens. A currency is usually native to a blockchain. Since BCT builds upon cryptography, such a currency is also called a cryptocurrency. The primary example is Bitcoin, which is the native currency of the Bitcoin blockchain [13]. A token is not native to a blockchain but is created on top of a blockchain and governed by a smart contract. Smart contracts govern most tokens following the common standard [36].

According to *Weingärtner*, the categorization of tokens into fixed classes is more difficult because any variation or functional enhancements can be programmed using smart contracts [37]. Freni et al. [38] have analyzed eight different token classification approaches from the literature to derive a morphological token classification framework. A useful distinction of tokens can be made by using the following five domains: The *technical* domain includes all technical properties of the token. All native functional properties are gathered in the *behavior* domain. The *inherent value* domain describes the economic value of the token. Furthermore, the *inherent value* domain investigates how the value of the token is created and influenced. The coordination of stakeholders depends on certain characteristics of the token, which are summarized in the *coordination* domain. The last domain, *pseudo-archetypes*, gathers all combined token frameworks. The derived 16 dimensions of the morphological token framework are summarized in dependence of the domains in Table 6.1.

With the use of tokens that represent the ownership of scarce digital resources, actors can be coordinated in a network. A proper design of a token is achieved by

Table 6.1 Morphological token classification framework [38]

	Domain			
	Technical	Token behavior	Inherent value	Coordination
Dimension	Accessibility *(Permissioned, Permissionless)*	Burnability *(Burnable, Non-burnable)*	Underlying Value *(Asset-backed, Network Value, . . .)*	Economic Value Driver *(Demand, Supply,. . .)*
	Chain *(New Chain, Forked Chain,. . .)*	Expirability *(Expirable, Non-expirable)*	Value generated by *(Effort of others, Effort of holders)*	Role *(Store of Value, Voting right, Value Exchange,. . .)*
	Layer *(Blockchain, Protocol Application)*	Spendability *(Spendable, Non-spendable)*		Monetary Policy *(Schedule-based, Pre-mined, . . .)*
	Number of Blockchains *(Single Chain, Cross Chain)*	Token Type *(Fungible, Non-fungible, Hybrid)*		
	Representation Type *(Common, Unique)*	Token Unit *(Fractional, Whole, Singleton)*		
		Tradability *(Tradable, Non-tradable, Delegable)*		

aligning the different types of incentives within a token-built ecosystem. The token can then be used to orchestrate the creation and control the development of these protocols. The use of tokens can accelerate and improve the development of an incentive-based decentralized network. These incentives bring different stakeholders' interests to a common denominator. Thus, token-based networks can strengthen competitive collaboration in the long run in a digital age [11, 39, 40].

Hülsemann and Tumasjan [12] simulated the effects of different designs. The designs of cryptocurrencies, network tokens, and investment tokens were examined in the context of prediction markets. The research concluded that network tokens provide the largest incentives for actors to join and remain in the network ecosystem over the long term. For example, network tokens can provide services within a system. In addition, the authors found that investment tokens provide the smallest incentives [12].

6.2.3 DLT-Based Supply Chain Traceability

To strengthen supply chain management, various blockchain projects have been initiated that also aim to create incentives through the use of tokens. In this chapter, we briefly summarize two well-known projects and present the differences and similarities in the token design.

VeChainThor was developed as a public blockchain to simplify supply chain management. Originally, it was developed to determine whether a real product is a fake or not, so that fraud and counterfeiting can be eliminated. In the meantime, the blockchain is used by large companies for supply chain traceability. VeChainThor designed a bi-token system that includes VeChain Token (VET) and VeThor Token (VTHO). The function of VET is to serve as a value-transfer medium to enable rapid value circulation within the VeChainThor network. VTHO represents the underlying cost of using VeChainThor and is consumed, or in other words, burned by performing on-chain operations. Furthermore, VTHO is generated from holding VET, which is established to allow any user with VET to make transactions at no extra cost if the user holds the VET tokens for long enough. The goal of this token design is to prevent transaction fees from being directly exposed to the volatility of VET's price, making the VeChainThor blockchain more suitable for conducting business or financial activities for both individuals and companies. The demand for VTHO arises from the execution of smart contracts and payment transactions. To stabilize transaction costs and maintain the balance of supply and demand for VTHO, the Foundation closely monitors the market and estimates the demand for VTHO based on the activity of applications running on the VeChainThor blockchain and token transfers [41]. The token classification of VET and VTHO is summarized in Table 6.2.

Waltonchain is a supply chain solutions platform built on the blockchain. It offers decentralized product tracking that tracks QR and RFID codes on the blockchain. Waltonchain (WTC) is a cross-chain ecosystem where the parent chain and child

Table 6.2 Token classification VeChainThor

	Domain			
	Technical	Token behavior	Inherent value	Coordination
Dimension	Accessibility: *Permissionless*	Burnability: *Burnable (VTHO), Non-burnable (VET)*	Underlying Value: *Network Value*	Economic Value Driver: *Demand*
	New Chain, New Code	*Non-expirable*	Value generated by: *Effort of Holders*	Role: Value Exchange, Reward Potential (VET), Access to Service (VTHO)
	Layer: Blockchain	*Spendable*		Monetary Policy: Pre-mined *Schedule-based*
	Single Chain	*Fungible*		
	RepresentationType: *Common*	Token Unit: *Fractional*		
		Tradable		

chains serve as the framework. In this cross-chain ecosystem, data circulation and value transfer can be realized between child chains. Using the cross-chain mechanism, child chain tokens are exchanged for the WTC token and can be further exchanged for other child chain tokens; thus, value circulates on the blockchain [42]. The Waltonchain ecosystem is currently under steady development and in order to make WTC more stable in the cross-chain IoT ecosystem, Waltonchain Foundation issues the Waltonchain Autonomy token (WTA) on the Waltonchain mainnet. To incentivize the community members, the Foundation launches tasks which can be completed by the community. By completing tasks the community members can earn points, which can be converted to WTA at a certain ratio. Furthermore, the WTA token can be used to deduct service fees during the token exchange between WTC and child-chain tokens [43]. The token classification of WTC and WTA is summarized in Table 6.3.

The two blockchain projects investigated, VeChainThor and Waltonchain, show similar patterns in token design. Both projects use two different tokens each in an attempt to get actors to join the ecosystem and act in the interests of the network. Both projects use a second token (VTHO and WTA) to maintain stability in the system and avoid the volatility of market behavior. However, the difference between the designs lies in the incentivization. While the VTHO token is designed to eliminate the cost of transactions within the blockchain, the WTA token aims to ensure that actors act in the spirit of the ecosystem. Moreover, the introduction of these two tokens induces actors to hold VET and WTC tokens in order to receive VTHO or WTA. In this context, the VET and WTC tokens constitute the role of value exchange, while WTC can be used for cross-chain transactions. These

Table 6.3 Token classification Waltonchain

	Domain			
	Technical	Token behavior	Inherent value	Coordination
Dimension	Accessibility: *Permissionless*	Burnability: *Non-burnable*	Underlying Value: *Network Value*	Economic Value Driver: *Demand*
	New Chain, New Code	*Non-expirable*	Value generated by: *Effort of Holders*	Role: Value Exchange, Reward Potential (Staking)
	Layer: Blockchain	*Spendable*		Monetary Policy: Pre-mined, Schedule Distribution
	Cross Chain	*Fungible*		
	Representation Type: *Common*	Token Unit: *Fractional*		
		Tradable		

exemplary findings should be enriched by empirical data in future studies so that further token designs can be evaluated in terms of incentivization.

As the value creation process more and more often takes place in enterprise networks—rather than in individual enterprises—also the combination and enrichment of data takes place by various actors in data spaces. For this reason, we investigate the use of Tokenomics in the context of Data Spaces in the next section.

6.3 Tokenomics in the Context of Personal Data Markets

Tokenomics is commonly applied in the framework of Personal Data Markets. In order to shed light on the interplay of these related concepts, we investigate the motivational factors of market operators for using a Tokenomics approach when building Personal Data Markets (PDMs). Furthermore, we classify token design patterns with regard to PDMs and examine whether there are "specific token designs" commonly applied for Personal Data Market infrastructures.

6.3.1 Personal Data Markets

Nowadays, individuals create tremendous amounts of valuable data [44] while interacting in their lives using devices equipped with sensors such as mobile phones, tablets, smart home systems, or computers. Even though this personal data belongs to the individuals, data producers commonly give up all rights concerning their data by agreeing to terms and conditions, which are required of them before using certain

services [45]. Thus, service providers can easily get access to such valuable personal data. Subsequently, they might use this personal data for further product and service development or even sell it to various third parties making substantial profits [46, 47]. The individuals from whom data were collected are left without any control or profit. Against this background, the concept of Personal Data Markets (PDM) has been raised in numerous discussions during the past years as an appropriate solution medium. PDMs are believed to lead to a fair(er) data economy and, hence, foster innovation in all areas of application through efficient and large-scale data sharing between private data owners and organizations of all kinds [48]. Consequently, the expectation is that they can contribute to societal benefit. Despite the potential benefit PDMs promise for the entire society, many marketplace providers are frequently forced to their knees as a variety of challenges for PDMs exist [49]. These challenges originate mainly from technological, legal, economic, or ethical domains [49], resulting in short life cycles of many marketplace providers, e.g., Data Fairplay, Datareum, Datatrade, Synapse AI, and MYBS. However, there are PDMs which seem to operate successfully as they have existed for some years now. When analyzing a sample of such "resistant" PDMs, it becomes clear that token systems play a major role in order to master the economic challenge of incentivizing market participants, especially the owners of personal data, for sharing their data.

6.3.2 Motivational Factors for Tokenomics Approach in Personal Data Markets

In the context of PDMs, tokens can be defined as units of value the market operator creates and emits in order to self-govern its business model, and to empower all participants of the marketplace to interact with each other, while facilitating the distribution of rewards and the sharing of benefits to all stakeholders. Thereby, the concept of Tokenomics has arisen from game theory, mechanism design, and monetary economics [50]. It is vital for a general understanding about the interplay between the concepts of Tokenomics and PDMs to examine the motivational factors convincing market operators for the application of token systems in the framework of their technical marketplace architectures. For this purpose, existing PDMs were analyzed giving special attention to technical architecture whitepapers of marketplace operators. The results of the sample examined are summarized below.

Airbloc relies on two kinds of tokens with distinct purposes. Firstly, this market operator offers the token ABL which is a tradable ERC20 token used as a means of participating in the network in order to settle payments for data exchange as well as staking register and maintaining certain services on nodes. Thus, ABL serves for payment, settlement, and participation. Secondly, Airbloc runs the virtual and non-tradable AIR token, primarily used for providing rewards to network participants. AIR tokens are one-way convertible into ABL only. Hence, AIR supports

productive behavior within the network by assessing participant's reputation and contribution.

The *Datum Network* is built upon the DAT token as utility token in order to facilitate transactions, especially data sharing, among data owners and buyers by providing the medium for payment settlements. It is a smart token enabling users to buy and sell stored data while enforcing data usage policies set by the data owner. The latter is controlled by underlying smart contracts running on blockchain. Furthermore, the token grants access to certain privileges in the network such as data storage and participation in the data market.

Madana uses tokens based on the Lisk Blockchain in order to support consistency, transparency, and co-determination as the PAX token holder has the opportunity to vote for a global data model ensuring consistency across the entire ecosystem. In doing so, the token holder determines the way data must be offered in order to participate. Furthermore, the token system serves to handle rewards for contributions of participants to the Madana platform in terms of data or services functioning as a payment vehicle. Thus, Tokenomics incentivizes provision of analytical services and data sharing on the platform.

In the *OSA DC* network, participants receive tokens for offering services such as collecting, cleaning, and enriching data and offering data storage or analytical services. Furthermore, the OSA tokens function as a payment medium, buying and selling transactions among participants. Additionally, the token system is applied to provide a reward system where data providers receive token rewards for each action in the ecosystem and data consumers may also be rewarded for purchasing specific products and services. Thus, OSA assigns three main purposes to its OSA token system: facilitation of service offerings, payment medium in the ecosystem, and incentivization to contribute to the network.

VLD tokens are functional utility smart contracts within the *VETRI* platform where users are remunerated in VLD tokens based on the desirability of their data shared perceived by data consumers. Thereby, the tokens enable to buy and sell transactions on the marketplace. Furthermore, the tokens function as payment and settlement medium for secure data storage and services platform participants can purchase. Above all, Vetri's token system aims to provide an incentivization mechanism for all stakeholders and to support price stability.

Based on our findings summarized in Table 6.4, we derive the following main motivational factors for the application of token systems within Personal Data Market infrastructures: payment medium, incentivization mechanism, access and usage control, and facilitation of transactions on the marketplace (including services).

6.3.3 Token Design Principles for Personal Data Markets

With the insights gained from the analysis of motivational factors for applying Tokenomics in PDMs, the following section introduces common design patterns

Table 6.4 Motivational factors for Tokenomics approach in existing Personal Data Markets

Motivational factors	Token	PDM
ABL • Payment medium and settlement • Access to and participation in network *AIR* • Assessment of productive behavior and contribution to network	ABL AIR	Airbloc
• Facilitation of transactions • Payment medium • Enforcement of data usage rights • Access to privileges in network	DAT	Datum Network
• Support of consistency, transparency, and co-determination • Payment medium • Incentivization for contributions to network	PAX	Madana
• Facilitation of service offerings • Payment medium • Incentivization for contributions to network	OSA	OSA DC
• Access to data storage and services • Payment medium for data • Facilitation of transactions • Incentivization and compensation of all participants	VLD	Vetri

of tokens in literature [36]. Subsequently, we try to highlight overarching token design principles chosen by PDMs while relying on the taxonomy of token classification of Oliveira et al. [36]. In this taxonomy we suggest common token design principles, typically applied in Personal Data Markets based on our previously analyzed PDM sample.

The first token attribute is the *Class* of tokens which is commonly used for the distinction of tokens as it differentiates cryptocurrencies (digital money), digital shares including entitlements for profit sharing (tokenized security), and remaining crypto-assets based on tokens with attached utility (utility tokens) [36]. According to our sample, utility tokens are the dominant design of *Class* in PDMs. The *Purpose* of tokens distinguishes between tokens represented uniquely as an asset (asset-backed), tokens combined with an access permission (usage token), and tokens storing value to reward or incentivize behavior (work token) [51, 52]. In our sample, most PDMs rely on the design characteristics of usage token and work token as they are applied, for instance, as a right to access the marketplace or the entire platform as well as for incentivization of a certain behavior. Mougayar [53] classified tokens into *Roles*, representing a right for the owner (right), a unit of value exchange within a system (value exchange), a fee for access (toll), a tool to enrich user experience (function), a de facto payment method (currency), or a right for the data owner to receive a share of the profit (earnings). In our sample, the main *Roles* appeared to be rights, mediums for value exchange, access, and currency. Glatz [54] classified tokens according to their *Representation*. According to the author, they can be designed as pure digital assets (digital), bound to physical objects (physical), tied to objects from virtual reality (virtual), or state legal permissions and rights granted by law or the network

(legal). The major *Representations* of token designs in PDMs we examined were digital assets. Furthermore, they were commonly combined with rights granted by the network. Chen [35] divided tokens into their *Supply* which can either be fixed and distributed once (fixed) or accorded to a certain schedule (schedule-based). Most tokens in PDMs are offered once and subsequently burned over time, meaning that supply is fixed. Another attribute refers to the behavior tokens aim to incentivize. Lena and Oxana [55] called this attribute *Incentive System* where the token design can incentivize to enter, to use, or to stay long term in a system or on a platform. Incentivization of behavior plays a key role in PDMs when designing tokens. However, according to the whitepapers we analyzed, the incentivization of usage is the most dominant design principle in terms of *Incentive System* applied by market operators. Lena and Oxana [55] also suggested the attribute *Transactions* distinguishing spendable and non-spendable tokens. In PDMs, tokens are usually spendable on a platform, e.g., when reception or execution of payment transactions is concerned. Yadav [56] considered *Ownership* as an attribute since a token can either be tradable or non-tradable. In PDMs, the former design principle is usually the case. Similar to *Supply*, Oliveira et al. [36] defined the attribute *Burnability* reflecting the possibility for purposely burning tokens in order to create artificial scarcity or to express the extinction of access rights bound to the token. According to the sample analyzed, burning tokens over time is a commonly applied technique in PDMs. Just like *Ownership*, Glatz [54] defined *Fungibility* of tokens which addresses either purely equal (fungible) tokens or non-fungible ones due to their distinct characteristics ensuring their uniqueness. The dominant design principle of PDMs in this regard are fungible tokens. Furthermore, the attribute *Layer* refers to the distinction based on the location of tokens. Thereby, Little [51] differentiates between tokens native to blockchain, issued on top of a protocol or placed on the application layer [51]. The PDMs of our sample under study mainly run on a blockchain as the first layer carrying a token system as a second layer. Finally, token design is also affected by the *Chain* the system relies on. Srinivasan [57] differentiated the design pattern of new chains on new code, new chains on forked code, forked chains on forked code, or issued on top of a protocol. We state that the latter appears as the dominant design principle for PDMs relying on the information from our whitepapers.

6.3.4 Derivation of Token Archetypes for PDMs

Oliveira et al. [36] defined eight token archetypes depending on their specific purposes where several tokens exhibit a multipurpose ability, that is to say they serve more than one purpose simultaneously (see Table 6.5). The archetypes are cryptocurrency token, equity token, funding token, consensus token, work token, voting token, asset token, and payment token. According to our previously defined motivational factors for the application of tokens in PDMs and some frequent PDM

Table 6.5 Taxonomy of tokens design characteristics from Oliveira et al. [36] applied to PDMs

Design Attribute	Design Characteristics					
Class	Cryptocurrency		*Utility Tokens*		Tokenized Security	
Purpose	Asset-backed Tokens		*Usage Tokens*		*Work Tokens*	
Role	*Right*	*Value Exchange*	*Toll*	Function	*Currency*	Earnings
Representation	*Digital*	Physical	Virtual	*Legal*		
Supply	*Fixed*		Schedule-based			
Incentive System	Enter Platform		*Use Platform*		Stay-long term	
Transactions	*Spendable*		Non-Spendable			
Ownership	*Tradable*		Non-Tradable			
Burnability	*Burnable*		Non-burnable			
Fungibility	*Fungible*		Non-Fungible			
Layer	Blockchain (Native)		*Product (Non-Native)*		Application (dApp)	
Chain	New chain, new code	New chain, forked code	Forked chain, forked code	*Issued on top of protocol*		

The highlighted characteristics of tokens in PDMs are based on our previous sample only

token designs analyzed, we suggest the following main corresponding archetypes for token design in PDMs.

As we assigned the archetype with the best thematically fit to the derived motivational factors, further adjustments to our work are possible and recommended. Furthermore, we emphasize that both Tables 6.2 and 6.6 present suggestions we give based on findings and assumptions derived from our analysis of PDMs in practice. We encourage future research to empirically examine design patterns of token systems in PDMs and to define more justified archetypes in order to extend the still limited (design) knowledge in research about the mutual relation of the domains Tokenomics and Personal Data Markets.

Table 6.6 Suggestions of token archetypes of PDM token systems relying on the archetypes defined by Oliveira et al. [36]

Motivational factor	Archetype	Main purpose	Token description
Payment medium	Payment token	Payment	Token used as internal payment method in the system/platform
Incentivization mechanism	Work token	Work reward	Token used as reward for users completing actions or exhibiting certain behavior
Access and usage control	Asset token	Asset ownership	Token representing asset ownership
Transaction and service facilitation	Consensus token	Validation reward, store of wealth	Token used as reward to nodes/participants ensuring certain services such as data validation or consensus

6.4 Conclusions

A significant challenge in bootstrapping a jointly used infrastructure such as Data Spaces is to incentivize the future participants of this infrastructure to invest funds or effort in setting up the infrastructure in the first place. In this chapter, we investigated this challenge and possible solutions, focusing on an approach called "Tokenomics" which has the goal to address this challenge using a token-based incentivization scheme.

To derive possible design options for Tokenomics approaches, we examined different token frameworks and their impact on participants. Furthermore, we investigated which token designs provide smaller or larger incentives in order to join or maintain a DLT-based ecosystem. This investigation was done in the context of two use cases where the use of blockchain-empowered tokens makes it possible to address the described challenges: As part of supply chain management we discussed two projects that use token deployment to create incentive systems. The two projects were then compared in terms of similarities and differences. In the last part of this chapter, we used a similar approach to examine the Personal Data Market (PDM) use case. Toward this, we presented five different Tokenomics approaches and derive a suggestion for token archetypes of PDM token systems.

Based on these investigations we can conclude that the Tokenomics approach appears suitable for use in addressing the challenge of incentivizing the future participants of this infrastructure to invest funds or effort in setting up the infrastructure. We therefore recommend that this approach be further elaborated at a technical level to be used for this purpose.

References

1. Munoz-Arcentales, A., López-Pernas, S., Pozo, A., Alonso, Á., Salvachúa, J., & Huecas, G. (2020). Data usage and access control in industrial data spaces: Implementation using FIWARE. *Sustainability, 12*(9), 3885. https://doi.org/10.3390/su12093885
2. Jarke, M., Otto, B., & Ram, S. (2019). Data sovereignty and data space ecosystems. *Business & Information Systems Engineering, 61*(5), 549–550. https://doi.org/10.1007/s12599-019-00614-2
3. Eberl, P., & Kabst, R. (2006). Vertrauen, Opportunismus und Kontrolle-Eine empirische Analyse von Joint Venture-Beziehungen vor dem Hintergrund der Transaktionskostentheorie. In J. Sydow (Ed.), *Management von Netzwerkorganisationen. Beiträge aus der "Managementforschung". 4., aktualisierte und erw. Aufl* (pp. 107–142). Gabler; VS-Verl.
4. Milberg, J. (Ed.). (2002). *Erfolg in Netzwerken* (1st ed.). Springer.
5. Auer, S., Jürjens, J., Otto, B., Brost, G., Lange, C., Quix, C., Cirullies, J., Lohmann, S., Schon, J., Eitel, A., Mader, C., Schulz, D., Ernst, T., Menz, N., Schütte, J., Haas, C., Nagel, R., Spiekermann, M., Huber, M., … Pullmann, J. (2017). *Reference architecture model for the industrial data space*. White Paper. Fraunhofer.
6. Otto, B., Jürjens, J., Schon, J., Auer, S., Menz, N., Wenzel, S., & Cirullies, J. (2016). *Industrial data space: Digital sovereignty over data*. White Paper, Fraunhofer.
7. Hardin, G. (1968). The tragedy of the commons. *Science, 162*(3859), 1243–1248.
8. Moulin, H. (1995). *Cooperative microeconomics: A game-theoretic introduction*. Princeton University Press.
9. Lerner, J., & Tirole, J. (2005). The economics of technology sharing: Open source and beyond. *Journal of Economic Perspectives, 19*(2), 99–120.
10. Heise, C., Poddey, A., & Jürjens, J. (2020). *Strong long-term incentive design in GAIA-X based on tokenomics*. Position Paper—GAIA-X TaskForce Tokenomics.
11. Lamberty, R., de Waard, D., & Poddey, A. (2020). Leading digital socio-economy to efficiency—A primer on tokenomics. https://arxiv.org/abs/2008.02538
12. Hülsemann, P., & Tumasjan, A. (2019). Walk this way! Incentive structures of different token designs for blockchain-based applications. Available online at https://core.ac.uk/download/pdf/301384295.pdf.
13. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. In *Manubot*. Available online at https://bitcoin.org/bitcoin.pdf. Checked on 26 Jan 2021.
14. Sussan, F., & Acs, Z. J. (2017). The digital entrepreneurial ecosystem. *Small Business Economics, 49*(1), 55–73. https://doi.org/10.1007/s11187-017-9867-5
15. Atlam, H. F., Azad, M. A., Alzahrani, A. G., & Wills, G. (2020). A review of blockchain in internet of things and AI. *BDCC, 4*(4), 28. https://doi.org/10.3390/bdcc4040028
16. Bitkom. (2019). *Online-Konsultation zur Erarbeitung der Blockchain-Strategie der Bundesregierung*. Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. Available online at https://www.bitkom.org/sites/default/files/2019-04/190408_stellungnahme_blockchain-strategie-konsultation_online.pdf. Checked on 2/5/2021.
17. Dabbene, F., Gay, P., & Tortia, C. (2014). Traceability issues in food supply chain management: A review. *Biosystems Engineering, 120*, 65–80. https://doi.org/10.1016/j.biosystemseng.2013.09.006
18. Westerkamp, M., Victor, F., & Kupper, A. (2018). Blockchain-based supply chain traceability: Token recipes model manufacturing processes. In: *2018 IEEE international conference on internet of things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData)*. IEEE.
19. Gualandris, J., Klassen, R. D., Vachon, S., & Kalchschmidt, M. (2015). Sustainable evaluation and verification in supply chains: Aligning and leveraging accountability to stakeholders. *Journal of Operations Management, 38*(1), 1–13. https://doi.org/10.1016/j.jom.2015.06.002

20. European Parliament. (2002). EU, 2002. Regulation (EC) No. 178/2002 of the European Parliament and of the Council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety. *Official Journal of the European Communities, L31*, 1–24.
21. Appelhanz, S., Osburg, V.-S., Toporowski, W., & Schumann, M. (2016). Traceability system for capturing, processing and providing consumer-relevant information about wood products: system solution and its economic feasibility. *Journal of Cleaner Production, 110*, 132–148. https://doi.org/10.1016/j.jclepro.2015.02.034
22. Pal, K., & Yasar, A.-U.-H. (2020). Internet of things and blockchain technology in apparel manufacturing supply chain data management. *Procedia Computer Science, 170*, 450–457. https://doi.org/10.1016/j.procs.2020.03.088
23. Brettel, M., Friederichsen, N., Keller, M., & Rosenberg, M. (2014). *How Virtualization, decentralization and network building change the manufacturing landscape: An industry 4.0 perspective*. https://doi.org/10.5281/ZENODO.1336426.
24. Onufrey, K., & Bergek, A. (2020). Second wind for exploitation: Pursuing high degrees of product and process innovativeness in mature industries. *Technovation, 89*, 102068. https://doi.org/10.1016/j.technovation.2019.02.004
25. Banafa, A. (2017). IoT and blockchain convergence: benefits and challenges. Available online at http://iot.ieee.org/newsletter/january-2017/iot-and-blockchain-convergence-benefits-and-challenges.html.
26. Jakob, S., Schulte, A. T., Sparer, D., Koller, R., & Henke, M. (2018). *Blockchain und Smart Contracts: Effiziente und sichere Wertschoepfungsnetzwerke*. With assistance of Michael ten Hompel, Michael Henke, Uwe Clausen.
27. Nofer, M., Gomber, P., Hinz, O., & Schiereck, D. (2017). Blockchain. *Business & Information Systems Engineering, 59*(3), 183–187. https://doi.org/10.1007/s12599-017-0467-3
28. Pilkington, M. (2016). Blockchain technology: principles and applications. In *Research handbook on digital transformations*. Edward Elgar.
29. Kuo, T.-T., Kim, H.-E., & Ohno-Machado, L. (2017). Blockchain distributed ledger technologies for biomedical and health care applications. *Journal of the American Medical Informatics Association : JAMIA, 24*(6), 1211–1220. https://doi.org/10.1093/jamia/ocx068
30. Mazonka, O., Tsoutsos, N. G., & Maniatakos, M. (2016). Cryptoleq: A heterogeneous abstract machine for encrypted and unencrypted computation. *IEEE Transactions on Information Forensics and Security, 11*(9), 2123–2138. https://doi.org/10.1109/tifs.2016.2569062
31. Lumineau, F., Wang, W., & Schilke, O. (2020). Blockchain governance—A new way of organizing collaborations? *Organization Science*. https://doi.org/10.1287/orsc.2020.1379
32. Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017). An overview of blockchain technology: architecture, consensus, and future trends. In *2017 IEEE international congress on big data (bigdata congress)*. IEEE.
33. Lemieux, V., Hofman, D., Batista, D., & Joo, A. (2019). Blockchain technology & recordkeeping. Available online at http://armaedfoundation.org/wp-content/uploads/2019/06/aief-research-paper-blockchain-technology-recordkeeping.pdf.
34. Francisco, K., & Swanson, D. (2018). The supply chain has no clothes: Technology adoption of blockchain for supply chain transparency. *Logistics, 2*(1), 2. https://doi.org/10.3390/logistics2010002
35. Chen, Y. (2018). Blockchain tokens and the potential democratization of entrepreneurship and innovation. *Business Horizons, 61*(4), 567–575. https://doi.org/10.1016/j.bushor.2018.03.006
36. Oliveira, L., Zavolokina, L., Bauer, I., & Schwabe, G. (2018, December 12–16). To token or not to token: Tools for understanding blockchain tokens. In. *International conference of information systems (ICIS 2018). San Francisco, USA*. University of Zurich. Available online at https://www.zora.uzh.ch/id/eprint/157908/.
37. Weingärtner, T. (2019). Tokenization of physical assets and the impact of IoT and AI. Available online at https://blockchain.pwias.ubc.ca/sites/blockchain.pwias.ubc.ca/files/report-files/weingaertner_tokenization_iot_ai%20(1).pdf.

38. Freni, P., Ferro, E., & Moncada, R. (2020). Tokenization and blockchain tokens classification: A morphological framework. In: *2020 IEEE symposium on computers and communications (ISCC)*. IEEE.
39. Henke, M. (2002). *Strategische Kooperationen im Mittelstand. Potentiale des Coopetition-Konzeptes* \NLN\. Zugl.: München, Techn. Univ., Diss., u.d.T.
40. Henke, M. (2003). *Strategische Kooperationen kleinerer und mittlerer Unternehmen (KMU) unter besonderer Berücksichtigung des Coopetitions-Ansatzes*. Wissenschaft & Praxis (Schriftenreihe managementorientierte Betriebswirtschaft, 4).
41. VeChain Foundation. (2019). VeChain Whitepaper 2.0. Available online at http://www.vechain.org/qfy-content/uploads/2020/01/VeChainWhitepaper_2.0_en.pdf. Checked on 26 Jan 2021.
42. Waltonchain. (2018). White Paper V2.0. Available online at http://www.waltonchain.org/en/Uploads/2020-06-11/5ee1c6b5b10cf.pdf. Updated on 26 Jan 2021.
43. Waltonchain. (2019). Waltonchain Global Community Autonomy Blue Paper. Available online at http://www.waltonchain.org/en/sys/list/52.html. Updated on 26 Jan 2021.
44. Lupton, D. (2018). How do data come to matter? Living and be coming with personal data. *Big Data & Society, 5*(2), 2053951718786314.
45. Metzger, A. (2020). *A market model for personal data: State of play under the new directive on digital content and digital services*. Data as Counter-Performance–Contract Law 2.0.
46. Conger, S., Pratt, J. H., & Loch, K. D. (2013). Personal information privacy and emerging technologies. *Information Systems Journal, 23*(5), 401–417.
47. Malgieri, G., & Custers, B. (2018). Pricing privacy–the right to know the value of your personal data. *Computer Law & Security Review, 34*(2), 289–303.
48. Spiekermann, S., & Korunovska, J. (2017). Towards a value theory for personal data. *Journal of Information Technology, 32*(1), 62–84.
49. Spiekermann, S., Acquisti, A., Böhme, R., & Hui, K. L. (2015). The challenges of personal data markets and privacy. *Electronic Markets, 25*(2), 161–167.
50. Au, S., & Power, T. (2018). *Tokenomics: The crypto shift of blockchains, ICOs, and tokens*. Packt.
51. Little, W. (2017). A primer on blockchains, protocols, and token sales. *Hackernoon (blog)*. Available at: https://hackernoon.com/a-primer-on-blockchains-protocols-and-token-sales-9ebe117b5759. Accessed 22 Dec 2020.
52. Tomaino, N. (2017). Tokens, tokens and more tokens. Available at: https://thecontrol.co/tokens-tokens-and-moretokens-d4b177fbb443. Accessed 21 Dec 2020.
53. Mougayar, W. (2017). Tokenomics—A business guide to token usage, utility and value. Available at: https://medium.com/@wmougayar/tokenomics-a-business-guide-to-token-usage-utility-and-value-b19242053416. Accessed 22 Dec 2020.
54. Glatz, F. (2016). A blockchain token taxonomy. Available at: https://medium.com/@heckerhut/a-blockchain-token-taxonomy-fadf5c56139a. Accessed 21 Dec 2020.
55. Lena and Oxana. (2017). What are you token about? Blockchain token economics and rights. Available at: https://hackernoon.com/token-economy-4a38ad02a239. Accessed 23 Dec 2020.
56. Yadav, M. (2017). Exploring signals for investing in an initial coin offering (ICO). Available at SSRN 3037106. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3037106.
57. Srinivasan, B. (2017). Thoughts on tokens. news. 21. co, 27. Available at: https://news.earn.com/thoughts-on-tokens-436109aabcbe. Accessed 21 Dec 2020.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part II
Data Space Technologies

Chapter 7
The IDS Information Model: A Semantic Vocabulary for Sovereign Data Exchange



Christoph Lange, Jörg Langkau, and Sebastian Bader

Abstract The Information Model of the International Data Spaces (IDS-IM) is the central integration enabler for the semantic interoperability in any IDS ecosystem. It contains the terms and relationships to describe the IDS components, their interactions, and conditions under which data exchange and usage is possible. It thus presents the common denominator for the IDS and the foundation for any IDS communication. As such, its evolution cycles are deeply related with the maturity process of the IDS itself. This chapter makes the following contributions related to the IDS Information Model: a brief overview of the vocabulary, its guiding principles, and general features is supplied. Based on these explanations, several upcoming aspects are discussed that reflect the latest state of discussions about the declaration and cryptographic assurance of identities and decentralized identifiers, and how these need to be treated to ensure compliance with the IDS principles.

In addition, we explain the latest developments around the IDS Usage Contract Language, the module of the IDS-IM that expresses Usage Contracts, and data restrictions. These definitions are further implemented with infrastructure components, in particular the presented, newly specified Policy Information Point and the Participant Information Service of the IDS.

C. Lange (✉)
Fraunhofer FIT, Sankt Augustin, Germany
e-mail: christoph.lange-bever@fit.fraunhofer.de

J. Langkau
nicos AG, Münster, Germany
e-mail: jlangkau@nicos-ag.com

S. Bader
SAP SE, Walldorf, Germany
e-mail: s.bader@sap.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_7

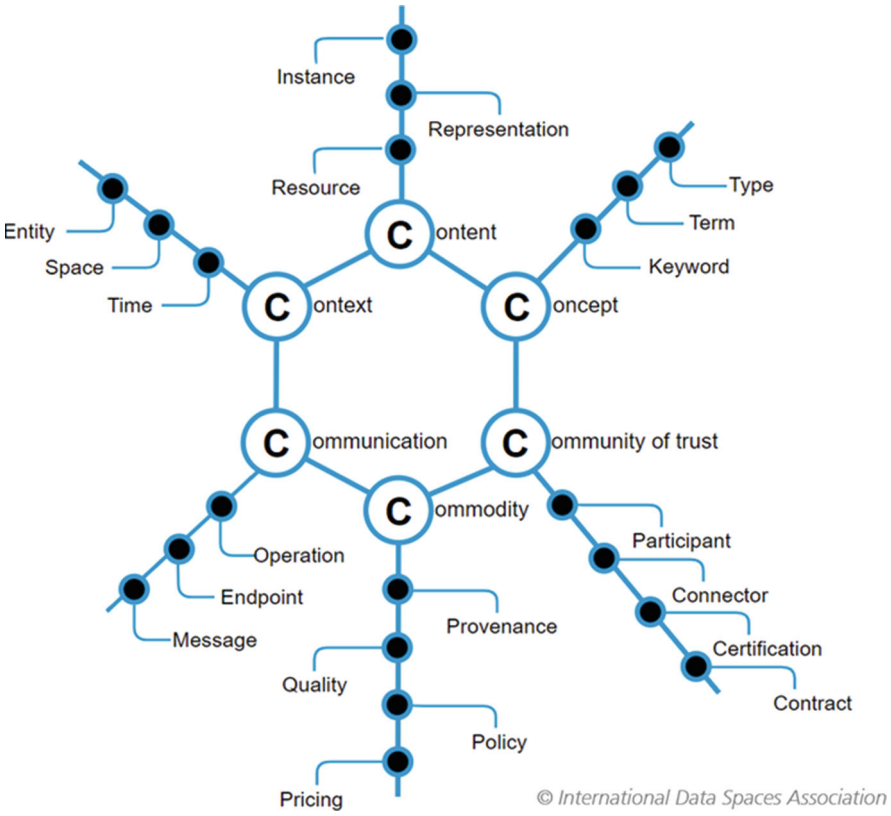


Fig. 7.1 The six concerns defined by the IDS-IM [2] (©2019, International Data Spaces Association)

7.1 Introduction

The declarative representation of the IDS-IM [1] is a vocabulary for the graph-based RDF data model, provided as a lightweight ontology in a GitHub repository[1] under an open license. It serves as the data schema for the Information Layer of the IDS, connecting the conceptual functionalities, roles, and processes with the implementations in the Connector interfaces and endpoints. Figure 7.1 illustrates its structure and the addressed topics grouped into six major concerns.

This chapter focuses on the *Community of Trust* concern. While the IDS regards the digital sovereignty of its participants as its main priority, the latest developments in cloud computing and recent specifications on the decidability of formal usage restriction statements have led to a new understanding how resources are identified

[1] https://github.com/International-Data-Spaces-Association/InformationModel/

and how detailed, machine-readable contracts need to be formalized and embedded into the IDS infrastructure. These evolutionary enhancements of the IDS-IM are introduced in the following.

Section 7.2 explains the extensions of the IDS Resource concept toward the principles of Self-Sovereign Identity (SSI), which is realized by adopting the Decentralized Identifier (DID) and Verifiable Credential (VC) standards. Section 7.3 further describes the formal specifications for the sovereign data exchange in the IDS by presenting the latest concepts of the IDS Usage Contract Language, one core building block of the *Community of Trust* concern. The therefore necessary infrastructure, in particular the so-called Policy Information Point (PIP), is further specified in Sect. 7.4, followed by the Participant Information Service (ParIS) as the dedicated PIP for business-critical metadata in Sect. 7.5. Finally, we conclude in Sect. 7.6.

7.2 Evolving Trust in the IDS Toward Self-Sovereign Identity

Participants in an IDS-driven data space describe themselves, the infrastructure components they operate, and the Resources they provide. Other Participants, e.g., data consumers, derive their trust into these self-descriptions from their trust into the participant that provides them. In version 3 of the IDS Reference Architecture Model [2], trust-related attributes, such as security status or certification for compliance with certain criteria, are assigned to an IDS Connector, the key interface every Participant has to operate to exchange data, in a dynamic manner by a central Dynamic Attribute Provisioning Service (DAPS), a part of the IDS Identity Provider component.

More recently, there is a push toward decentralized identity and trust management in the sense of Self-Sovereign Identity (SSI). The development toward SSI in IDS is most strongly influenced by its discussion in the scope of the Gaia-X federated data infrastructure [3], which builds on many IDS principles [4]. The Gaia-X Identity and Access Management Framework supports an SSI approach based on Decentralized Identifier (DID [5]) and Verifiable Credential (VC [6]). A DID is a special case of a URI and thus compatible with the identifiers used in the IDS. The additional indirection provided by the mechanism of resolving a DID to obtain information about the subject it identifies facilitates privacy and enables verification of identities by cryptographic proofs such as digital signatures. DID resolution may be implemented in a trusted way using, e.g., distributed ledgers.

The information that participants in an SSI setting provide about themselves can be made tamper-evident and more trustworthy by wrapping it into VCs. VCs state claims about a subject and carry cryptographic proofs, such as digital signatures. For example, a credential stating that a Participant has been certified to comply with a certain standard might be signed by an evaluation facility. In Gaia-X, the

self-descriptions of Participants and anything they provide are comprised of such VCs, i.e., the proofs of any claim about a subject reside decentrally with the subjects themselves, like physical ID cards residing in a person's wallet. Once, after successful verification, one Participant Alice decides to trust another Participant Bob's credentials, Alice may continue to operate on the raw claims Bob or other trusted parties made about himself, by extracting them from Bob's credentials and merging all of them into a single, easy-to-query metadata graph, which has the same structure as a plain, untrusted self-description.

Thus, SSI can be realized on top of the IDS-IM in a backward compatible way, which we are planning to pursue in the course of co-evolving IDS and Gaia-X in an aligned way [4].

7.3 Definition of Contract Clauses: The IDS Usage Contract Language and Its Core Concepts

The IDS-IM provides the means to describe and exchange datasets and services. In particular, it combines several prominent standards, for instance the Data Catalogue Vocabulary (DCAT) and the Open Digital Rights Language (ODRL), and further specifies them toward unambiguous definitions of its meaning and implications. The IDS Usage Contract Language, a part of the *Community of Trust* concern, allows the expression of usage restrictions in the form of machine-interpretable data objects. It thus bridges the gap between textual, human-readable contracts and decidable logical statements.

A similar approach is conducted with the Solid framework and its access control model. Both define the restrictions in RDF resources, and both use semantic expressions to indicate allowed operations and to prohibit unintended requests. Still, the two differ significantly in their scope and also the expressiveness of their supplied concepts. These differences are explained in the following and form the basis for the later outlined extensions of the conceptual foundations of the IDS Usage Contract Language and its regarded dimensions and operators.

7.3.1 The Solid Access Control Model vs. IDS Usage Contract Language

The IDS Usage Control Language comprises the necessary concepts and attributes to describe rich usage permissions and how to enforce them. These concepts follow the holistic scope of an overall usage control perspective. Solid, the approach for a decentralized social network of self-controlled data pods, aims at giving human users the sovereignty of their personal data. The data protection solution promoted by Solid—the Web Access Control (WAC [7]) language—differs slightly from the IDS

concepts. First, its scope is restricted to pure access control. The expressiveness of the protection rules, an Access Control List (ACL) or *acl:Authorization*, ends as soon as the data resource is requested by the consumer. Subsequent usage activities and even the further distribution of the resource are not reflected.

Nevertheless, both approaches follow the same identification pattern. The data resources and the Authorization (Usage Contract) related to them follow the URI schema, usually as http URIs. While ACLs in general are used also in many other environments and are open to any—also non-URI—identifier scheme, the special usage of WAC in Solid requires compliance with the Linked Data Platform (LDP) specification. *acl:Authorizations* determine the permitted RESTful operations on the (LDP) target resource and can also be accessed as own LDP Resources. As such, one can even think about *acl:Authorizations* managing the access to other *acl:Authorizations*, something that is not possible for IDS Usage Contracts.

WAC does not allow the specification of a resource owner. This is the same approach as in the IDS model. The authorized entity that has the technical ability to control the resource itself and define its usage permissions is implicitly assumed to be the sovereign of this resource—whereas, for example, Verifiable Credentials held by one entity (e.g., a parent) may make claims about another entity (e.g., their child). The WAC procedure addresses the challenge of ownership by relying on the technical capabilities of users or user agents. One has to note that such indirect proceedings are necessary as an ownership concept—like for physical goods—is not possible or applicable for digital data. The thereby created implication is that the Data Sovereign is the one who has the technical capability and permission to manage the access to a resource, and that the access permissions and control rights are granted to the Data Sovereign. This logical circle requires more in-depth examination.

One further challenge is the definition of user groups. The range starts with a group of the size one, which is equivalent with an individual user. In an WAC statement, this user is identified by its WebID, i.e., a URI. On the other end of the possible descriptions is the group including everyone, which is expressed by the class *foaf:Agent*. As everyone, either human or nonhuman, organization or individual, is by definition an instance of *foaf:Agent*, a permission assigned to this class applies also to everyone.

Between these trivial cases are the more relevant defined groups, for instance the members of a certain organization or company. The best practice for encoding such groups is to use URLs. The Web resource of the group URL shall provide a list of its members. The access management service therefore needs to dereference the group identifier, and thereby receives all required information to accept or decline the access request. The IDS handles such groupings differently. Its business-to-business focus leads to the view of all users or organizations as instances of the generic *ids:Participant* class. Recursive declarations then enable the description of membership. A department is modeled as a Participant, which is part of the bigger company, which is also an IDS Participant.

The core challenge for both approaches is the clarification of whether a user is a valid member of a group or organization. In the terminology of XACML [8], a

Policy Information Point for this type of information is needed. The WAC specifications solve it by using the identifier also as the reference to the PIP interface—the group URI is also the pointer to the Web resource containing its members. That is not possible in the IDS where membership information requires the highest protection level and must not be open to arbitrary Web requests and does not reflect group members. The respective PIPs in an IDS ecosystem are therefore globally known components such as the Participant Information Service or company-specific like an LDAP system.

The WAC scope is limited to describing the following access operations: read, write, append, and control. The IDS Usage Control language must also express activities after the granted request, for instance to not distribute, log, notify about usages, or even delete the resource from the receiving system. Obviously, such obligations cannot be ensured by the providing server but require a trusted system at the receiving party. While the IDS includes specifications and guarantees to achieve such usage policy enforcement, it is not in the scope of Solid or the WAC language.

An additional difference can be found in the inheritance regime of the WAC and the IDS Usage Control language. Solid ACL files are interpreted relative to their location in LDP Containers. That means that the same permissions apply for the contained container and resources. As the IDS does not follow a strict container model, such passing of rules is not possible. Each *ids:Resource* needs its own assigned usage policy. Nevertheless, the IDS follows a similar concept when it comes to the concrete appearances of a data asset. By default, policies specified for the *ids:Resource* are propagated through the related *ids:Representation*s in particular formats to the final *ids:Artifact*s, i.e., their materialized instances. Solid misses this differentiation and consequently has only one way to describe the target data asset.

Another significant difference lies in the embedding into HTTP headers for the runtime discovery of ACL files. As the IDS supports several protocol bindings, the limitation to RESTful discovery operations (GET, HEAD) is not sufficient. Potential consumers therefore need to use infrastructure services, in particular the IDS Metadata Broker, or directly the resource self-description at the hosting connector to find the applying Usage Policies. As the Solid interactions are only enabled through the LDP operations, the discovery mechanism using Link headers are sufficient and no third-party component is required.

7.3.2 Usage Control Dimensions

The general syntax of Usage Control contracts as proposed in the section before enables systems to form expectations about the order and relations of their clauses and parts. The formulation in RDF further defines serialization formats and allows their exchange via messages. Nevertheless, the involved systems also need to understand their content, their implications, as well as the limitations of their own

Table 7.1 Building blocks for feature dimensions (incomplete list)

Feature of interest	Degr.	Reference (anchor)	Datatype/vocabulary
Logic	1	Absolute (Boolean algebra)	*xsd:boolean*
Set	1	Relative	*rdf:List*
Arithmetic	1	Absolute (0)	*xsd:double*
	2	Relative (counter)	*xsd:double*
Text	1	Absolute	*xsd:string*
	2	Relative (pattern matching)	*xsd:string* incl. RegEx pattern
Time	1	Absolute (UTC)	*xsd:dateTimeStamp*
	1	Relative	*xsd:duration*
	2	Absolute (UTC)	Interval (*xsd:dateTimeStamp* × *xsd:dateTimeStamp*)
SpatialEntity	1	Qualitative	e.g., Wikidata Entities
Geometry (point)	2	Relative (local origin)	Coordinates (xsd:float × xsd:float)
Geometry (n-edge polygon)	2n	Relative (local origin)	Point × . . . × Point
GeoSpatial (SpatialThing)	2	Absolute (WGS 84)	WGS84 Geo Positioning
GeoSpatial (n-edge polygon)	2n	Absolute (WGS 84)	SpatialThing × . . . × SpatialThing
Organization	1	Hierarchical	Organization Ontology [9]
Org. and Membership	2	Hierarchical	Organization Ontology
Org and Memb. and Site	3	Hierarchical	Organization Ontology
Network Space	1	IP address (relative)	*xsd:string*
	1	Domain (absolute)	*xsd:string* (DNS scheme)
	1	MAC (relative)	*xsd:string*
Events	?	Relative	Cloud Events
. . .			

capabilities. A pure syntactic or data-oriented view on the contracts directly leads to situations, in which undesired clauses are acknowledged.

Table 7.1 presents the sorted list of feature dimensions agreed throughout the IDS community. Higher-ranked features are assumed to be more relevant, or more commonly used, than lower-ranked ones. For instance, nearly every Usage Control use case requires basic logical statements (something is *true* or *false*), while geospatial relations are less important. In addition, higher-ranked features tend to be better understood and have more mature definitions than lower-ranked ones. Due to their wider usage, more research works have treated them and examined their implications. This assumption is further supported by the observation that, while the "Network Space" and "Events" are crucial for many IDS use cases, still no widely accepted modeling scheme or data format for them can be found in the literature. The Question marks in the table indicate where the specification is still work in progress.

The varying complexity of the different features is a further gap in the literature. To the best of our knowledge, no recent work aims to examine the different requirements on an abstract level. It is important to understand the different kind of input requirements for each dimension. For instance, basic temporal comparisons like *X is before Y* or *X is after Y* are, given that *Y* is a known and fixed value, dependent on the one free variable *X*. Questions like *X is in the three-months testing period*, however, require not only the given period length but also its starting date. Consequently, two input factors need to be known before the statement can be evaluated. The column *Degree of Freedom* (Degr.) contains the analysis for each feature dimension and thereby poses requirements to the actually usable contract clauses.

Definition A degree of freedom is the minimal number of differing input parameters that are necessary to unambiguously define a statement of the respective feature of interest.

That is important to understand why certain contract clauses, in particular Constraints in this work, are complete and others not. Continuing the example of a testing period, a clause like *use only for three months* is not sufficiently defined (start date of the period is missing) and must be rejected by every enforcement system. Understanding such requirements, and linking them to the defined features of interest, is the only way to prevent unsolvable situations at the evaluation time of a contract.

Similar important to the Degrees of Freedom are the anchors of each scale. An intuitive example is the xsd:dateTime datatype defined by XML Schema 1.1. The anchor is the Coordinated Universal Time (UTC) based on the Gregorian calendar. While this is a universal, worldwide identical anchor, this is not necessarily the case for other features. Organizational relations, most importantly the role of an employee in the organizational hierarchy, are different in every company. It is therefore impossible to define such anchors at the design time of a system, and the related information needs to be supplied at the evaluation time. Somehow in between are clauses about geolocations. While there is a widely used coordinate system (WGS 84) based on the GRS 80 reference ellipsoid, other coordinate systems such as ETRS89 can have deviations and need to be mapped beforehand. That implies that—even though coordinates are syntactically valid, and the evaluations on top can be calculated—also the anchors must be the same to gain the same results.

The last column of Table 7.1 contains the specified datatypes for simple types and the used reference for complex ones. Whenever possible, RDF recommendations have been used. Where necessary, well-defined reference ontologies provide the clauses, for instance the ORG ontology [9] for relations and roles between organizations—which *ids:Participant* is also derived from. The table, of course, only represents a first common basis that will certainly need extensions in the future.

7.3.3 Operators for Usage Control Rules

The presented feature dimensions from the previous section define the space in which the Usage Contracts can be applied. The individual clauses in this space are built using operators to create meaningful statements. These operators are limited to two input parameters per clause. While certainly more parameters are possible without affecting the applicability of the operators in general, this limitation reduces the implementation load for applications and has been proven as a suitable trade-off between expressiveness and implementation complexity.

Definition A Usage Control Operator $\Diamond$ is a binary operator identified by a URI, deriving a Boolean value from exactly two input parameter sets A and B:

$$\Diamond : \mathsf{A} \times \mathsf{B} \rightarrow \{\text{true}, \text{false}\}$$

An operator is therefore defined on the input dimensions *A* and *B* and its derivation function. The identifier, a character sequence forming a valid URI, is required for the reference in the rules. As far as the IDS use cases are affected, URIs in the HTTP scheme have proven their applicability and are the preferred pattern.

Table 7.1 aligns the operators with their corresponding feature dimension. The previously mentioned clause—*X is before Y*—can be modeled using the $\Diamond_{\text{BEFORE}}$ operator[2] and *X* and *Y* as time instances. Using graph patterns, we can write:

$$x \Diamond_{\text{BEFORE}} \ \text{y} \rightarrow \{\text{true}, \text{false}\}, \quad x, \text{y} \in \text{TemporalInstance}$$

We can assign a definite value using a date-time stamp for *y*, for instance *"2020-12-31T23:59:59+01:00"^^xsd:dateTimeStamp* or any other unambiguous reference to a temporal instant. Reflecting its position in the statement and according to the ODRL terminology, we call *y* a *RightOperand*. The *LeftOperand*, *x*, may also have a fixed dateTimeStamp, for instance *"2021-01-01T00:00:00+01:00"^^xsd:dateTimeStamp*. In that case however, the clause is trivial and therefore uninteresting, as it always evaluates to *true*. More relevant for real-world use cases are references to the current point in time, for instance to express that a *now* must be *before* a certain time instance, as the contract shall limit the usage time. It is therefore necessary to define a set of references, which the Usage Control system is aware of and can use to evaluate the applicability of the contracts.

This demand is reflected by the list of *LeftOperands*. *Now* is represented by the *POLICY_EVALUATION_TIME* instance. Additional instances, like *PAY_AMOUNT*, allow the description of usage fees or make references to certified security characteristics (*SECURITY_LEVEL*) of the regarded application. Combining all these parts allows the data sovereign to state its intentions and gives the

[2] $\Diamond_{\text{BEFORE}}$ is identified by the URI *https://w3id.org/idsa/code/BEFORE* in the IDS-IM.

Table 7.2 Binary operators for Usage Control constraints

Feature of interest	Compared values	Operators
Logic	Boolean × Boolean	EQUALS, NOT
Set	Element × Collection	IN
Arithmetic	Double × Double	EQ, LT, GT, LTEQ, GTEQ
Text	String × String	STRING_EQ, STRING_CONTAINS, STRING_IS_CONTAINED
Regex	String × RegEx	MATCHES
Time	Instant × Instant	BEFORE, TEMPORAL_EQUALS, AFTER
	Instant × Interval	BEFORE, MEETS (= STARTS), TEMPORAL_EQUALS (= CONTAINS), MET_BY (= FINISHES), AFTER
	Interval × Duration	LONGER, LONGER_EQ, AS_LONG, SHORTER_EQ, SHORTER
	Interval × Interval	BEFORE, MEETS, OVERLAPS, DURING, STARTS, TEMPORAL_EQUALS, CONTAINS, OVERLAPPED_BY, MET_BY, FINISHES, AFTER
	Duration × Duration	LONGER, LONGER_EQ, AS_LONG, SHORTER_EQ, SHORTER
Geospatial	Point × Point	DISJOINT, SPATIAL_EQUALS
	Polygon × Point	DISJOINT, SPATIAL_MEET, CONTAINS
	Polygon × Polygon	DISJOINT, SPATIAL_MEET, SPATIAL_OVERLAPS, INSIDE, COVERED_BY, SPATIAL_EQUALS, COVERS, CONTAINS
Organization	User × Organization	MEMBER_OF
	User Org. × Role	HAS_MEMBERSHIP (includes MEMBER_OF)
	User Org. × Role × Site	HAS_SITE (includes HAS_MEMBERSHIP)
Network . . .	Network Location × Network	INSIDE_NETWORK (?), . . .
Event	Instant × Event	?
. . .		

Operators expand to URIs in the *https://w3id.org/idsa/code/* namespace

downstream consuming systems the information how they need to handle the incoming data. The according behavior of their applications can be certified and proven to remote parties using trustworthy tokens. The IDS promoted Dynamic Attribute Tokens (DAT) contain such claims, combining the results of certification processes with cryptographic signatures; in Sect. 7.2, we have explained how to further decentralize this approach to Self-Sovereign Identities. The result is an ecosystem where usage intentions can be expressed, exchanged, and implemented on a technical level (Table 7.2).

The outlined definitions to merge Usage Control and Data Sovereignty with decidable constructs show a way of how to create an interoperable and at the same

time protected ecosystem for data. The provided assumptions and statements are intended as starting points for further discussions. We believe that a consolidation of the thereby affected approaches and concepts is necessary, and a trade-off between formalization and expressed details on the one hand and adoption on the other is indispensable.

Still, the demand for more and more autonomously acting systems enforces overhead in terms of data models and implementations. IDS Usage Contracts show how the different specifications can be combined to a comprehensive interpretation. We have observed that similar ideas and pattern appear in the different domains. The integrated approach can therefore also serve as a bridge between the communities. The result however has the potential to disrupt the way we treat digital information and how trustworthy systems can technically enforce the initial restrictions.

7.4 The Policy Information Point

The concept of Policy Information Points, or PIPs, depicts one of several roles in a decision system as shown in Fig. 7.2. Similar to the other policy points, a PIP is an abstract role assignment rather than a concrete software asset. That means that a single application can serve, for instance, as a PIP, a Policy Decision Point (PDP), and a Policy Execution Point (PEP) at the same time. Still, the separation of the different terms helps to define the different capabilities and responsibilities in a decision workflow and to describe which tasks are executed by which component.

In the IDS, the PIP provides the capabilities to provide the information that is necessary to evaluate whether the conditions of a formal contract are met or not. For instance, if the usage is restricted to a certain duration, the PIP supplies the elapsed

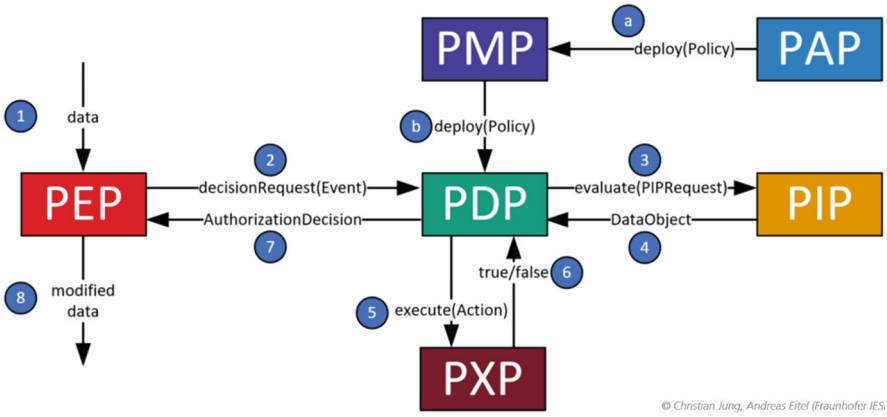


Fig. 7.2 The PIP in the workflow between the different policy points (illustration from [10]) (©2021, International Data Spaces Association)

usage time. It is important to note that the PIP itself does not derive any decision itself. It only provides the interface to, in this example, a time service without any further indication.

Obviously, PIPs can appear in many different realizations in an IDS ecosystem. As explained, one appearance can be an API to Connector-internal functionalities. A second option is the realization as a stand-alone application in the network, serving required information for several Connectors and their usage enforcement systems. A further distinction can be made by the operation mode of a PIP component. While for the former use case the PIP is part of the functional architecture of the Connector itself, or operated close to it, the provisioning of reliable and accurate information can as well present a business model that justifies a certain fee. Such PIPs are operated as independent components of an IDS infrastructure, and their integration depends on the configurations of a usage contract. Therefore, the communication is initiated at the runtime of the consuming connectors, which implies the usage of standardized interfaces and defined data objects both for the incoming request and for the returned response. The following explanations discuss these features and present the latest state of the community discussion.

Definition An *External PIP* is an IDS Connector that supplies at least one attribute of interest for the evaluation of Usage Contracts. It must outline its role in its self-description together with the path to the endpoints where the provided attributes can be accessed through IDS-specified communication channels.

In particular, the applied definition demands that PIPs operate valid DATs and are capable of communication in at least one supported IDS message protocol. This especially does not enforce the support for all protocols. For instance, one PIP may only have one endpoint for the IDS Communication Protocol (IDS-CP), while another also supports the IDS REST binding. A requesting Connector cannot, for now, expect that its protocol of choice is possible without further lookups in the PIP self-description.

Furthermore, a PIP is not obliged to answer an incoming request. Possible rejection reasons could be an expired or otherwise invalid DAT of the requesting Connector, or open fees that need to be paid beforehand. In any case, the PIP must—as every other Connector—respond with an according rejection message indicating the issue.

The information request can happen in different ways, for instance in a message-oriented way using an *ids:RequestMessage*. The currently recommended pattern, however, uses the IDS REST binding that models the PIP as a Linked Data Platform server with so-called container resources. A container in that sense serves both as an endpoint for incoming requests but also as the destination for lookups. It provides descriptions about its functionality, the required input and output parameters, and indicates its capabilities using the IDS-IM.

An IDS Connector discovering a PIP can therefore traverse through its descriptions and recognize endpoints and the information offers supplied therein on the fly. This is especially relevant as in a common scenario, the data provider demands the

usage of a certain PIP to evaluate the applicability of a Usage Contract. The consuming Connector usually is not aware of the existence of this specific PIP and consequently needs to integrate its endpoint during its runtime. The execution of a sequence of *read* requests is at that point significantly easier for the consuming Connector than the on-the-fly integration of a message endpoint.

An External PIP as regarded by the IDS strictly distinguishes between the provisioning of information and its meaning and implications at the time of the evaluation of a contract. The PIP therefore must not anticipate the effects of an information request and act as a neutral and unbiased source. As such, it combines the conceptual role from the usage control perspective, which requires a standardized entity for attribute queries, with the communication interfaces of a Connector.

The PIP obviously needs to also implement the features required of any Connector, in particular the provisioning of a self-description conforming to the IDS-IM but also a valid and verifiable DAT, which also relies on IDS-IM attributes. PIPs however easily go beyond the expressiveness of the Information Model, for instance when domain-specific data attributes are supplied. Such information must be integrated with the technology stack of the IDS, in particular the usage of RDF. Nevertheless, the proper selection of the data format alone is not sufficient. To ensure the possible incorporation of the provided attributes, their semantic meaning also needs to be specified. Open catalogues like ECLASS or from the Linked Open Data Cloud contain the required terms and concepts and often provide lookup facilities. A client connector using a PIP can thereby not only automatically traverse the PIP itself and discover its provisioning on the fly but also understand the external concepts used by applying the same operations. This significantly simplifies the integration efforts on the developer side and paves the way for a broad adoption of the IDS PIP concept.

7.5 The Participant Information Service (ParIS)

One of the most important value propositions of the IDS is the enablement of business interactions between previously unrelated Participants. That aims at companies that have not met before in the digital or non-digital world but now start business agreements solely relying on the IDS. The therefore necessary trust in the opposite party is technically achieved by a verifiable identity management process through the Identity Provisioning Service and the DAPS. Both components equip each Participant with the necessary attributes and cryptographic proofs for the IDS handshakes.

The establishment of a secure and uncompromised communication channel is however only the necessary requirement for a business interaction. In addition, the respective Participants need to understand their opposite's state in regard to business workflows. For instance, every business actor needs to know its customers' tax identification or VAT number to create correct invoices. Furthermore, the registered

address is critical to understand the responsible jurisdiction for the unfortunate cases when only courts can solve conflicts.

Such information is provided and maintained by the Support Organization in an IDS, the legal entity that administers the ecosystem. The Support Organization introduces a new Participant by creating its digital identity and at the same time registers security-critical attributes at the DAPS and business-relevant attributes at another technical component: the Participant Information Service (ParIS), a PIP for Participant attributes. The ParIS provides access to these attributes to the other IDS Participants and components and connects the unique Participant identifier—a URI—with additional metadata. Usually, each IDS ecosystem operates only a small number of ParIS instances, most often just one. IDS Participants therefore know the location where to ask for more information about a potential business partner and can decide whether to start a data exchange.

Different from other IDS components, the trustworthiness of ParIS' provisioned information is not grounded on technical measures such as signatures or certificates, but on the administrative process controlled by the Support Organization. A direct consequence of this process is the necessity that each change request is manually verified before being added to the ParIS database.

The initial population of a Participant entry is conducted directly after the certification and identity creation process is finished. The Support Organization is informed about the successful steps and provided with the corresponding metadata about the new entity (cf. Table 7.3). The provisioning of this information is not part of the IDS interactions yet and must be managed through traditional communication measures. The Support Organization verifies the correctness of the claims, verifies the information, and—together with other additional steps—equips the dedicated ParIS with the new IDS Participant instance. It is further recommended that each Participant also hosts this self-description on a publicly accessible HTTP server of its choice. Preferably the locator of this self-description document, a HTTP URL, is identical with the used Participant URI. This best practice, one of the "Linked Data" principles, enables the lookup or dereferencing of the Participant Identifier through every HTTP client and thus eases the discovery of relevant information. Nevertheless, in case the own supplied Participant self-description and the metadata at the ParIS deviate, the latter is more trusted, as its claims have been verified through the Support Organization beforehand.

Requests to the ParIS itself follow the IDS Message Model and are described, among others, in the IDS Communication Guide and the IDS Information Model. The most common call, the request for a description of an identified IDS Participant, is executed using the *ids:DescriptionRequestMessage* with the Participant Identifier as the value of the *ids:requestedElement* attribute. The response message contains a JSON-LD representation of the Participant instance or returns an error message in case the requested identifier is not known to the server.

The attributes provided by the ParIS (see Table 7.3) are in general not critical for the security aspects of the IDS but highly relevant for every business interaction. While for instance IDS handshakes can be executed without knowing the physical location of the opposite party, business processes need to know about the respective

Table 7.3 Attributes of a Participant as provided by a ParIS

Attribute	Datatype/ target class	Description
ids:version	*xsd:string*	Version identifier of the Managed Entity
gr:legalName	*xsd:string*	The complete legal name of the IDS Participant, for instance "International Data Spaces e. V." is the legal name for the IDSA
gr:name	*xsd:string*	The commonly used name or term for the IDS Participant, for instance "IDSA" for the legal entity "International Data Spaces e. V."
ids: corporateEmailAddress	*xsd:string*	Email address for contacting the participant on a general level
ids:corporate Homepage	*xsd:anyURI*	General official homepage of the participant
ids:member Participant	*ids: Participant*	Indicates that a participant has a member which is again a participant. This is useful for defining hierarchical relations in a participant's organization as well as identifying groups of participants to capture, e.g., members of a collaboration
ids:participant Refinement	*ids:Abstract Constraint*	Conditions that need to be satisfied for a single Participant to be seen as a member of the subject Participant. For instance, all Participants with their headquarter in Europe might be potential consumers of GDPR-related data
ids:memberPerson	*ids:Person*	Indicates membership of a person to an organization
ids:participant Certification	*ids:Participant Certification*	Certification issued for the given Participant
ids:primarySite	*ids:Site*	Indicates a primary site for the organization, this is the default means by which an organization can be contacted and is not necessarily the formal headquarters
gr:vatID	*xsd:string*	The value-added tax ID of the IDS Participant. See http://en.wikipedia.org/wiki/Value_added_tax_identification_number for details
gr:taxID	*xsd:string*	The tax /fiscal ID of the IDS Participant, e.g., the TIN in the USA or the CIF/NIF in Spain. It is usually assigned by the country of residence
ids:registryCourt	*xsd:string*	The responsible court for the organization. Usually, a city and country
ids:legalForm	*xsd:string*	The legal form of the IDS Participant
ids:hasRole	*xsd:anyURI*	The internal organizational roles that the Participant wants to announce to business partners to restrict data usage. For instance, if the organization has dedicated "risk managers," Data Providers can state that only users with this role can see the IDS Resource. A role must be encoded with a valid, unambiguous URI, for instance https://<company-domain>/role#<role-name>

country and tax regulations, or which is the complete, legally valid name to create a proper invoice. Table 7.1 contains the currently defined properties of an IDS Participant that can be used to model the metadata for a ParIS. The described properties represent the minimal set of attributes each ParIS implementation must accept. Nevertheless, specialized implementations may also accept and provide additional properties, dependent on the domain-specific requirements or its technical capabilities.

The IDS ParIS is an infrastructure component in the IDS general architecture with a similar functionality as the IDS Metadata Broker. While the latter indexes Data Resources and Connectors, the ParIS makes Participant information available. Further extensions of its specification may include a more fine-grained access and update model. This could include the ability for each Participant to update noncritical attributes of its own description, for instance the corporate email address or the member persons. Other attributes, such as the VAT number or the legal name, will still require a verification process through the Support Organization. Realizing such options will enable a faster and more accurate provisioning of descriptive metadata while the trust level for critical attributes stays the same. The IDSA Architecture Working Group structures these activities and prepares the further development of the ParIS specification.

7.6 Conclusion: The IDS-IM as the Bridge Between Expressions, Infrastructure, and Enforcement

This chapter presents several aspects of the IDS Information Model. It explains the new state of discussions to combine decentralized identification provisioning methods with the rich expressiveness of the already elaborated vocabulary. The IDS-IM is supplied as an RDF ontology conforming to the conventions and best practices of open data resources and maintained in a transparent process. One of the activities therein is the integration of defined logical statements to formally express the intentions of usage restrictions in machine-interpretable data objects. Different from previous approaches, the IDS-IM not only extends the descriptive representations to access decision, like the WAC vocabulary, but faces the significantly more complex challenge of enforcements in the data usage phase.

This is only possible with defined infrastructure components that provide reliable and standardized attributes in IDS-compliant manners. The IDS interpretation of the PIP fulfills this role and enhances the reference architecture with the capabilities to execute applicable usage enforcement into data ecosystems. In particular, the ParIS further specifies the PIP functionality for Participant attributes and provides all Connectors with the necessary standardized interfaces to retrieve business-critical information from the IDS Identity Provider components.

As such, the latest development of the IDS-IM incorporates the requirements of state-of-the-art cloud computing with the expressiveness needed to implement true

data sovereignty. Its next evolvements include the implementation in cloud-native environments.

References

1. Bader, S., Pullmann, J., Mader, C., Tramp, S., Quix, C., Mueller, A., Akyürek, H., Böckmann, M., Imbusch, B., Lipp, J., Geisler, S., & Lange, C. (2020). The international data spaces information model. In *Proceedings of the International Semantic Web Conference*. Springer.
2. Otto, B., Steinbuß, S., Teuscher, A., Lohmann, S., & et al. (2019). *IDS reference architecture model, version 3.0*. International Data Spaces Association.
3. Gaia-X. (2021). Gaia-X Architecture document, release 21.03, Gaia-X, European Association for Data and Cloud, AISBL, Brussels. https://gaia-x.eu/pdf/Gaia-X_Architecture_Document_2103.pdf.
4. Otto, B. (ed.) (2021). *GAIA-X and IDS*. Position paper, version 1.0, International Data Spaces Association. https://www.internationaldataspaces.org/wp-content/uploads/2021/01/IDSA-Position-Paper-GAIAX-and-IDS.pdf.
5. W3C. (2021). *Decentralized Identifiers (DIDs) v1.0 – Core architecture, data model, and representations*. W3C Candidate Recommendation Snapshot. Accessible at https://www.w3.org/TR/2021/CR-did-core-20210318/.
6. W3C. (2019). *Verifiable credentials data model 1.0 – Expressing verifiable information on the Web*. W3C Recommendation. Accessible at https://www.w3.org/TR/2019/REC-vc-data-model-20191119/
7. Solid Community. (2021). Web Access Control (WAC), version 0.5.0. Available online: https://solid.github.io/web-access-control-spec/ (accessed on 04.02.2021: https://github.com/solid/web-access-control-spec/tree/58eae0f548a11c02c30bf2f4a1d620f5ed147490).
8. OASIS eXtensible Access Control Markup Language (XACML) TC. (2013). Accessible at https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml.
9. W3C. (2014). *The organization ontology*. W3C Recommendation, D. Reynolds (editor), accessible at http://www.w3.org/TR/2014/REC-vocab-org-20140116/.
10. Eitel, A., Jung, C., Brandstädter, R., Hosseinzadeh, A., Bader, S., Kühnle, C., Birnstill, P., Brost, G., Gall, M., Bruckner, F., Weißenberg, N., & Korth, B. (2021). *Usage control in the international data spaces*. Position Paper, Version 3.0, International Data Spaces Association. https://internationaldataspaces.org/wp-content/uploads/IDSA-Position-Paper-Usage-Control-in-the-IDS-V3.0.pdf

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 8
Data Usage Control



Christian Jung and Jörg Dörr

Abstract Data-driven business models are based on sharing and exchanging data. However, to establish a trustworthy and secure data exchange between different organizations, we have to tackle several challenges. Data sovereignty, for instance, is an essential prerequisite to empower data-driven business models across different organizations. The International Data Spaces provide solutions for data sovereignty to implement a secure and trustworthy data economy.

In this chapter, we focus on data usage control and data provenance as building blocks to solve data sovereignty challenges. We introduce concepts and technology for realizing usage control and describe the differences between usage control and access control as well as other related concepts such as digital rights management or user managed access. We present the implementation of data sovereignty in the International Data Spaces starting from the formalization of data usage restrictions as policies (i.e., the policy specification) to the technical compliance and adherence of the data usage restrictions (i.e., the policy enforcement). In doing so, we present the transformation of data usage restrictions to machine-readable policies that can be enforced by the systems. Different technologies, such as the MY DATA Control Technologies can be used to implement the enforcement of data sovereignty in a technical manner and discuss future expansion stages of implementing data sovereignty.

8.1 Introduction

Today's data-intensive business is driven forward by continuously exchanging critical and sensitive data between business partners. Data is typically secured by access control mechanisms, which means that data can be used (e.g., altered, copied, disseminated) without further restrictions after access to the data has been granted.

C. Jung (✉) · J. Dörr
Fraunhofer Institute for Experimental Software Engineering IESE, Kaiserslautern, Germany
e-mail: christian.jung@iese.fraunhofer.de; joerg.doerr@iese.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_8

However, controlling only access to data is not sufficient to establish a trustworthy and secure data economy and to realize innovative business models.

Therefore, the International Data Spaces (IDS) establish a virtual data space, where businesses can exchange and exploit the potential of data in a trustworthy and secure manner. Data sovereignty is a key success factor to establish trust between all involved business partners for implementing data-driven business. Data sovereignty provides data owners full control over their data, which includes the (restricted) usage of their data by data consumers [1]. Hence, access control is extended by mechanisms to control the usage of data (i.e., data usage control) after access is granted.

We differentiate between intrinsic and extrinsic motivations to implement data sovereignty. On the one hand, organizations realized that more and more businesses are spurred by data. Hence, their own data gets more valuable and must be protected to prevent misuse. This also includes the protection of their intellectual property. Protecting their data by completely locking it is not a solution as it disables business performance. Therefore, organizations have an intrinsic motivation to implement data sovereignty to benefit from data-driven business while keeping control over their own data. On the other hand, domain requirements, standards, rules, and legal obligations constrain organizations in their possibilities to use their data. Above all, data protection regulations such as the European Union General Data Protection Regulation (EU-GDPR) [2] must be mentioned here as its implementation is often difficult and data protection incidents can be very costly. We call this extrinsic motivation, because it is induced by external factors.

We structure the remainder of the document as follows. Section 8.2 presents the general concept of data usage control and explains the difference to access control and its relation to other concepts. We address policy specification, i.e., the specification of usage control rules, in Sect. 8.3, followed by technologies to implement data sovereignty. In Sect. 8.4, we briefly conclude with a discussion about future expansion stages for data sovereignty.

8.2 Usage Control

Usage control is an extension to traditional access control (see Fig. 8.1). It is about the specification and enforcement of restrictions regulating what must (not) happen to data. Thus, usage control is concerned with requirements that pertain to data processing (obligations), rather than data access (provisions). Usage control is relevant in the context of intellectual property protection, compliance with regulations, and digital rights management.

In this section, we describe access control (as the basis) and usage control, followed by subsections describing the different building blocks to implement data usage control.

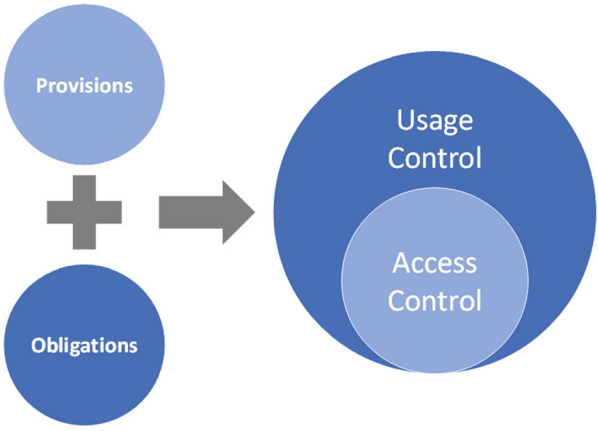


Fig. 8.1 Usage control extends access control

8.2.1 Access Control

In general, access control restricts access to resources. The term authorization is the process of granting permission to resources. Several access control models exist, such as Discretionary Access Control (DAC), Mandatory Access Control (MAC) [3], Role-Based Access Control (RBAC) [4], Attribute-Based Access Control (ABAC) [5], etc. Although such a plethora of access control models exists, RBAC and ABAC are most commonly used.

We use the XACML (eXtensible Access Control Markup Language) Standard [6] to introduce commonly used terms in the field of access control. XACML is a policy language to express ABAC rules. The main building blocks of the language are subject, action, resource, and environment. The subject describes who is accessing a data asset (e.g., a user). The action describes what the subject wants to perform on the data asset (e.g., read or write). The resource describes the data asset. Finally, the environment specifies the context (e.g., time, location).

Figure 8.2 illustrates the data-flow model of XACML and the main actors or components to implement it: Policy Enforcement Point (PEP), Policy Decision Point (PDP), Policy Information Point (PIP), and Policy Administration Point (PAP).

Access control in the IDS is a resource-centric regulation of access requests from subjects (i.e., IDS participants) to resources (i.e., data services). Resource owners define attribute-based access control policies for their endpoints and define the attribute values a subject must attest in order to grant access to the resource. These attributes may include:

- Specific identity of a connector (only access requests from a specific connector will be granted)
- Connector attributes (only access requests from a connector that possesses specific attributes will be granted)

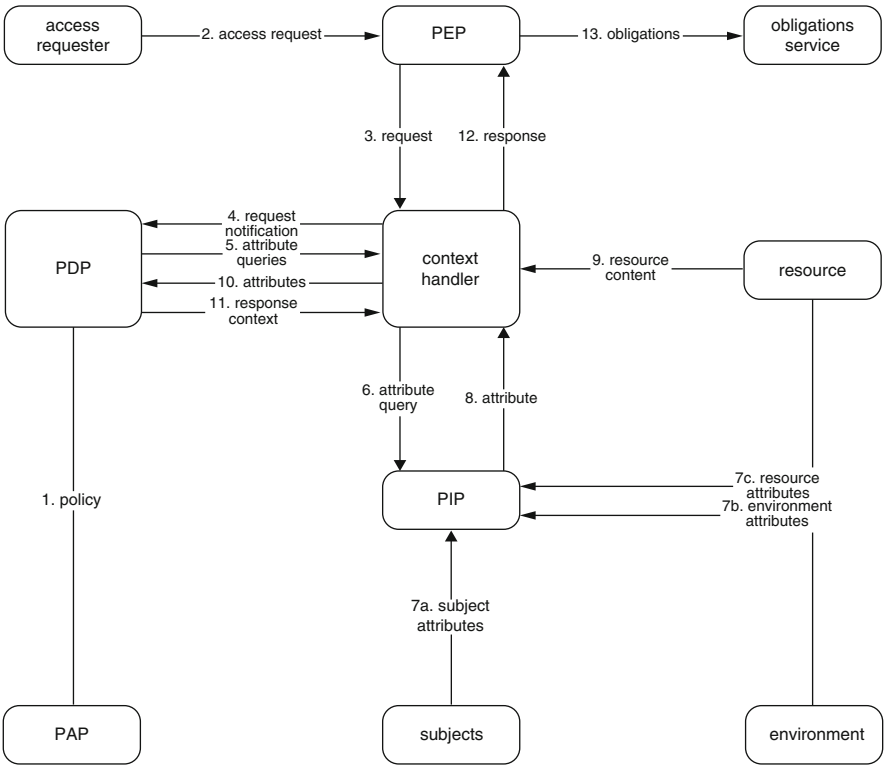


Fig. 8.2 XACML data flow illustration [6], © 2013, OASIS Standard

– Security profile requirements (only access requests from a connector that fulfills specific security feature requirements will be granted, e.g., having a TPM ≥ 1.2 and doing application isolation)

The IDS security architecture does not dictate a specific access control enforcement language or implementation.

8.2.2 Usage Control

In contrast to access control, where access to specific resources (e.g., a data service) is restricted, the IDS architecture additionally supports data-centric usage control. The overall goal is to enforce usage restrictions for data after access has been granted. Therefore, the purpose of usage control is to attach policies to data that specify their usage restrictions and to continuously control the way how data is processed, aggregated, or disseminated within the IDS. This data-centric perspective

allows data owners to specify how their data is used, rather than only controlling access to the data services.

At runtime, the usage control enforcement prevents IDS connectors from treating data in an undesired way, for example, by disseminating personal data to public endpoints. Thus, usage control empowers IDS participants to ensure they are not building an architecture that violates data sovereignty. In addition, usage control mechanisms monitor data flows, which can be used as an audit mechanism to create evidence of a compliant data usage.

In the following examples, we illustrate data sovereignty requirements that cannot be achieved using traditional access control, but rather require data-centric usage control:

- **Data Secrecy:** Classified data can only be forwarded to nodes and services with the respective clearance or security level.
- **Data Integrity:** Critical data can only be modified by trusted nodes or services.
- **Time to Live:** Data must be deleted or altered after a given period of time.
- **Anonymization:** (Personal) Data has to be anonymized before use, for instance by aggregation or replacement.
- **Separation of Duty:** Data sets, for instance from competitive organizations, must be kept separated (e.g., no joining operation or processing within the same service).
- **Usage Scope and Purpose:** Data can only be used within trusted nodes or services. In addition, data can only be used for specific usage purposes.
- **Context Awareness:** Data can only be used by meeting specific contextual conditions (e.g., only within company premises or only within a specific geolocation).

It is important to note that the purpose of usage control is to allow the specification of such constraints and enforcing them in the running system. It is a prerequisite to usage control that the enforcement mechanism itself is trusted, i.e., usage control itself does not establish trust in an endpoint. It rather builds upon an existing trust relationship and facilitates the enforcement of legal or technical requirements as well as data privacy regulations.

8.2.3 Usage Control Components and Communication Flow

For enforcing usage restrictions, data flows are monitored and intercepted by control points (i.e., Policy Enforcement Point, PEP), as illustrated in Fig. 8.3. Information about the intercepted data flows is sent from the PEP component to the decision engine (i.e., Policy Decision Point, PDP) for allowing or denying the data flow. The decision from the PDP may include instructions to modify the data of an allowed data flow.

The PDP takes the responsibility to answer incoming permission requests for data flows from a PEP. The decision from the PDP is based on the evaluation of the

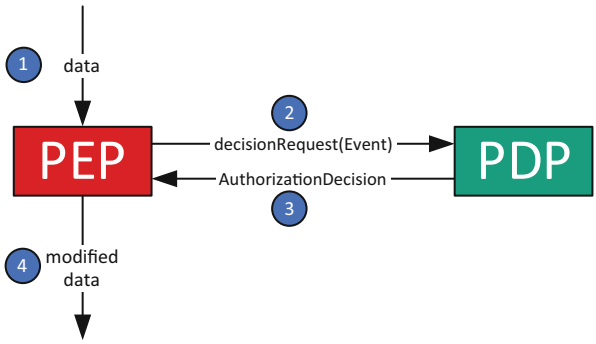


Fig. 8.3 Usage control communication flow (simple)

deployed usage restrictions (i.e., policy evaluation). The PDP decisions are received by the PEP component, which also encapsulates the enforcement of a PDP decision.

Figure 8.3 depicts the communication flow between PEP and PDP.

The policy evaluation may depend on information that is not directly present in the intercepted data flow. Such information may be contextual information such as the geolocation of the organization or membership information from a directory service. A Policy Information Point (PIP) takes responsibility to provide missing information for the policy evaluation during the policy evaluation phase in the PDP. It offers a standardized interface to the PDP and encapsulates the necessary logic to perform information retrieval.

Finally, there is the concept of Policy Execution Point (PXP). A PXP is used to perform additional actions based on the deployed policies. Examples for such additional actions are the sending of an email when data is used or sending a log message to a third-party logging system. Figure 8.4 illustrates an exemplary sequence of all processing steps (incl. optional PIP and PXP steps) to enforce usage control restrictions on a data flow.

In the following, we describe all processing steps in more detail: The PEP intercepts the data flow in the target system (1). Data is extracted from the data flow and prepared as a decision request for the PDP. The PEP requests a decision

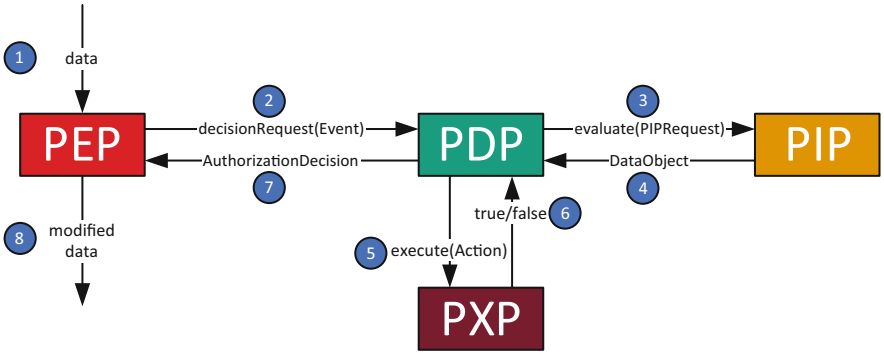


Fig. 8.4 Usage control communication flow (extended)

from the PDP (2). By receiving the decision request, the PDP starts its internal policy evaluation. In doing so, the PDP processes all deployed policies to draw a decision. Depending on the policy evaluation, the PDP may need additional information and invokes a PIP (3). The PIP processes the evaluation request, which may result in an interaction with further external systems. After information retrieval, the PIP responds with a data object to the PDP including the requested information (4). The PDP further processes the policy evaluation. If it includes the execution of additional actions, the PDP invokes a PXP (5) including parameters to execute the action. The PXP processes the action and returns to the PDP whether the action succeeded or not (6). The steps to retrieve additional information from a PIP (i.e., (3) and (4)) or to perform an additional action with a PXP (i.e., (5) and (6)) can be done several times, depending on the policy. The final step of policy evaluation is the preparation of the authorization decision and sending it back to the PEP (7). The PEP receives the decision and denies or allows the data flow (8). For the allow case, the authorization decision may include additional information to modify the data in transit, which is then performed by the PEP. Detailed information about policy specification and the resulting behavior during runtime can be found for the MY DATA Control Technologies at their developer website [7].

8.2.4 Specification, Management, and Negotiation

Another important aspect of usage control is the specification and management of usage restrictions. Data providers express the usage restrictions for their data in a more or less formal way. For a technical enforcement, the result of a policy specification must produce a machine-readable output. In order to simplify the specification process for the end-user, a Policy Administration Point (PAP) is used as an entry point for specification of usage policies, usually providing a user-friendly graphical interface.

A Policy Management Point (PMP) administers the usage restrictions. Hence, the component is concerned with the policy life cycle. This includes the instantiation, negotiation, deployment, and revocation of usage restrictions, as well as conflict detection between policies and, if necessary, the resolution of the conflicts. PAP and PMP completes the usage control communication flow as presented in Fig. 8.5.

We differentiate two ways to locate and transfer usage restrictions. First, usage restrictions can be attached to the data, which is also called sticky policy [8]. Sticky policies are one way to cope with the distribution of usage restrictions. In this approach, machine-readable usage restrictions (policies) stick to data when they are exchanged. There are different realization possibilities. Usually, data is encrypted and can only be decrypted when adherence to usage restrictions is guaranteed. The main advantage, the distribution of usage restriction, comes free of charge as it is transferred with the data. However, it is also the main disadvantage as data and policy are transferred before adherence to the policy can be guaranteed.

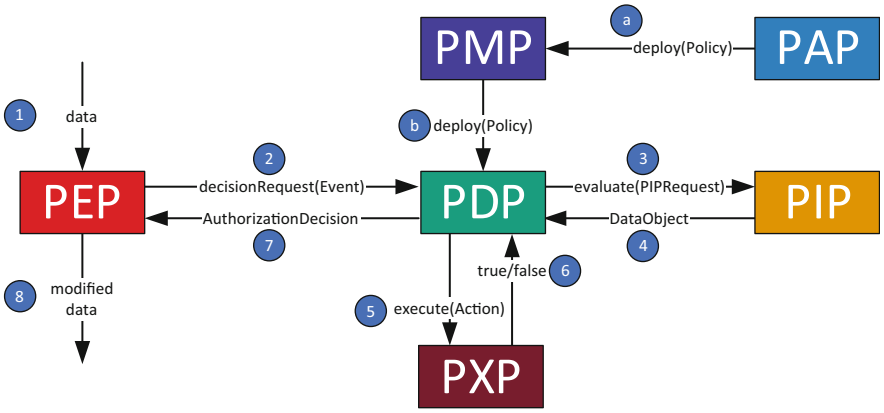


Fig. 8.5 Usage control communication flow (complete)

Moreover, usage restrictions cannot be altered without revoking all instances of the data/policy tuple.

Second, policies can be stored independently from the data in a central component. For instance, policies are directly stored in the PMP or in a dedicated storage such as a Policy Retrieval/Repository Point (PRP). In this case, the management component is responsible to exchange the usage restrictions between different systems. Hence, we have the disadvantage to cope with the distribution of usage restriction, but the advantage to transfer data only, if the usage restrictions are fulfilled by the target system.

The management of usage policies becomes especially important when exchanging data across system boundaries. Every time data crosses system boundaries, the target system must be prepared for the protection of incoming data. Therefore, the corresponding policies need to be negotiated, instantiated, and deployed. Hence, policy negotiation is also part of policy management. As enforcement mechanisms may work differently in various technical environments, policies must also be instantiated on the target system. And finally, before establishing data exchange, the policies have to be deployed, i.e., activated, on the PDP.

8.2.5 Related Concepts

In addition to access control, there are more related concepts to cope with data sovereignty challenges. In order to make the concept of usage control more clear, we present related concepts to usage control such as Data Leak or Loss Prevention (DLP), Digital Rights Management (DRM), User Managed Access (UAM), and Windows Information Protection (WIP). These concepts are no replacement for usage control, but rather supplementary technologies to support data sovereignty.

8.2.5.1 Data Leak/Loss Prevention

Data Leak/Loss Prevention (DLP) technologies detect and prevent potential data breaches by monitoring sensitive data. Commonly used are Endpoint DLP solutions that run on the client's operating system (e.g., as extension or feature of a security suite). In addition, there are also DLP solutions available that are monitoring the network or access to central storage devices.

Hence, the solutions are rather preventing unintended data leakages, but do not control the adherence to usage restrictions during data usage.

8.2.5.2 Digital Rights Management

The term Digital Rights Management (DRM) is frequently used in the area of protecting digital content from unintended use, modification, and distribution. Different DRM technologies exist to protect multimedia content such as movies (e.g., DVD, Blu-ray), music (e.g., audio CDs, Internet music), television, or E-books. In addition, there are DRM technologies to protect digital documents (e.g., MS Word, PDF) within enterprises. This kind of DRM is also known as enterprise rights management (ERM) or information rights management (IRM) and aims to control access and use of corporate documents.

DRM is closely related to data usage control, especially when focusing on enforcing usage restrictions on clients. However, usage control can be implemented at the client as well as at infrastructure components such as the IDS Connector. Hence, it is more flexible and extensible.

8.2.5.3 User Managed Access

The purpose of User Managed Access (UMA) is to empower the resource owner to control the authorization of data sharing. It is often used to protect resources between online services on behalf of the owner. OAuth-based access management systems are representations of UMA [9]. Several open-source implementations exist that follow the UMA core protocol [10].

UMA and related solutions such as OAuth focus on the authorization and the access, but rather neglect the usage. Usage control explicitly focuses on the usage, rather than access.

8.2.5.4 Windows Information Protection

Microsoft introduced several technologies to establish a comprehensive information protection in their operating system and software such as Microsoft Office (e.g., BitLocker, Windows Information Protection (WIP), Office 365, and Azure

Information Protection (AIP)) [11]. WIP, for instance, is an integral part of Windows 10. The goals of the WIP are to protect data on its own devices, to separate private and business data (data separation), to prevent unauthorized access and use (data leakage protection), and to protect data when shared. WIP-protected documents can only be used in WIP-compliant apps. For example, WIP prevents pasting sensitive information (e.g., by using ctrl+c and ctrl+v) to non-WIP-compliant apps.

WIP and AIP are very closely related to usage control concepts, but with the disadvantage of a technology and vendor lock-in. However, it can complement existing usage control solutions. Moreover, usage control is more flexible and extensible.

8.3 Usage Control in the IDS

We have two activity streams within the IDS to implement data usage control: First, a policy language to express data usage restrictions. The policy language is descriptive and technology independent and initially based on the Open Digital Rights Language (ODRL). To express usage restrictions within the IDS, we developed several predefined policy classes that express the most common data usage restrictions from our IDS use cases. Second, we developed different usage control technologies to enforce these usage restrictions at a technical level. We differentiate between proactive and retrospective technologies. The proactive technologies enforce provisions and obligations across system boundaries during runtime. These technologies control the data usages during runtime and ensure compliant usage with respect to usage restrictions (i.e., prevent data misuse). The retrospective technologies are rather monitoring and recording technologies, but do not prevent or actively control any data usages. Therefore, it does not prevent undesired data usages and is called detective enforcement.

In the following section, we first address the usage control policies, the IDS policy classes, and necessary steps from the specification of usage restrictions to their negotiation between data provider and data consumer. Second, we present the different technologies to enforce usage control (preventive and detective enforcement).

8.3.1 Usage Control Policies

The IDS include several use cases such as IDS Connector, Digital Supply Chain, Smart Urban Mobility, Intelligent Sensor, Interconnected ESN, and many more [12, 13]. These use cases lead to a variety of usage restrictions and finally to a set of formal policies. The policies were categorized into a set of classes in order to define the previously mentioned templates. They are building blocks for expressing data sovereignty requirements with a common understanding such as duration-

restricted or purpose-restricted data usage. Hence, data provider and data consumer have a common understanding of usage restriction that applies to the exchanged data. In addition, the policy classes are expressed in the IDS contract as IDS policy (an ODRL-like language) as part of the IDS Information Model and can be transformed to machine-readable policies that can be technically enforced by usage control technologies such as the MY DATA Control Technologies [7]. We call the technology-independent policies "Specification Level Policies" (SLP) and the technology-dependent policies "Implementation Level Policies" (ILP) [14]. SLPs directly refer to the aforementioned policy classes. ILPs have a direct mapping to the individual usage control technologies for technical enforcement.

As a starting point, the data providers of the IDS need to specify their data usage restrictions as policies. A policy editor or PAP can support the data provider in the policy specification process. It offers various pre-filled templates that refer to classes of data usage control policies. The data providers can choose a template, fill it with the required information, and receive the corresponding IDS policy, which they can use for a usage control technology to enforce the policy in the target system and respectively protect their data.

The policy transformation can be a part of the policy editor (i.e., PAP) or part of the policy instantiation process, usually supported by a PMP. The latter is the most common case as the target system and environment of the data consumer are usually not known during the specification of usage restriction by the data provider.

We will have a look at the policy classes in the next subsection, followed by a subsection presenting the policy negotiation. Finally, we will present the entire process from policy specification by the data provider, negotiation between data provider and data consumer, the transformation of the SLP to ILP, and finally the deployment of ILPs at the data consumer.

8.3.1.1 Policy Classes

A data usage control policy, in general, may provide permission to an IDS data consumer to operate specified action(s) over a data asset or prohibit the operation of that specified action(s). As well, a policy may require the operation of a specified action under specific circumstances. Providing permission or prohibition of an operation is extended to a variety of actions. A policy can be specified to provide permission to use the data. The action of using the data can cover various operations over that piece of data such as displaying it, printing it, making calculations over it, and so on. In addition, a policy may address only a particular fine-grained action. For example, a policy that permits reading data allows the act of obtaining the data asset from the data source without further restrictions; however, the action of printing data is not permitted.

The usage control technologies offer whitelisting, blacklisting, or both approaches to support data sovereignty and to protect the data. In a whitelisting approach, any data usage is denied unless there is a policy allowing the usage.

Contrary to a blacklisting approach, any data usage is allowed unless there is a policy denying the usage.

Thus, it is possible to specify policies that either allow or deny the data usage. Obviously, there might be conflicts among the specified policies. Regardless of the whitelisting or blacklisting approach, a conflict detection and resolution strategy are needed. For example, while a policy provides permission to use the data, another policy might deny the data consumer to print the data. A conflict detection method must realize that the action of printing the data has been permitted once; likewise, it has been prohibited. A conflict resolution method in the context of IDS might always decide to prohibit the action in a case of a conflict; however, the exact definitions of the data usage actions and the concepts of conflict detection and resolution are still evolving.

The IDS supports 21 predefined policy classes such as purpose or role-restricted data usage policies. IDS association members can access the policy classes via the IDSA JIVE [15]. In addition, the IDS Lab offers a PAP to create IDS policies following the aforementioned IDS policy classes. In addition, it includes the transformation from an IDS policy to technology-dependent MY DATA policy.

8.3.1.2 Policy Negotiation

Now that we explained the basics of policy classes, we would like to introduce the concept of policy negotiation. It is a common activity in business to negotiate a (nontechnical) contract between provider and consumer. The same holds for data-driven business in the IDS, but in the end the policy shall be enforced technically. Hence, data provider and data consumer have to negotiate a data usage contract to establish their data economy.

A negotiation process takes care of two aspects: the technical mapping of usage restrictions toward the internal system landscape and the potential bargaining of the usage conditions. The mapping itself targets the challenge of instantiating the stated requirements to decidable features according to the deployed systems. For instance, the data provider is usually not supposed to know the types and variants of the IT architecture of the data consumer. Furthermore, neither the data provider nor the data consumer is willing to reveal more information about their local settings than necessary. However, any automatically enforceable restriction must state the exact parameters which, through a binary decision process, deterministically conclude whether any possible action is allowed or not.

The second aspect of a negotiation step is the bargaining of the actual conditions. When the usage restrictions are specified, the requirements and preferences of data consumers are usually unknown. Following a simple accept or reject pattern drastically reduces the number of potential data consumers and thereby reduces business opportunities. In addition, fixing obligations without knowing the context and implementation details of the potential usage is not reasonable as the information gap between specification and implementation time leads to unforeseen mismatches and conflicts. Therefore, an interested data consumer should be enabled to respond to

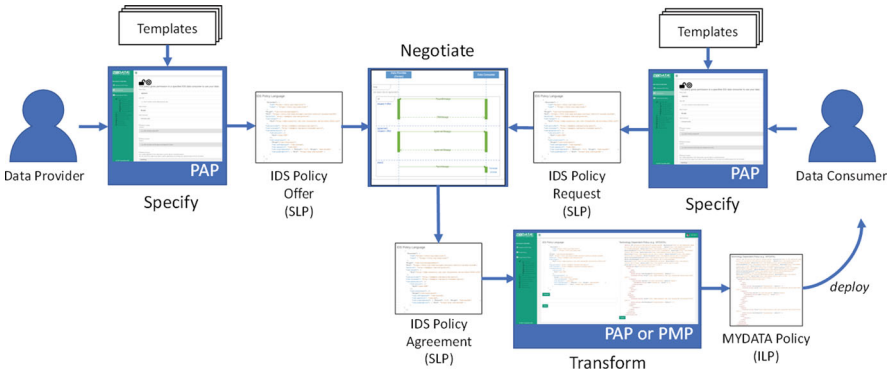


Fig. 8.6 Policy specification, negotiation, and transformation

a usage offering with a slightly adjusted counter offer. Still, it must be always in the authority of the data sovereign to accept or reject the request, or even make an additional offer regarding the details of the received counter-offer.

In Fig. 8.6, we depict the entire negotiation process. We start with the specification of the IDS policy by the data provider using a PAP equipped with the templates. The output of the first processing step is an IDS policy offer. After the specification, we have to negotiate between data provider and data consumer. As part of the negotiation, the data consumer also uses a PAP to specify his or her IDS policy request. If offer and request match, the negotiation terminates, and we will have an IDS policy agreement. The IDS policy agreement is still a technology-independent policy (i.e., SLP) that has to be transformed and instantiated at the data consumer. The transformation and instantiation are usually supported by a PAP or PMP. The result is a technology-dependent policy (i.e., ILP) such as a MY DATA policy (see the next section). The technology-dependent policy can then directly be deployed on a target system at the data consumer to enforce data usage control restrictions.

For more details, we refer the interested reader to the position paper "Usage Control in the International Data Spaces" [16].

8.3.2 Usage Control Technologies

In this section, we briefly present the integration concept for usage control in the IDS and the different technologies.

8.3.2.1 Integration Concept

In the IDS, all communication between data provider and data consumer is handled by connectors. A data provider connects a data source with the connector that

delivers the data to the data consumer connector that in turn connects to a data sink. Data source and data sink can be any kind of system or application. In order to add usage control to this data flow, we integrate control points (i.e., PEPs) to intercept the data flow. At the data provider, we are using a routing pattern, and at the data consumer, we are using an interceptor pattern. In doing so, we ensure to hook into every data flow at the data consumer and apply our data usage control policies. Figure 8.7 illustrates the entire communication flow.

8.3.2.2 MY DATA Control Technologies

MY DATA Control Technologies (MY DATA for short) is a technical implementation of data sovereignty, which represents an essential component for informational self-determination. It is a technical implementation of the conceptual IND^{2}UCE framework for data usage control developed by Fraunhofer IESE.

In general, MY DATA implements data sovereignty by monitoring or intercepting security-relevant data flows. This enables fine-grained masking and filtering of data flows in order to make them anonymous, for example. Compared to classical access control systems, MY DATA can additionally enforce partial filtering and masking of data, context and situation restrictions, as well as restrictions on the purpose of use.

MY DATA offers a decision service (PDP), a management service including a policy editor (PMP and PDP), and an Open Source Software Development Kit for creating control points (PEPs), execution points (PXPs), and information points (PIPs).

MY DATA is integrated in the IDS Connector as part of the Usage Control Container (as illustrated in Fig. 8.7). However, the MY DATA Open Source Software Development Kit allows easy integration in any system.

8.3.3 Logic-Based Usage Control (LUCON)

LUCON (Logic-Based Usage Control) is a policy language for controlling data flows between endpoints. The IDS Trusted Connector uses Apache Camel to route messages between services. LUCON controls how these messages are processed and passed around between services. The LUCON policy language comes with an Eclipse plugin for syntax highlighting, code completion, and compilation into a format that is understood by the LUCON PDP within the Connector.

8.3.3.1 Degree (D°)

While LUCON and MY DATA aim at providing usage control for existing applications and workflows, D° takes another approach. D° is a domain-specific language

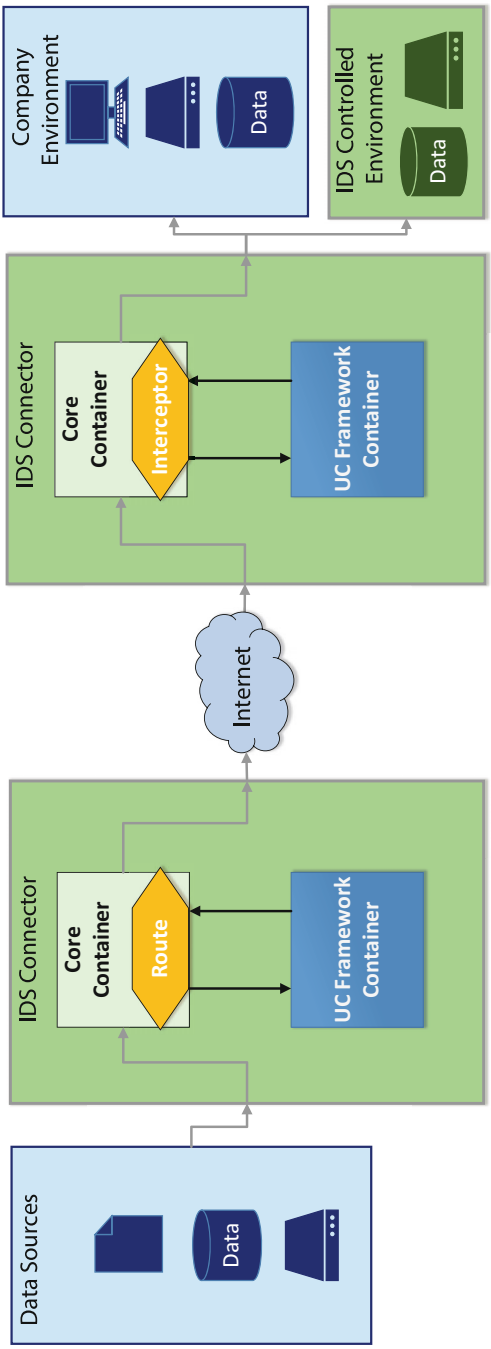


Fig. 8.7 Usage control integration

(DSL) for the development of data processing applications (so-called Data Apps) and takes usage control into account from the beginning of the development. D° uses Java as host language. Through the use of Model-Driven Software Development (MDSD), Data Apps which are developed with D° are transformed into Java applications which are finally compiled into executable applications.

D° is a programming language and runtime environment that orchestrates, isolates, and controls all running data apps. The fulfillment of the IDS policies is ensured by directly integrating them during app compilation into executables.

8.3.3.2 Data Provenance Tracking

While distributed data usage control is concerned with the enforcement of provisions and obligations when exchanging data across system boundaries, the focus of data provenance tracking is on transparency and accountability. Hence, data provenance tracking is closely related, but also complementary to distributed data usage control.

Instead of using PEPs in a proactive manner such as MY DATA, data provenance tracking uses PEPs to passively monitor the data flow. More precisely, it observes, interprets, and logs data transactions and usages for retrospective examination. A data provenance tracking infrastructure can be built upon the same PEPs as distributed data usage control. Furthermore, data provenance tracking does not require a policy specification language, but rather a specification of how observed actions are to be interpreted in terms of data transactions or data usage.

From a technical perspective, data provenance tracking introduces local components to store data flow tracking information (i.e., a provenance storage) offered by PEPs. The provenance storage regularly updates central components (i.e., the provenance collection). The central component can then be equipped with a provenance dashboard to illustrate data flows in human-understandable manner. In addition, we can perform queries on the provenance data for auditing or reporting purposes.

8.4 Conclusion

Data-driven business models are based on sharing and exchanging data. Data sovereignty is a key success factor for data-driven business. In this chapter, we presented how the IDS use usage control concepts and technologies to tackle data sovereignty challenges. The main building blocks are the specification of usage restrictions as IDS policies and their adherence to technical enforcement. Although a lot of work has already been done, it will take a while until digital ecosystems will fully exploit the power of data usage control in order to realize a comprehensive and complete data sovereignty. It is rather a transition process that includes organizational, technical, and legal issues.

In addition, data usage control has to be integrated into and implemented for different positions of the entire digital ecosystem systems to ultimately enable a

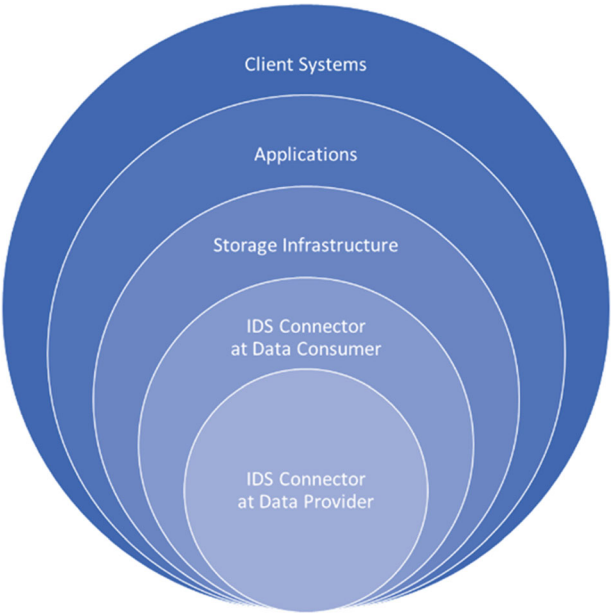


Fig. 8.8 Illustration of the usage control onion

comprehensive control of all data usages. We call this the usage control onion, illustrated in Fig. 8.8. The inner part of the onion represents the starting point for usage control; it is the IDS Connector at the data provider. Every additional onion shell represents an expansion stage for the implementation of data usage control such as integrating usage control technologies into applications or client systems to enable a comprehensive control. There are different architecture options to integrate and implement data usage control [17], resulting in a tradeoff between costs and control capabilities.

References

1. Otto, B., Lohmann, S., Auer, S., Brost, G., Cirullies, J., Eitel, A., Ernst, T., Haas, C., Huber, M., Jung, C., Jürjens, J., Lange, C., Mader, C., Menz, N., Nagel, R., Pettenpohl, H., Pullmann, J., Quix, C., Schon, J., . . . Wenzel, S. (2019). *Reference architecture model for the industrial data space 3.0*. Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. and Industrial Data Space e.V..
2. European Parliament and the Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Brüssel. https://eur-lex.europa.eu/eli/reg/2016/679/oj.
3. Sandhu, R. S. (1993). Lattice-based access control models. *Computer, 26*(11), 9–19.

4. Sandhu, R. S. (1998). Role-based access control. *Advances in Computers, 46*, 237–286. Elsevier.
5. Kuhn, D. R., Coyne, E. J., & Weil, T. R. (2010). Adding attributes to role based access control. *Computer, 43*(6), 79–81.
6. Parducci, B., Lockhart, H., & Rissanen, E. (2013, January 22). eXtensible Access Control Markup Language (XACML) Version 3.0. *OASIS Standard* [Online]. Available: https://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html. Last visited: 16 Aug 2019.
7. MYDATA Control Technologies. (2021). *MYDATA Control Technologies—Developer documentation*. Fraunhofer IESE. https://developer.mydata-control.de. Last visited: 13 Jan 2021.
8. Mont, M. C., & Pearson, S. (2011). Sticky policies: An approach for managing privacy across multiple parties. *Computer, 44*(9), 60–68.
9. Hardjono, T., Maler, E., Machulak, M., & Catalano, D. (2015). User-managed access (UMA) profile of OAuth 2.0. Version 1.0. https://docs.kantarainitiative.org/uma/rec-uma-core-v1_0.html. Last visited: 14 Apr 2021.
10. Kantara Initiative. UMA implementations. https://kantarainitiative.org/confluence/display/uma/UMA+Implementations. Last visited: 14 Apr 2021.
11. Mercer, N. (2018, February 28). *Introducing Windows information protection*. Microsoft [Online]. Available: https://techcommunity.microsoft.com/t5/Windows-Blog-Archive/Introducing-Windows-Information-Protection/ba-p/166564. Last visited: 13 Jan 2021.
12. International Data Spaces Association. (2018). *Use Case overview: International data spaces: Our use cases make it happen!* International Data Spaces Association. https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/Use-Case-Brochure_2018.pdf.
13. International Data Spaces Association. Use cases are IDS in action. https://internationaldataspaces.org/make/use-cases-overview/. Last visited: 16 Apr 2021.
14. Rudolph, M., Moucha, C., & Feth, D. (2016). A Framework for generating user-and domain-tailored security policy editors. In *IEEE 24th international requirements engineering conference workshops, Beijing, China* (pp. 56–61).
15. Hosseinzadeh, A. (2020, October 13). *Policy classes*. Fraunhofer IESE. https://industrialdataspace.jiveon.com/docs/DOC-3412. Last visited: 13 Jan 2021.
16. Eitel, A., Jung, C., Brandstädter, R., Hosseinzadeh, A., Bader, S., Kühnle, C., Birnstill, P., Brost, G., Gall, M., Bruckner, F., Weißenberg, N., & Korth, B. (2021). *Usage control in the international data spaces*. International Data Spaces Association. https://internationaldataspaces.org/wp-content/uploads/IDSA-Position-Paper-Usage-Control-in-the-IDS-V3.0.pdf
17. Zrenner, J., Möller, F., Jung, C., Eitel, A., & Otto, B. (2019). Usage control architecture options for data sovereignty in business ecosystems. *Journal of Enterprise Information Management, 32*(3), 477–495.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 9
Building Trust in Data Spaces



Monika Huber, Sascha Wessel, Gerd Brost, and Nadja Menz

Abstract Data is becoming increasingly valuable and must be protected. At the same time, data becomes an economic asset and companies can benefit from exchanging data with each other. The International Data Spaces enable companies to share data while ensuring data sovereignty and security.

Data providers can keep control over the processing of their data by utilizing usage control policies, including the verification that these usage control policies are enforced by the data consumer. For this, data processing devices, called connectors, must prove their identity and the integrity of their software stack and state.

In this chapter, we present the overall security concept for building trust in data spaces enabling data sovereignty and usage control enforcement. The concept builds on a certification process for components and operational environments utilizing the multiple eye principle. This process is technically mapped to a public key infrastructure providing digital certificates for connector identities and software signing. Finally, the third building block is the architecture and system security of the connectors where usage control must be enforced, the identity and integrity of other connectors and their software stack and state must be verified, and the actual data processing happens.

9.1 Introduction

Data is an important asset for many companies. In particular, data collected during manufacturing or usage of their products could provide insights in business secrets. Such data is therefore often carefully protected and not shared with others. On the

M. Huber (✉) · S. Wessel · G. Brost
Fraunhofer AISEC, Garching, Germany
e-mail: monika.huber@aisec.fraunhofer.de; sascha.wessel@aisec.fraunhofer.de; gerd.brost@aisec.fraunhofer.de

N. Menz
Fraunhofer FOKUS, Berlin, Germany
e-mail: nadja.menz@fokus.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_9

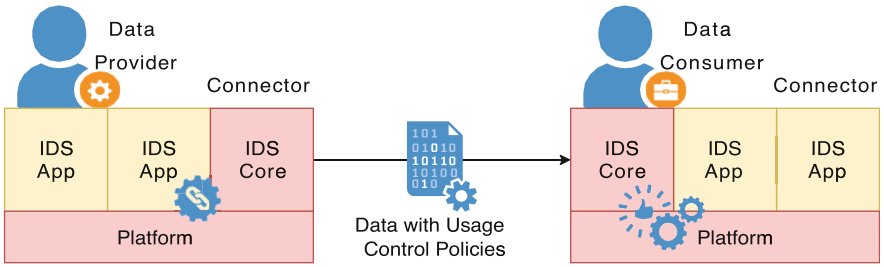


Fig. 9.1 Data exchange in the International Data Space

other hand, companies can benefit from exchanging data with each other and processing combined data sets as a result. Due to the importance of the confidentiality of the data, such a data exchange requires a high trust level between the communication partners.

For that purpose, the International Data Spaces (IDS) enable companies to share data while ensuring data sovereignty and security. A typical scenario for the data exchange in the IDS is depicted in Fig. 9.1. A data provider offers data that is then shared with a data consumer. The data exchange and processing of data is conducted by IDS connectors on both sides of the communication. Further details on this scenario are provided in Sect. 9.2.

In the following, we introduce how trust is established in the IDS ecosystem with the focus on the communication between data providers and data consumers. One key value of the IDS is the possibility for data providers to keep control over the processing of their data by utilizing usage control policies. Data providers can verify that their usage control policies are enforced by their communication partners. For that purpose, connectors can prove their identity and the integrity of their software stack and state. Data consumers can utilize the same mechanisms to ensure the trustworthiness of the data provider. For this, the following three aspects are required in the IDS:

- *Certification Process*: Processes providing verified information concerning the utilized components and involved companies based on an evaluation conducted by independent trusted third parties.
- *Connector Identities and Software Signing*: Mechanisms to technically represent the results from the certification process in order to make the identity and integrity of components verifiable.
- *Connector System Security*: Security of the connector to fulfill the requirements of the certification and to properly conduct the verification of the component's identity and integrity.

In the following, we first describe the overall concept in more detail in Sect. 9.2. Afterward, we explain the certification process in Sect. 9.3, the utilized identity infrastructure and trust model in Sect. 9.4, and the connector system security in Sect. 9.5. Finally, we conclude in Sect. 9.6.

9.2 Data Sovereignty and Usage Control

In the following, we first provide details regarding the communication between data providers and consumers in Sect. 9.2.1. Based on the attacker model described in Sect. 9.2.2, we afterward describe the building blocks for establishing trust and security in the IDS in Sect. 9.2.3.

9.2.1 Data Provider and Data Consumer

One goal of data sovereignty is to allow data owners to control the usage and processing of their data. The IDS offer a solution for companies to keep their data sovereignty even after transferring data to others. The data providers can define usage control policies which the data consumers will have to fulfill. Examples for such policies are the restriction of processing to defined time periods or countries as well as requiring anonymization of data before further processing.

In order to technically implement usage control, both communication partners must implement a trustworthy system that supports the definition of usage control policies (data provider) and the verifiable enforcement of those policies (data consumer). Those systems are called connectors in the IDS.

Figure 9.1 depicts a minimal scenario for a communication in the IDS where one data provider wants to transfer data to one data consumer using IDS connectors. Every connector consists of a Platform providing a runtime to run data processing IDS Apps and implementing system security mechanisms, as well as one dedicated IDS Core App implementing core features for secure connector-to-connector communication. The Platform, the IDS Core, and one data processing IDS App have full access to the confidential data and must therefore be trusted. More details on connector architectures are described in Sect. 9.5.1. The blue gears in Fig. 9.1 symbolize the usage control policies, which are defined by the data provider on the left side, transmitted to the data consumer, and enforced by the data consumer's connector on the right side.

9.2.2 Protection Goals and Attacker Model

A connector must guarantee data confidentiality and the integrity of the connector stack, which implements the enforcement of usage control policies, as its primary *protection goals*. The secondary protection goal is the integrity of the data. Availability is only ranked third, as within the IDS confidentiality is always prioritized higher than availability.

Depending on the business case and the type and amount of data, the data itself can be of different value. Therefore, the IDS consider different attacker models, and

each company can decide against which type of attacker the protection goals must be defended.

In the IDS the following two classes of attackers are addressed:

- *Remote attacker*: The attacker can eavesdrop and modify communication between connectors. The attacker cannot break cryptography if it is state of the art. The attacker does not have access to the connector.
- *Local attacker*: In addition to the previous attacker class, the attacker has full control over the connector. This includes the capability to access unprotected data on the connector or modify unprotected software on the connector. However, the attacker cannot break state-of-the-art cryptography and cannot own the hardware-based root of trust for measurement. The attacker does not have access to private keys, e.g., used for decryption or authentication.

9.2.3 Building Blocks

To reach the protection goals defined in Sect. 9.2.2, the IDS is built on the following three modules depicted in Fig. 9.2 which must fit together and are described in detail in the following sections:

- Section 9.3—Certification Process: The *Certification Process* implements the evaluation of participating companies and connectors utilizing a multiple eye principle to gain verified information regarding the functionalities and security of connectors and companies. In order to address different use cases, the IDS certification offers different security and assurance levels for both types of certification.
- Section 9.4—Connector Identities and Software Signing: In order to technically represent the *Certification Process* described above, the IDS define a Public Key Infrastructure (PKI) for managing the identities of persons (users) as well as devices in the IDS. The user certificates are used to sign the results from the



Fig. 9.2 Building blocks for building trust in data spaces

certification process in the form of company descriptions and software manifests. Together with the device certificates, those signed descriptions are used to prove the identity and integrity of the IDS connectors.

- Section 9.5—Connector System Security: The *Certification Process* defines numerous security requirements for IDS connectors. Depending on the architecture of the connector, different components might have access to the data. All these components must fulfill these requirements and must be evaluated by trusted third parties. In order to ensure that a communication partner really utilizes those certified software components, the IDS offer an attestation protocol for remote identity and integrity verification. This protocol uses the device certificates, company descriptions, and software manifests introduced before.

9.3 Certification Process

The IDS utilize a certification scheme [1] to get confirmed and verifiable information on the connectors and the companies operating them. The underlying principle for this certification is described in Sect. 9.3.1. Afterward, we provide details concerning the *Component Certification* focusing on the connector implementations in Sect. 9.3.2 and the *Operational Environment Certification* for the companies operating IDS components in Sect. 9.3.3.

9.3.1 Multiple Eye Principle

All connectors as well as their operators must provide a sufficient level of (data) security and must correctly implement the IDS standards. Participants need to be able to trust that other participants previously unknown to them truly fulfill these requirements. Therefore, the IDS utilize a certification scheme [1] which ensures that the needed information concerning the communication partners is verified by multiple independent parties. The procedure for certification is depicted in Fig. 9.3 with

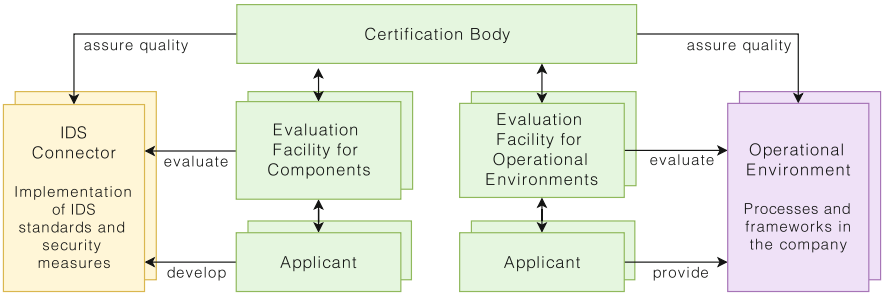


Fig. 9.3 Certification process for components and operational environments

the left side showing the certification of IDS connectors and the right side illustrating the operational environment certification. As the figure shows, the three involved parties and the procedures for the two different certification types are equivalent. The roles of the depicted entities and their responsibilities in the certification process are described in the following:

- The *applicant* is the person or organization that starts the certification process by contracting an evaluation facility and providing it with all necessary information and evidence for conducting the evaluation. For components this role is typically filled by the developer and for operational environments by (a representative of) the company that wants to operate an IDS component.
- An *evaluation facility* is responsible for assessing whether a connector or operational environment fulfills the specified requirements for the intended security level. The evaluators perform a thorough evaluation of the provided information following a standardized evaluation procedure. As a result, they provide an evaluation report to the certification body which describes the conducted evaluation work and its results. The concrete depth and scope of the evaluation depend on the desired level of security and assurance as described in Sects. 9.3.2 and 9.3.3.
- The IDS *certification body* is responsible for ensuring the comparability and quality of all conducted evaluations. For that purpose, the certification body must assess the competence of evaluation facilities and approve them before they are allowed to conduct evaluations in the IDS. For each conducted evaluation, the certification body reviews the provided evaluation report to ensure the correct execution of the evaluation. The certification body makes the final decision about the award or denial of each certificate.

9.3.2 Component Certification

Connectors play an essential role in the IDS. They are responsible for executing the communication and data processing in the IDS and must therefore fulfill many requirements concerning functionality and security. Consequently, the certification of IDS connectors focuses on interoperability and security. In order to address the different use cases in the IDS, the certification is designed to offer different security profiles and assurance levels as depicted in Fig. 9.4.

The three security profiles define an increasing set of security measures and requirements:

- The *Base Security Profile* defines a set of minimal requirements for participating in the IDS which focus on protection against a remote attacker. It allows companies to try out the IDS and can be used for handling open data.
- The *Trust Security Profile* defines more thorough security requirements to be fulfilled. The goal is to protect data from a remote attacker and provide all necessary means to realize usage control on the connector. However, for local

attack scenarios the profile only aims to prevent accidental misuse from an administrator but does not consider the administrator as a potential attacker.
- The *Trust + Security Profile* fills this gap. It has almost equivalent (security) requirements as the *Trust Security Profile*, but in addition aims to keep all the required capabilities and assurances against a powerful local attacker such as a malicious administrator.

The criteria catalog for these three security profiles [2] combines IDS-specific functional requirements aiming to achieve conformity with the IDS Reference Architecture Model [3], requirements from the industrial security standard IEC 62443-4-2 [4], and requirements for secure software development.

In addition to security profiles, the IDS certification offers three assurance levels that define the depth of the evaluation:

- As a low access barrier, the IDS certification allows a *Checklist Approach* as an assurance level for the *Base Security Profile*. The applicant fills out a questionnaire covering all applicable requirements and an automated test suite is used to run a set of interoperability and security tests. No evaluation facility is involved, but the questionnaire and the result from the automated test suite are reviewed by the IDS Certification Body.
- For a *Concept Review*, the applicant is required to provide the evaluation facility with detailed documentation for the connector as well as a working instance of the connector implementation. The evaluation facility conducts a review of the provided concepts and security measures as well as practical testing concerning the functionality and security of the implementation.
- The *High Assurance Evaluation* builds on the *Concept Review* evaluation and extends it with in-depth source code reviews and an on-site visit to the development site of the connector.

The described assurance levels and security profiles can be combined into six possible ways as depicted in Fig. 9.4. Whenever applicants want to get a connector certified, they can decide which security profile and assurance level they want to



Fig. 9.4 Connector certification levels

target. In reverse, each IDS participant can decide which expectation it has on the certification of its communication partner's connector as detailed in Sect. 9.5.2.

9.3.3 *Operational Environment Certification*

In addition to the certification of technical components, the IDS certification scheme [1] defines an operational environment certification that is required for each company offering connectors in the IDS. It requires companies to fulfill defined standards for management processes and infrastructure in order to ensure a secure operation of IDS components. Analogous to the connector certification, the operational environment certification is offered with different security and assurance levels as depicted in Fig. 9.5.

Depending on the role a company wants to fulfill in the IDS, there are different security levels with increasing requirements:

- The *Entry Level* is meant for companies that want to give the IDS ecosystem a try and gain access with minimal effort. Thus, this level only includes minimal requirements for the procedures and security mechanisms in the company.
- The *Member Level* defines the standard requirements that are meant to be met by all companies participating in the IDS. This level is recommended for data providers and data consumers introduced in Sect. 9.1.
- The *Central Level* is designed for companies that take on specific trust-building tasks in the IDS. An example is the issuing of identity certificates described in Sect. 9.4.1. These companies must meet a more extensive set of requirements than member-level participants, since any security breaches on their side could affect broad parts or even the whole IDS ecosystem.

In order to reduce the effort and costs for the certification, the requirements are based on existing schemes and standards for information security (management), namely, IEC 27001 [5] and BSI C5 [6].

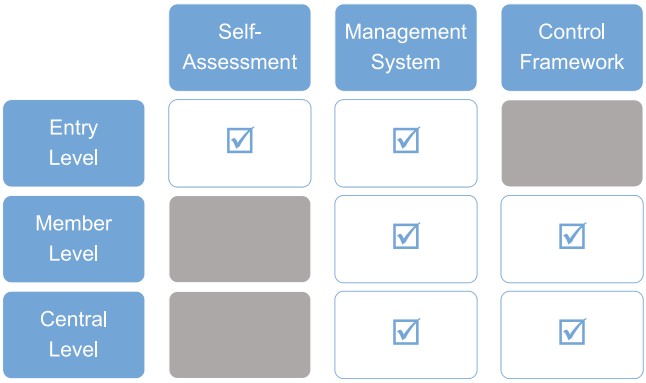


Fig. 9.5 Operational environment certification levels

In addition to security levels, the IDS certification offers three assurance levels that define the depth of the evaluation:

- As a low access barrier, the IDS certification allows the *Self-Assessment* as an assurance level for the *Entry Level*. The applicant fills out and signs a self-assessment concerning the fulfillment of the defined requirements, and, without involving an evaluation facility, the IDS Certification Body reviews the document.
- For a *Management System* evaluation, the evaluation facility assesses whether a company has defined the processes and structures necessary and whether the employees are aware of and following those processes. The evaluators review documentation and process definitions, conduct interviews, and audit the company during on-site visits.
- Companies can utilize control frameworks in their company to ensure and document adherence to the defined management processes. After a certain amount of time after such control mechanisms have been established at the company, their correct application can be verified by a *Control Framework* evaluation. For that purpose, the evaluation facility conducts a *Management System* evaluation as described above and additionally reviews the control mechanisms and evidence for the application of these mechanisms based on random samples.

The described assurance levels and security profiles can be combined in six ways as depicted in Fig. 9.5. Whenever applicants want to get certified, they can decide which security and assurance level they want to address. In reverse, each IDS participant can decide which level they expect from their communication partners as detailed in Sect. 9.5.2.

9.4 Connector Identities and Software Signing

The IDS connect different organizations and components with each other. The trust required for the conducted communication is established by a PKI that manages identities for users and devices and is utilized to implement a technical representation of the certification process described previously in Sect. 9.3.

Figure 9.6 shows the PKI which builds the identity infrastructure necessary for achieving the following purposes:

- Create and manage digital identities for persons, e.g., for employees of an evaluating company.
- Create and manage digital identities for devices.
- Assign ecosystem roles to person identities, e.g., as evaluator of components.

The digital identities of users and devices are bound to the possession of a private key matching the public key contained in their X.509 identity certificate. In the depicted identity ecosystem, the *User Certificate Authority (CA)* issues user

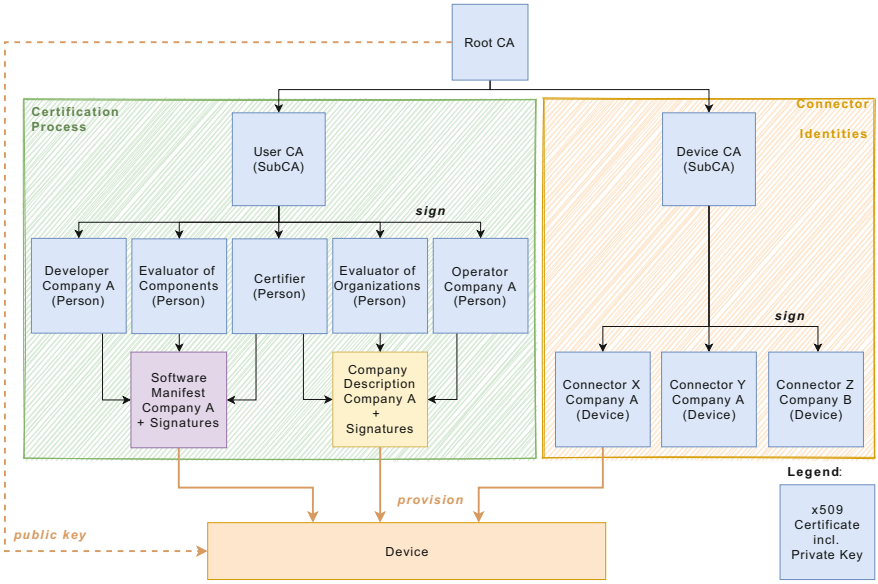


Fig. 9.6 Public Key Infrastructure for building trust in data spaces

identities and assigns ecosystem roles as described in Sect. 9.4.1. The *Device CA* issues device identities which are combined with signed company descriptions as described in Sect. 9.4.2. The signed software manifests necessary for ensuring the connector integrity are detailed in Sect. 9.4.3.

The combination of device identities, signed company descriptions, and software artifacts enables building of trust in the ecosystem. Every device is provisioned with its device certificate, a company description of the operator, the related software manifests for all deployed components, and the public key of the *Root CA* to validate certificates of other party's devices. Optionally, there might be more than one *Root CA* in the ecosystem.

9.4.1 Technical Implementation of the Certification Process

The certification process described in Sect. 9.3 is used to ensure that all components and organizations in the IDS adhere to defined standards. For the verifiability of successful certifications, the IDS include a technical implementation of the certification process that is depicted in the green box in Fig. 9.6. Every person participating in this process receives an identity certificate and owns a private key that allows signing of software manifests and company descriptions. A successful certification process is proven by having these information artifacts signed by multiple parties. As an example, the steps for the component certification are detailed in the following:

- The developer of a software artifact signs the initial software manifest in its role as developer and certification applicant.
- The evaluator of the component signs the software manifest after successful evaluation.
- The certifier signs the software artifact after validating the evaluation report.

The certification for operational environments is conducted in a comparable way to receive a company description with multiple signatures.

The process is designed to create a chain of trust rooted in the unforgeable identity of the applicant. The additional layers of trust are created by an evaluator who verifies that all requirements for the desired trust level are satisfied. Another layer is added by the certifier that ensures the quality and conformity of the evaluation.

9.4.2 Connector Identities and Company Descriptions

Each connector needs to be able to uniquely identify the other connectors in the IDS. For that purpose, each connector instance is provided with a technical identity. This identity is issued by a dedicated *Device (Sub)*CA. Each device is owned by a specific company. The device certificate creates the link to the company description through the company identity. This company description ensures a trustworthy operational environment and is described in the next section. Certified software components can be deployed on multiple devices. Overall trust relies on a combination of device identities, company descriptions, and software artifacts including software signatures.

The company responsible for hosting a connector is described in a company description that has been validated and signed by three independent parties, i.e., the applicant, evaluator, and certifier. It must, at minimum, contain these items:

- The company name.
- The location of the head or branch office.
- The certification level derived from the operational environment certification process.
- The expiry date of the company description.
- An endpoint which may be queried to get a current status of the company description to allow the revocation of the description.

9.4.3 Software Signing and Manifests

For assessing the trustworthiness of the software running on another connector, it is necessary to get information regarding the software stack and the certification it passed. Thus, each software component on a connector needs to be uniquely identifiable using a software manifest that has been validated and signed by all

three parties directly involved in the component's certification process. This manifest contains details to allow an assessment of the components and utilizes cryptographic hashes to uniquely identify the software component belonging to it. In detail, the software manifest must contain at least the following items:

- Cryptographic hash(es) to identify the components which are described by this manifest.
- A unique identifier for the developer of this software artifact.
- An artifact identifier that allows tracking of different versions.
- A version number.
- A classification of the artifact type, e.g., boot loader, kernel, and protocol adapter.
- Functionalities provided by this artifact.
- Usage control policies that can be enforced, e.g., deletion of data after a specified time period.
- The certification level derived from the component certification process.
- The expiry date of the software manifest.
- An endpoint which may be queried to get the current status of the manifest to allow the revocation of the manifest.

Based on the manifests, the software integrity of all component parts becomes verifiable to other connectors and the communication partners are able to assess the risk and possible consequences of sharing their data with this connector.

9.5 Connector System Security

In addition to the certification process as well as the connector identities and software signing, the third building block is the system security of the connectors where data is processed, and usage control must be enforced.

The four core functionalities of a connector are:

1. Providing a *runtime* for isolated data processing apps implementing the IDS information model.
2. *Communication* between connectors with authentication and remote integrity verification.
3. Persistent *storage* of confidential and integrity protected data.
4. *Usage control enforcement.*

These functionalities are implemented in a set of core components depending on the device type (embedded system, gateway or edge device, server, virtual machine, cloud) and the use case with its required security level. Details on possible architectures are provided in Sect. 9.5.1. All components with access to the asset *data* must be trustworthy and their trustworthiness must be remotely verifiable by other connectors before transmitting the data. The utilized communication protocol is defined in Sect. 9.5.2.

9.5.1 Trusted Computing Base

The Trusted Computing Base (TCB) of a connector is the set of all hardware and software components that are critical to the confidentiality and integrity of the transmitted and processed data. Typically, only one or at least a few vulnerabilities suffice to break the system's security. Therefore, a common goal is to reduce the complexity and the size of the TCB which also results in a reduced evaluation effort in the course of the certification process described in Sect. 9.3.

Figure 9.7 shows four common connector system architectures. The system architectures on the top use process isolation mechanisms to protect the data processing apps and IDS functionalities of a connector. The system architectures below use system virtualization with a hypervisor for isolation. The system architectures on the right side make use of an exemplary hardware feature to reduce the TCB. Components in *blue* directly process the data to be protected. Components in *red* are part of the TCB and components in *gray* are not part of the TCB. In the following, we first introduce the components followed by an introduction to the four common connector system architectures.

The components of a connector are as follows:

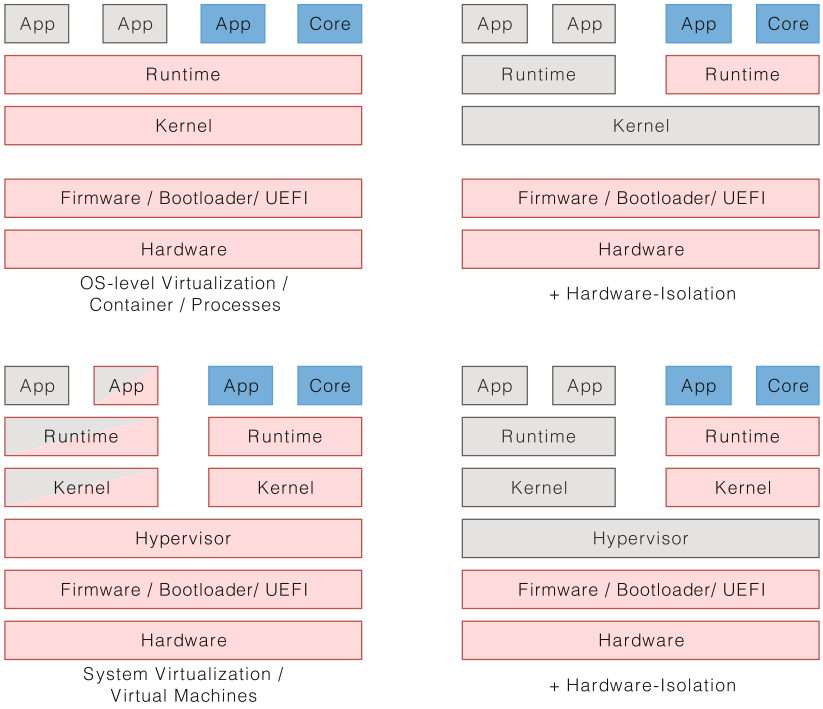


Fig. 9.7 Trusted computing base for IDS Apps and Core

- *Apps* implement all functionalities that are not needed to start other apps. This includes core functionalities required on all connectors and use case-specific features.
- The *Core* is a specific app. It implements the communication interfaces for data exchange with other connectors. This includes an implementation for remote attestation described in Sect. 9.5.2. Encrypted communication between connectors terminates here.
- The *Runtime* implements functionalities to bootstrap apps in a system. It might be small, implementing only a minimal set of functionalities used by the apps. The runtime is a user space component and always part of the TCB.
- The *Kernel* connects user space software to the hardware of a connector. It typically implements memory management, resource management, and device management. User space components use *system calls* to interact with the kernel.
- A *Hypervisor* allows to run multiple *virtual machines* on one physical machine providing a standardized interface to the kernel. In the IDS this may have advantages regarding the deployment, migration, and management of connectors.
- *Firmware/Bootloader/UEFI* includes all software components to bootstrap a system and to initialize hardware before the kernel is started. In interaction with the hardware, these software components bootstrap the trust in the system by implementing a trust anchor and building the root of trust.
- *Hardware* includes all hardware components in the system with access to unencrypted data of apps. This especially includes the processor. It might include a Secure Element (SE), e.g., implementing a Trusted Platform Module (TPM) [7]. It usually does not include hardware devices for persistent storage of data or network interfaces. It might not include Random Access Memory (RAM) if data is always stored encrypted. The hardware must always provide functionalities to bootstrap the trust in the system.

The four connector system architectures presented in Fig. 9.7 are:

- *OS-Level Virtualization/Containers/Processes*: The top-left subfigure shows a typical architecture using OS-level virtualization, i.e., containers. In this case, the *Runtime* includes user space processes running in the root namespace on top of the *Kernel*. The TCB includes *Apps*, *Core*, *Runtime*, *Kernel*, *Firmware/Bootloader/UEFI*, and *Hardware*.
- *OS-Level Virtualization/Containers/Processes + Hardware isolation*: The top-right subfigure shows a possibility to reduce the TCB of the previous architecture using a Trusted Execution Environment (TEE) like Intel Software Guard Extensions (SGX) [8] or ARM TrustZone [9]. RAM used by the TEE is encrypted respectively only accessible by the software components inside the TEE. Functionalities like drivers for persistent storage or network interfaces are kept outside of the TCB. In this case, the TCB includes *Apps*, *Core*, the *Runtime* inside the TEE, *Firmware/Bootloader/UEFI*, and *Hardware*.
- *System Virtualization/Virtual Machines*: The bottom-left subfigure shows a typical architecture using system virtualization, i.e., virtual machines running on top of a hypervisor. In this case, the TCB includes *Apps*, *Core*, *Runtime*, *Kernel*,

Hypervisor, *Firmware/Bootloader/UEFI*, and *Hardware*. Depending on the implementation, the hypervisor might be integrated into a *Kernel* or one virtual machine might be privileged in the system. In such cases, the *Kernel*, *Runtime*, and one or more *Apps* of this privileged virtual machine might have access to the data processed in the virtual machine running the IDS *Apps* and *Core* and therefore might be part of the TCB as well.

- *System Virtualization/Virtual Machines + Hardware isolation*: The bottom-right subfigure shows a possibility to reduce the TCB of the previous architecture using extensions like Intel Trust Domain Extensions (TDX) [10] or AMD Secure Encrypted Virtualization (SEV) [11]. RAM used by the virtual machine is encrypted respectively only accessible by the software components inside the virtual machine. In this case, the TCB includes *Apps*, *Core*, the *Runtime*, *Kernel*, *Firmware/Bootloader/UEFI*, and *Hardware*.

9.5.2 Remote Attestation

In Sect. 9.3 we described the certification and evaluation process of components and operational environments, and in Sect. 9.4 we described how this is mapped to a PKI, manifests, and cryptographic signatures. Afterward we described in Sect. 9.5.1 which components of a connector are critical to its security. The final step is the verification of the identity and trustworthiness of the connector's TCB before transmitting data to it. To achieve this, both communicating connectors perform a remote attestation before any data is exchanged.

Figure 9.8 shows a simplified sequence diagram for the connection establishment up to the first data exchange. The following steps are performed:

- *Hello Message (incl. Nonce)*: In the first step, a hello message including a random number is sent. This number is used to guarantee freshness for the following messages and, thus, prevent replay attacks.
- *Proof of Connector Identity*: In the next step, the connectors identify themselves by using their X.509 certificate and their company description. The private keys of their identity certificates are used to establish the communication channel. Both the signatures of the connector-specific X.509 certificates and the signatures of the company description can be verified up to the *Root* CA. After this step, the identity and the operational environment certification level are known.
- *Proof of Connector Integrity*: In this step, the integrity of both connectors is verified. Each connector implements one *prover* to prove its integrity and multiple *verifiers* to verify the integrity of all sorts of connectors. The prover sends all signed manifests describing its connectors TCB and a prover-specific description of the system state. This typically includes a hash chain beginning with the root of trust for measurement followed by measured hash values representing all components in the secure boot chain. This information is then signed by the prover with its private key in a secure manner. The verifier will then verify the signature,

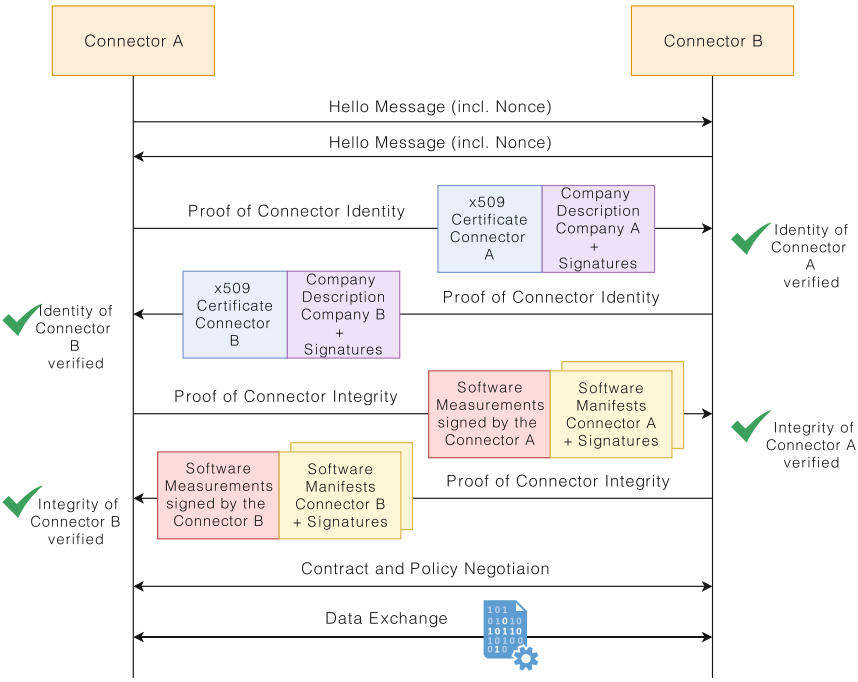


Fig. 9.8 Communication between connectors in the IDS

the freshness, and the hash chain of the software artifacts. Afterward, for each artifact the manifest including the signatures from the developer, evaluation facility, and certification body is verified up to the *Root* CA. After this step, the trustworthiness and the component certification level of the whole TCB are known.

- *Contract and Policy Negotiation*: Based on the established secure and trustworthy communication channel, the two communication partners can negotiate the (legal) terms for data exchange and the required usage control policies.
- *Data Exchange*: In the final step, data can be exchanged.

9.6 Conclusion

In this chapter, we presented the overall security concept for building trust in data spaces enabling data sovereignty and usage control enforcement. We defined data confidentiality and the integrity of the connector stack, which implements the enforcement of usage control policies, as our primary protection goals. To reach these goals, we introduced the three building blocks:

- Certification process.
- Connector identities and software signing.
- Connector system security.

The IDS certification process described in Sect. 9.3 implements a multiple eye principle to gain comparable and trustworthy information concerning the fulfillment of certification requirements for connectors and companies. Both types of certification are technically mapped to a PKI and digital signatures for company descriptions and software manifests in Sect. 9.4. In combination with an X.509 identity certificate for each connector instance, the company description can be used to prove the identity of the connector and its operating company. Additionally, the usage of certified software can be proven to other connectors by using remote attestation utilizing software measurements and signed software manifests. Based on the verification of the identity and integrity of the communication partner, each participant in the IDS is enabled to sovereignly decide with whom and under which conditions they want to share their data. Finally, we have shown multiple connector architectures with small TCB in Sect. 9.5 to reduce costs for the evaluation and certification.

References

1. IDSA. (2019a). *Framework for the IDS certification scheme, Version 2*.
2. DIN. (2020). *DIN SPEC 27070:2020–03 Anforderungen und Referenzarchitektur eines Security Gateways zum Austausch von Indu triedaten und Diensten*.
3. IDSA. (2019b). *IDS reference architecture model version 3.0*.
4. ISO. (2019). *IEC 62443–4-2:2019 Security for industrial automation and control systems - Part 4–2: Technical security requirements for IACS components*.
5. ISO. (2013). *IEC 27001:2013 Information technology — Security techniques — Information security management systems — Requirements*.
6. BSI. (2018). *Cloud Computing Compliance Controls Catalogue (C5)*.
7. TCG. (2014). *Trusted Platform Module (TPM) 2.0*. Library Specification, Family 2.0, Level 00, Revision 01.16.
8. Costan, V., & Devadas, S. (2016). *Intel SGX explained* [Cryptology ePrint Archive, Report 2016/086].
9. Pinto, S., & Santos, N. (2019). Demystifying arm TrustZone: A comprehensive survey. *ACM Computing Surveys., 51*, 1–36.
10. INTEL. (2020). *Intel Trust domain extensions. Technical report.*
11. AMD. (2020). *Secure encrypted virtualization API version 0.24. Technical report.*

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 10
Blockchain Technology and International Data Spaces



Wolfgang Prinz, Thomas Rose, and Nils Urbach

Abstract The core objective of the concept of International Data Spaces (IDS) is to enable controlled exchange and sharing of data between organizations, regardless of the type of data. Sharing of data will generate services that become an asset while data providers maintain their sovereignty. IDS furnish a technology enabler for implementing data economies to exchange data and knowledge, which are according to usage policies. Thus, data turns into an economic asset. However, once data have been provided toward IDS, sovereignty of data owners is of pivotal importance, as well as the question of its use and the transfer of incentives to providers. At this point, blockchain technology enters the ballpark. It is instrumental for the implementation and operation of clearing houses as trading platform for data provision and knowledge utilization. The aim of this chapter is to examine and discuss the role of blockchain for IDS. Next to general blockchain foundations and potentials, blockchain's specific potential for IDS is discussed and its application is demonstrated by four compelling use cases.

10.1 Introduction

The International Data Spaces (IDS) concept is a virtual data space leveraging accepted standards, technologies, and governance models of the data economy to facilitate secure and standardized data exchange and data linkage in a trusted business ecosystem. It provides the foundation for creating smart-service scenarios

W. Prinz · T. Rose
Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

RWTH Aachen University, Aachen, Germany
e-mail: wolfgang.prinz@fit.fraunhofer.de; thomas.rose@fit.fraunhofer.de

N. Urbach (✉)
Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

FIM Research Center, Frankfurt University of Applied Sciences, Frankfurt am Main, Germany
e-mail: nils.urbach@fit.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_10

and enabling innovative cross-organizational business processes, while at the same time guaranteeing sovereignty for data providers [1]. Hence, IDS is the technology enabler for the monetization of data [2].

The core purpose of IDS is to enable controlled exchange and sharing of data between organizations, regardless of the type of data (e.g., structured data, streaming data). In IDS use cases, two basic data sharing patterns can be identified. First, data is shared to feed new, data-driven services, such as apps, smart algorithms, and other digital services. Second, data is shared for some form of business process synchronization such as using the data to execute transactions, enable production, or synchronize processes [1]. The generation of knowledge is one specific type of service that also builds upon a shared data space. This knowledge can be utilized to improve production processes once generated. PigFarm illustrates a vital life cycle of knowledge for the optimization of pork production processes [3]. Information does become the currency of the digital age as pointed out earlier by Carly Fiorina.[1] In many of these cases, sharing of data enables transactions with data as services, which become shared assets with responsibility for the participating organizations [1]. Data analysis for the generation of knowledge is one specific instance of this case. The question arises on how the generation and trade of such data-driven assets are monitored and how incentives toward knowledge producers as well as invoice users are rewarded. Thus, accounting methods come up as one particular use dimension of blockchain for IDS.

Methods for an accounting and clearing of data and knowledge utilization would complement the already existing capabilities of the IDS architecture to share (potentially large) datasets with the help of IDS connectors. Blockchain has the potential to ensure data consistency and transparency in combination with the general IDS approach for data sovereignty as well as secure data exchange and sharing owners [4] in order to turn data into business assets [2, 5].

The aim of this chapter is to examine and discuss the role of blockchain technology for IDS. Therefore, we introduce the general foundations of the blockchain technology, describe the different design parameters for setting up blockchain implementations, explain the concept of smart contracts, and outline the general potential of blockchain. On this basis, we discuss the potential role that blockchain can play for IDS. We then illustrate the application potentials of blockchain for IDS by referring to four compelling use cases. Emphasizing the important role that blockchain can play to fully leverage the potentials of many IDS use cases.

Throughout the course of this chapter, we use the term *blockchain technology* consistently. Yet, blockchain is just a specific instance of *distributed ledger technology (DLT)*, which spans a wider spectrum of implementation technologies such as data structure for the ledger. The term DLT also expresses the nature of the distribution of ledgers across stakeholders to be more tangible and intrinsically

[1] Carly Fiorina, former CEO of Hewlett-Packard, Oracle Open World, San Francisco, 2004 http://www.hp.com/hpinfo/execteam/speeches/fiorina/04openworld.html

includes smart contracts as means for process automation. Basically, both terms can be used interchangeably in the context of this chapter.

10.2 Blockchain Technology

This section addresses the core concepts of blockchain technology and exposes specific characteristics that distinguish it from common data management methods.

10.2.1 Basic Concept

A blockchain is a distributed data structure that manages transactions transparently, chronologically, and immutably in a computer network. The core concept of databases with its centralized governance of transaction management in terms of ACID (atomicity, consistency, integrity, and durability) is replaced by maintaining data distribution [6].

A blockchain is a chronologically ordered chain of blocks, in which each block contains information about valid network activity since the addition of the previous block. The individual blocks are connected to each other by cryptographic hashes, creating a data structure that cannot be changed afterward. Each block contains a block header in addition to transactions within the network. The block header contains a reference to the previous block in the block chain, a timestamp, an arbitrary string (nonce), and the root of a Merkle tree, which represents a data structure that efficiently aggregates all transactions contained in the block via hashes [7, 8].

The blockchain together with the associated peer-to-peer network can be described as a blockchain system. Such systems use cryptography and peer-to-peer principles instead of a central authority to achieve a network-wide verification of the system's status by consensus [9]. The general transaction processing of a blockchain system can be illustrated by the example of the Bitcoin system. Here, a network node first sends a message combined with a digital signature including transaction details to the entire network. Network nodes that receive the message then check for duplicate use of the Bitcoins and then propagate the transaction and the evaluation result further in the network [10]. To coordinate the consensus, i.e., the common status of the current blockchain, the network nodes use a so-called proof of work mechanism [11]. In general, the proof of work (PoW) is intended to prevent excessive or improper use of a service [12]. The provision of a PoW requires a certain amount of effort; in the case of Bitcoin, a computationally intensive problem must be solved. A hash value must be changed until it matches a certain pattern [10, 13].

10.2.2 Design Parameters

Blockchain systems can be distinguished regarding several design parameters [14]. First, blockchain systems can be classified according to whether they are private or public [15]. The decisive factor here is by whom the systems can be used, i.e., who has access to the data or can propose new data inputs (e.g., transactions)? If this use is permitted to anyone, it is a public system. However, if it is restricted to a specific user group—e.g., an organization or consortium—the blockchain system is considered private [15]. Another possible distinction between different system types is whether permission is required to participate in the administrative process of the blockchain. In the permissionless Bitcoin system, this process is basically allowed to anyone without approval. However, if the network nodes that perform the corresponding validation are pre-selected by a consortium or a central authority, the system is a permissioned blockchain system [15]. Finally, blockchain systems can be distinguished by the way in which a consensus on the system status is reached. In addition to the PoW mechanism employed by Bitcoin, a number of alternative consensus mechanisms exist. An alternative approach is the use of a proof of stake. The basic idea here is that the blockchain is updated primarily by network nodes that hold a large share of the currency or other values in the blockchain, which creates an incentive for correct maintenance of the system [16]. Further alternatives are the use of a proof of activity [17], a proof of publication [18], or a proof of storage [16]

10.2.3 Smart Contracts

A blockchain is not just a distributed transaction manager, but a special form of database management system [6]. It can also be utilized for the automation of processes and in particular an automation of cooperation logics in terms of scripting languages. Such scripts coined smart contracts can be programmed to run arbitrary application logics on the nodes of the P2P network [19]. As early as 1997, the concept of Smart Contracts was introduced by Nick Szabo [20] and defined as a computer-based transaction protocol that implements the terms of a contract [21]. Blockchain technology for the first time offers suitable means for implementing such contracts. Smart contracts can be understood as computer programs that conduct computations once certain conditions are met [22]. To this end, the smart contract can use external information as input, which then causes a certain action to be taken via the rules defined in the contract. The corresponding scripts with the contract details are stored in a specific address of the blockchain for this purpose. If the specified external event occurs, a transaction is sent to the address as trigger, whereupon the terms of the contract are executed accordingly [8]. Therefore, smart contracts are instrumental for maintaining dependencies between data and processes consistently.

Smart contracts are therefore an effective means for the automation of interactions among systems, agents, and the environment. Contracts can be executed, enforced, verified, and inhibited by means of algorithms without any interventions and control by intermediaries. The range of possible applications is therefore very broad. For example, properties such as cars, bicycles, or apartments could be rented out via a smart lock and blockchain system without physical key transfer. To do this, the owner sets the deposit and rent in the smart contract. Therefore, the correctness of smart contracts, i.e., how to automize the cooperation's logic, is essential [23]. The smart contract also contains rules for access and usage authorization (e.g., the user can only open the lock after paying the deposit and rent). All interactions with the blockchain system, such as making payments, exchanging the digital key, or opening and closing the smart lock, can be performed by the tenant and user via smartphone. Incoming payments, authorization distribution and management, as well as deposit repayments are carried out transparently, securely, and unalterably via the blockchain [8]. Moreover, any type of information logistics and digital ecosystems can be maintained consistently due to process automation by smart contracts.

10.2.4 Opportunities of Blockchain Systems

Blockchain potentially offers at least a revolutionary element of change for the development of the Internet as a digital infrastructure. The first step was the Internet in its well-established understanding as "Internet of information." Recently, the "Internet of things" has emerged through the increasing networking of intelligent devices. However, transfer of assets is limited to those that are available in digital format such as data about container transports on inland waterways. A transfer of values—such as contract transaction for the transfer of physical premises—was only possible by involving a trustworthy third party and the use of common ledger technologies in public administrations. With the introduction of blockchain, value transactions without dependence on or trust in a third party have now become possible, the so-called the Internet of values and trust. Hence, knowledge about processes and their improvements can also become a valuable business asset with governance for its management and control of transactions.

In principle, blockchain technology can be used in the form of three generic roles: firstly, as an enhancer for optimizing existing processes that are already being handled digitally or non-digitally without intermediaries via bilateral (peer-to-peer) interfaces, *optimization of operational businesses*; secondly, as a transformer to streamline processes that were previously carried out with the involvement of traditional intermediaries, *change governance of partner networks*; and thirdly, as an enabler to foster collaborations that were previously not feasible due to the lack of an intermediary for organizational and technical integration, *innovate business partnerships to generate new markets*.

Possibilities for implementing new and innovative services and different application solutions are manifold, but certain patterns can be identified as landmarks: "neutral platform for business collaboration," "tamper-proof documentation," "payment transactions," "management of cross-organizational processes," "digital identity," "digital documents," "services without service providers," and "economically autonomous machines" [14, 24].

The use of a blockchain technology for a particular use case is often justified economically and organizationally. The identification of technological evidence is often rather sparse from an application [25] as well as a governance perspective [26]. Moreover, performance and scalability of blockchain-based solutions are, at least up to now, technically inferior to a centrally organized system with well-known database technology for the management of operational data that show up in production environments with high frequencies. Established database technology is certainly the asset for efficient data management, while blockchain technology can maintain audit records for the correctness of processes, i.e., *database technologies allow an efficient management of production data while blockchain technology offers auditing services for the correctness of business processes*. Moreover, blockchain technology offers the possibility to digitally implement processes that could previously only be realized by involving a trustworthy third party.

10.3 Blockchain in International Data Spaces

So far, potentials of blockchain technology from an application- and process-oriented point of view have been identified previously. Moreover, a distribution of concerns between an efficient management of operational data by DB (database) technology and an auditing of business processes by BC (blockchain) technology has been clarified. The following technical properties of a blockchain are relevant for the identification of its potential for accompanying an IDS:

- Information and transactions stored in a blockchain are irreversible and cannot be deleted or modified. Immutability of collaborations conducted is granted.
- Consensus about the information to be agreed upon is achieved within a public or permissioned network. Hence, partners and potential stakeholders can team up in an open fashion and they are no longer limited to organizational frameworks agreed upon in advance.
- Smart contracts promote the automation of cooperation logics in terms of (business) processes and elements in a trusted way. Trustworthiness of interaction is granted, and cooperation of partners can be based in more open networks in which trust is based on algorithms and organizational positions.
- Assets can be represented on a blockchain using tokens of various kinds. As such, tokenization can be applied to a multitude of business assets starting with raw data of production processes via data-driven knowledge about the optimization of

processes toward auditing machine data amid production processes and certificates for vocational training.

In light of these opportunities, we also have to consider limitations with regard to information governance:

- A blockchain is not suitable to store large amounts of data that are typical for operational and production environments. Therefore, a blockchain application often requires the combination and linkage of blockchain transactions with data managed in an off-chain information space.
- Transaction throughput of blockchain infrastructures is also limited compared to established database technology. Thus, applications with high-frequency transaction workload require special solutions with regard to data governance.

Consequently, blockchain applications require an integration with high-performance data management services such as the IDS. A separation of concerns between efficiency of data crunching and immutability of process auditing is an effective solution.

The IDS Reference Architecture Model already identifies the application of blockchain for data consistency and data transparency [1]. However, we believe that beyond these aspects, a number of additional integration and combination possibilities exist [4].

IDS themselves aim at meeting several strategic requirements of which the following are very much related to the basic blockchain concepts. The first is the requirement of trust which is satisfied by an evaluation and certification of each participant before granting access to the system. Identification management and certificate management form one of the primary application areas for blockchain technology. Several solutions are already available [27, 28]. In combination with the recently published W3C standards on Decentralized Identifiers [29] and Verifiable Credentials, the blockchain technology can provide an ideal solution for the decentralized and trusted management as well as fully automated verification of identities and credentials. A solution for the identification and the authentication of stakeholders for inland waterway transport has been developed for project Sinlog [30]. The core idea is that a written confirmation in terms of personally signed documents—perhaps scanned and forwarded as PDF by email for the sake of digitalization—can be replaced by certified signatures of humans as well as technical agents, e.g., the load of a barge has to be assessed, be it by a human with a yardstick or a technical sensor on board.

A blockchain network can provide the functionality of the certification authority identified in the IDS Reference Architecture Model. [1], which would also be in line with the architecture proposed by the W3C [29].

The aspect of security and data sovereignty as proposed is also very much in line with blockchain principles. The IDS apply the concept of usage contracts and usage policies to determine the permissions and obligations of a resource. The implementation of these policies can be defined as a smart contract in a blockchain that is directly attached to the information itself or a token that represents the information or

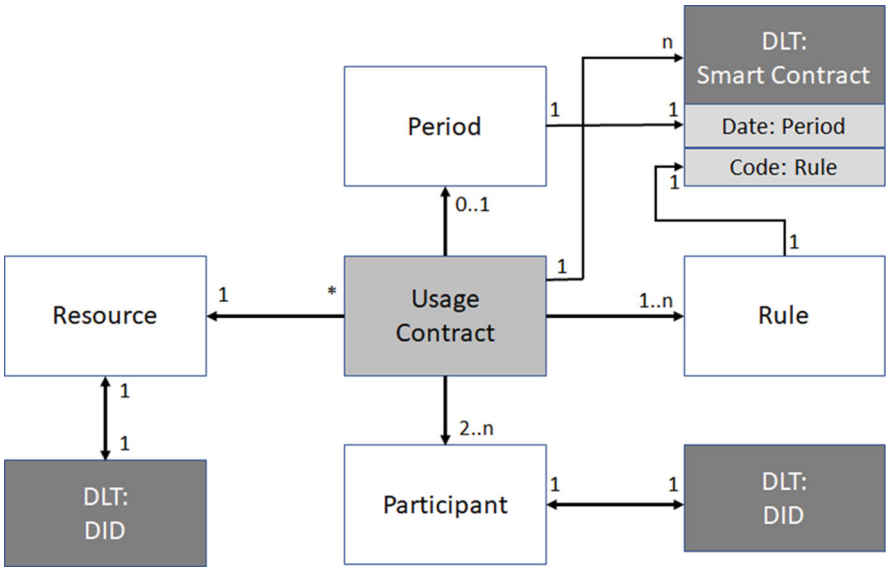


Fig. 10.1 Usage Contract concept from [1] extended with DLT concepts

any other asset. Furthermore, standardized smart contracts can provide means for trusted exchange and distribution of access keys to information as demonstrated in Weber and Prinz [31]. In combination with payment tokens or a planned digital euro, it can also support an integrated payment process.

Figure 10.1 illustrates the current data model of the usage concept. This can be mapped to a blockchain concept in the following way:

- The *Resource* is represented as an asset token.
- The *Participant* is identified by a DID.
- *Period* and *Rule* are stored as data in a smart contract.
- The *Usage Contract* is implemented as a smart contract.

With this approach, the usage contract can be implemented either as a generic smart contract that is able to interpret and execute the rules or as a smart contract that actually implements the rules. The former allows more flexibility as rules can be adopted but will require the implementation of a smart contract as a sophisticated rule engine. Rules that two partners have agreed upon will then be stored in a blockchain as the payload of a transaction or within a specific rule smart contract. This allows for more flexibility as rules can be changed during a cooperation process if both parties agree, which is similar to process modeling in smart contracts [32].

Vice versa the usage policy language proposed for the IDS provides a very good basis for the specification and automatic generation or configuration of smart contracts. In such a scenario smart contracts are being automatically generated from a usage policy and deployed in a blockchain. Thus, the smart contract can be

used to enforce a usage policy on data that is either stored in a blockchain or guarded by a blockchain through the provision of access rights or keys.

Another application of the IDS concepts is the use of the usage policies in the context of programmable money. Programmable money goes beyond digital money or a cryptocurrency not by just providing a tokenized asset but by adding rules and programmable capabilities to money. This enables the combination of a budget with a specific purpose. Examples for such rules are:

- These 50 euros can only be used to pay for food, but no alcohol.
- Only 20% of my donation shall be used to pay for administrative purposes.
- 10.000 euro of my heritage shall be transferred to my granddaughter.

These few examples show that this concept can have a strong impact on services such as auditing but also general trustees' services and they enable a whole new range of blockchain applications. However, they do also require a semantic modeling of services outside the blockchain environment to enable the validation of services in the context of the execution of programmable money rules. Obviously other alternatives for the description of these rules exist [33]. However, once the blockchain community would adopt the IDS specification, then the world of data and production would speak the same language as the world of finance.

The considerations of this section show that a combination of the blockchain and IDS technologies has more opportunities and potential than just using the blockchain as trust service for data consistency and data transparency. In the following section, we will illustrate these with application examples.

10.4 Application Examples: Industrial Use Cases

The previous section unveiled opportunities of the IDS to capitalize on blockchain technology. Usage policies are one instance, i.e., an automation of provenance governance. On the other hand, information use calls for rewards. Hence, usage policies can be seen in both directions between IDS and blockchain technology. This section highlights some projects that exemplify the advantages of blockchain for governing integrated data spaces and capitalize on data-driven services:

- *TrackChain*—Initial focus has been provenance of processes in the logistic domain. Integration of devices via Internet of Things (IoT) has allowed for several automation patterns in terms of smart contracts, while role-based access control supports flexibility of information visibility.
- *Silke*—Safety and traceability of food chains is of vital importance for all kinds of wholesalers and retailers in the food production industry. Silke illustrates the use of blockchain technology for secure process provenance. Although such tracing is mostly required upstream, i.e., from the retailer toward the producer, a reverse approach might also be beneficial when automating product handling by smart contracts.

- *Sinlog*—German inland waterway transport is the most environmentally sound means for transport, be it for container transports or bulky loads such as grain or coal. Digitalization across the transport routes is rather sparse and merely in place for local terminals or single operators. Cross-organizational exchange is sparse, and the question arises on how blockchain technology can impact the efficiency of transportation processes. Moreover, audit certificates for information are required when no more physical documents with handwritten signatures for authentication are in place.
- *BC for Production*—In production environments, i.e., sequencing of machinery for the fine-blanking of body parts for a car, fusion of information sources is essential for the optimization of production lines. Since sources can come from different operators, questions about a trustworthy information fusion arise, i.e., how to certify the correctness of information transmitted without unveiling raw data.
- *Blockchain for Education*—Many people intend to transfer the management of educational and vocational certificates to electronic deposit boxes for reasons of comfort. Hence, digitalization of certificates and a transfer to content management system are offered for such reasons. Moreover, such certificates are a constant source of fraud. Blockchain technology can be employed to assure the correctness of certificates issued by training institutions.

All projects share some conceptual features.

- *Fusion of heterogeneous data*—All applications are founded in the fusion of data originating from different data sources. Hence, sovereignty of data providers is a vital concern by nature.
- *Data governance*—Operational data such as university certificates or sensor data from machinery are maintained by database-oriented systems. Blockchain is merely used for auditing purposes with regard to data transferred or processes conducted.
- *Flexibility of process cooperation*—Due to advanced availability of digital information, some process optimizations can be implemented. Yet, blockchain technology has to assure the correctness of information transferred before official authentication arrives.
- *New governance structures*—Typical intermediaries are replaced by ledgers controlled by blockchain technology. In addition, new cooperation patterns among partners can be established.
- *Sensors for data capture*—Some of the systems built depend on sensor data. Once these sensing devices are certified with proven identities, then processes can be fully automatized without human certification.

10.4.1 TrackChain

TrackChain strives to improve the tracking of transports. Products are going to be shipped from manufacturers toward consumer sites. Goods are moved by haulers and each transport is insured. Thus, four stakeholders are involved in the process. Although they relate to the same objective of delivering a product, they certainly have different business objectives on their agenda. However, all parties share the interest to have an immutable documentation of the transport process. In addition, sensors inside transport containers are used to monitor each transport with regard to violations of temperature or concussions, which are both potential risks to the goods.

Parties certainly ought to have different views on the data of the process, which is provided by role-based access control. Yet, proper provision of data access has to be traced by a blockchain.

TrackChain illustrates the advantages of utilizing blockchain technology in several dimensions:

- A collaborative information flow about transport devices with goods from a production facility to customers including transport agencies and insurance companies is maintained by a blockchain.
- As such, the clearing of service use can directly be triggered by smart contracts.
- Transport procedures are automatically controlled by sensors. In case of violations of transport conditions, measures can directly be triggered. If goods arrive with damages at their final destination without any sensor alarms, then damages might originate from the sender. Hence, sensors are used to monitor the transport in an immutable fashion. If violations occur, measures can be initiated.

10.4.2 Silke

Silke addresses the safety of transport chains in food production processes. Food products are very sensitive and vulnerable against environmental conditions. During the transportation of meat, certain temperature constraints have to be assured and monitored for safety and liability reasons. Hence, transport monitoring is similar to TrackChain. In addition to TrackChain more processing agents are involved in the entire process. Thus, a number of agents and sensing devices increase. Yet, automation and control opportunities are pretty similar.

One difference revolves around the tracing of the transport chain. Typically, there is an upstream interest in the provenance of the process: the destination is interested in the correctness of the process starting from the source. As such, proper execution of transportation orders can be monitored. In a downstream scenario the source would strive for governing proper transport conditions in order to assure an outstanding quality of the product produced and delivered, e.g., certain handling conditions might be required. Again, smart contracts can be utilized to enforce

such constraints, e.g., temperature conditions or packaging instructions, down the supply chain.

10.4.3 Sinlog

Sinlog strives for the digitalization of documents in German inland waterway transportation. As of now, information logistic in this ecosystem of transportation services is characterized by the processing of PDF documents. Digital information, e.g., specific attributes of containers, is only available inside organizations such as terminals but not across organizations. An exchange of information is only possible when the full account of information for a processing step is available in terms of PDF documents as proof of activities. Imagine how overall process performance can be improved when pieces of information can be forwarded in advance. Simple examples include the improvement of planning processes for container placement in terminals.

By splitting information typically represented by one document into pieces of information, many process optimizations are possible due to an earlier availability of information as advocated by manifestos for re-engineering. In addition, process collaborations can be tracked, and proper process compliance can be proven. Hence, cooperation patterns can become more flexible. Planning of follow-up transportation is a grassroots example.

In addition, merely handling digital assets requires an authentication of data, processes, and tools, i.e., having 1000 tons of grain be loaded according to what meter. Thus, identity management of process participants is necessary [14], be they human or technical agents in terms of sensors such the ones for insertion depth as meter for the tonnage loaded.

When all documents revolving around the transport are available as electronic freight transport information (eFTI) according to European bylaws, then questions about correctness of information arise in general. Thus, information exchanged requires certification in terms of content as well as origin. Hence, digitalization of processes generates new requirements for identity management since the trustworthiness of exchanges via certified documents diminishes.

10.4.4 BC for Production

Data security and traceability while at the same time maintaining the sovereignty of data providers are becoming decisive factors for production processes in light of digitalization and globalization for manufacturing companies (manufacturing, factoring, supplier, etc.). Methods for an immutable and controllable documentation of production-related data are necessary. At the same time, access to information

sources has to be supported with a revision-safe documentation of any kind of utilization.

Intelligent management approaches for production data are required that offer related information in process-aware contexts. The IDS again serve as a concept for data fusion, while blockchain supports the clearing of data utilization and auditing of transactions.

In production industries several machines contribute to the manufacturing of goods. Along this production process data about production parameters are exchanged and integrated. A part passing different processing stations and perhaps also crosses enterprises can turn from a kind of raw material toward a product. Along this process of formation several information transfers from suppliers to manufacturers happen. However, suppliers do not intend to unveil their production parameters in detail but merely confirm the compliance of production constraints.

A similar pattern occurs for the transportation of bulk material in Sinlog. A delivery of grain has to guarantee certain levels of nutrition and moisture. Since the delivery will be a combination of several contributions, the trader has to guarantee quality indicators, but is certainly not willed to unveil details [34]. Hence, raw data will not be provided but the compound is supposed to meet the requirements. The trader merely encodes the aggregation of the data such that proper delivery can be proven. Such pattern of information exchange is a role model for many business collaborations. Our approach furnishes the technical means for such business collaboration and the tracking of its provenance.

10.5 Conclusion

We started with the challenge of marrying the concept of the IDS with the potential of blockchain technology. Data monetization and information economies accompanied by machine-to-machine collaboration appeared as full-fledged candidates to capitalize on the potentials of both lines of research and development, i.e.,

- Generating knowledge for process improvements on the basis of aggregated datasets and knowledge derived due to econometric analytics.
- Capitalizing on this knowledge by utilization policies and their economic clearing, i.e., monitor and bill the use of this knowledge while analysts and data providers ought to be awarded with incentives.

On the one hand, *data space concepts* are required to reconcile data stemming from a multitude of information spheres each perhaps representing a different view of related activities on a shared core process such as the fine-blanking of body parts for automotive vehicles or the optimization of food production processes.

Any knowledge economy is contingent upon insights derived from an analysis of data originating from different but comprehending data sources.

On the other hand, application of this knowledge generates a significant economic advantage as demonstrated by statistical models for improving farming efficiency

and animal welfare as shown by PigConomy [5]. Since management and use of this kind of knowledge cannot be controlled by intermediaries for economic reasons due its rather low market capitalization for each single transaction, automized procedures with an algorithmically controlled governance are necessary, i.e., DLT for the clearance of knowledge use. Knowledge users will be charged on a use case basis, while data providers and analysts will receive fair incentives.

Hence, both lines of research and development for data space management and transaction control of knowledge management are essential ingredients to foster data economies. The chapter has firstly shown a conceptual marriage of both lines as a complement. Moreover, it has shown an array of outstanding examples for capitalizing on knowledge due to the digitalization of processes ranging from the tracking of the transportation of sensible goods, via the integration of process data from different machinery parcs in the production domain and process automation for the operation of container shipments toward a trustworthy management of educational and vocational certificates. Furthermore, we identified how the blockchain community can adopt IDS concepts of usage policies to implement smart contracts in the context of programmable money.

To reiterate, the IDS provides a trusted operational basis for the integration of data sources, while DLT empowers knowledge transactions with trust across business partners.

Acknowledgments Parts of this work have been supported by the b-it foundation.[2] Project Sinlog is funded by the mFUND Programme of the German Ministry for Transport and Digital Infrastructure (Project id: 19F2099C).

References

1. Otto, B., Steinbuß, S., Teusher, A., & Lohmann, S. (2019). *Reference architecture model version 3.0*. https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-Reference-Architecture-Model-3.0.pdf
2. Trauth, D., Bergs, T., & Prinz, W. (2021). *Monetarisierung von technischen Daten – Innovationen aus Technik und Forschung*. Springer.
3. Rose, T., Gruber, J., & Hausmann, K. (2020). PigFarm – Developing decision support for the pork production industry. In *Proceedings EnviroInfo 2020, Digital twins for sustainability*, Nikosia, Zypern, September 2020.
4. Punter, M., Fournier, F., Skarbowski, I., Jürjens, J., Bernhard, H., & Steinbuss, S. (2019). *Blockchain technology in IDS – Position paper on the application of blockchain technology in the context of international data spaces*. International Data Spaces Association.
5. Rose, T., Gruber, J., Hausmann, K., & Osterland, T. (2021). PigConomy – Evidenzbasierte Analyse von empirischen Daten der Nutztierhaltung und deren Verwertung. In D. Trauth et al. (Eds.), *Monetarisierung von Technischen Daten*. Springer.
6. Özsu, M. T., & Valduriez, P. (1996). Distributed and parallel database systems. *ACM Computing Surveys, 28*, 125–128. https://doi.org/10.1145/234313.234368

[2] https://www.b-it-center.de/

7. Prinz W., & Schulte A., (eds.) (2018). *Blockchain and smart contracts: technologies, research issues and applications* (Blockchain und Smart Contracts – Technologien, Forschungsfragen und Anwendungen (fraunhofer.de), https://www.iuk.fraunhofer.de/content/dam/iuk/en/docs/Fraunhofer-Paper_Blockchain-and-Smart-Contracts_EN.pdf.
8. Schlatt, V., Schweizer, A., Urbach, N., & Fridgen, G. (2016). *Blockchain: Grundlagen, Anwendungen und Potenziale*. Projektgruppe Wirtschaftsinformatik des Fraunhofer-Instituts für Angewandte Informationstechnik, 2016.
9. Glaser, F., & Bezzenberger, L. (2015). Beyond cryptocurrencies - a taxonomy of decentralized consensus systems. In *Proceedings of the 23rd European Conference on Information Systems (ECIS 2015)*, May 26–29, Münster, Germany
10. Franco, P. (2015). *Understanding bitcoin: Cryptography, engineering, and economics*. Wiley.
11. Zohar, A. (2015). Bitcoin: Under the Hood. *Communications of the ACM, 58*(9), 104–113.
12. Dwork, C., & Naor, M. (1993). Pricing via processing, or, combatting junk mail, advances in cryptology. In *CRYPTO'92: Lecture Notes in Computer Science* (pp. 139–147). Retrieved from http://www.wisdom.weizmann.ac.il/~naor/PAPERS/pvp.ps
13. Prinz, W., Rose, T., Osterland, T., & Putschli, C. (2018). Blockchain. In R. Neugebauer (Ed.), *Digitalisierung: Schlüsseltechnologien für Wirtschaft und Gesellschaft* (pp. 311–319). Springer.
14. Osterland, T., & Rose, T. (2020). From a use case categorization scheme towards a maturity model for engineering distributed ledgers. In H. Treiblmaier et al. (Eds.), *Blockchain and distributes ledger technology use cases – applications and lessons learned*. Springer.
15. Peters, G. W., & Panayi, E. (2015). Understanding modern banking ledgers through Blockchain technologies: future of transaction processing and smart contracts on the internet of money. In P. Tasca, T. Aste, L. Pelizzon, & N. Perony (Eds.), *Banking beyond banks and money: a guide to banking services in the twenty-first century*. Springer International Publishing.
16. Narayanan, A., Bonneau, J., Felten, E. W., Miller, A., & Goldfeder, S. *Bitcoin and Cryptocurrency Technologies: 2015*. Abgerufen am 02.06.2016, von https://d28rh4a8wq0iu5.cloudfront.net/bitcointech/readings/princeton_bitcoin_book.pdf.
17. Bentov, I., Lee, C., Mizrahi, A., & Rosenfeld, M. (2015). Proof of activity: Extending Bitcoin's proof of work via proof of stake. *ACM SIGMETRICS Performance Evaluation Review, 42*(3), 34–37.
18. Tschorsch, F., & Scheuermann (2015). *Björn: Bitcoin and Beyond (2015): A technical survey on decentralized digital currencies*. https://eprint.iacr.org/2015/464.pdf.
19. Christidis, K., & Devetsikiotis, M. (2016). Blockchains and smart contracts for the internet of things. *IEEE Access, 4*, 2292–2303.
20. Wright, A., & de Filippi, P. (2015). *Decentralized Blockchain technology and the rise of Lex Cryptographia*. http://ssrn.com/abstract=2580664.
21. Szabo, N. (1997). Formalizing and securing relationships on public networks. *First Monday, 2*(9).
22. Kõlvart, M., Poola, M., & Rull, A. (2016). Smart contracts. In T. Kerikmäe & A. Rull (Eds.), *The future of law and eTechnologies* (pp. 133–147). Springer International Publishing.
23. Osterland, T., & Rose, T. II (2020). *Model checking smart contracts for Ethereum, mobile and pervasive computing, special issue on blockchain technology, 2020*.
24. Fridgen, G., Guggenberger, N., Hoeren, T., Prinz, W., Urbach, N., & et al. (2019). *Chancen und Herausforderungen von DLT (Blockchain)*. In Mobilität und Logistik, Bundesministerium für Verkehr und digitale Infrastruktur.
25. Klein, S., Prinz, W., & Gräther, W. (2018). A use case identification framework and use case canvas for identifying and exploring relevant blockchain opportunities. In *ERCIM Blockchain workshop 2018*. European Society for Socially Embedded Technologies (EUSSET).
26. Osterland, T., & Rose, T. (2018, May). *Engineering sustainable blockchain applications*. ERCIM Workshop on Blockchain Technology.

27. Gräther, W., Kolvenbach, S., Ruland, R., et al. (2018). Blockchain for education: Lifelong learning passport. In *ERCIM Blockchain Workshop 2018*. European Society for Socially Embedded Technologies (EUSSET).
28. Prinz, W., Kolvenbach, S., & Ruland, R. (2020). Blockchain in der Ausbildung: Ein lebenslanger, sicherer und nachvollziehbarer Lernausweis. In *Expedition: Werte, Arbeit, Führung 4.0* (pp. 107–111). TÜV Media GmbH.
29. W3C-DID, Decentralized Identifiers (DIDs) v1.0. https://www.w3.org/TR/did-core/ (Working draft 12.02.2021).
30. Osterland, T., & Rose, T. III (2020). Decentralised identity management for DLT-based cooperation support – Enabling new business relationships for inland waterway transport. In *Proceedings EnviroInfo 2020, Digital twins for sustainability*, Nikosia, Zypern, September 2020.
31. Weber, T., & Prinz, W. (2019). Trading user data: A Blockchain based approach. In *2019 Sixth International Conference on Internet of Things: Systems, Management and Security (IOTSMS)* (pp. 547–554).
32. Osterland, T., Rose, T., & Putschli, C. (2020). On the implementation of business process logic in DLT nodes. In *Proceedings International Artificial Intelligence and Blockchain Conference (AIBC 2020)*, Nagoya, Japan, May 2020.
33. Royal, D., Rimba, P., Staples, M., Gilder, S., Tran, A. B., Williams, E., Ponomarev, A., Weber, I., Connor, C., & Lim, N. (2018). *Making money smart - empowering NDIS participants with Blockchain technologies*. Commonwealth Bank of Australia, CSIRO.
34. Osterland, T., & Rose, T. (2021). Oracle-based process automation in DLT dominated ecosystems with an application to German waterway transportation. In *Proceedings 2nd Artificial Intelligence and Blockchain Conference (AIBC)*, Macau, China.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 11
Federated Data Integration in Data Spaces



Matthias Jarke and Christoph Quix

Abstract Data Spaces form a network for sovereign data sharing. In this chapter, we explore the implications that the IDS reference architecture will have on typical scenarios of federated data integration and question answering processes. After a classification of data integration scenarios and their special requirements, we first present a workflow-based solution for integrated data materialization that has been used in several IDS use cases. We then discuss some limitations of such approaches and propose an additional approach based on logic formalisms and machine learning methods that promise to reduce data traffic, security, and privacy risks while helping users to select more meaningful data sources.

11.1 Introduction

Data of all kinds and media formats constitute evidence about the state and events in the world. Analogous to legal procedures, these evidences must be checked for quality and synthesized in a semantically meaningful and purpose-oriented manner in order to assist human-machine learning and evidence-driven decision making. This is the task of data integration.

Transactional databases (ERP systems), huge social networks, and most recently billions of sensors in the Internet of Things produce an ever-growing volume and variety of such evidences in ever-growing velocity but sometimes doubtful veracity. It is therefore not surprising that, despite almost 40 years of research, data integration

M. Jarke (✉)
Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

Informatik 5 (Information Systems), RWTH Aachen University, Aachen, Germany
e-mail: jarke@dbis.rwth-aachen.de

C. Quix
Fraunhofer Institute for Applied Information Technology FIT, Sankt Augustin, Germany

Hochschule Niederrhein, University of Applied Sciences, Krefeld, Germany
e-mail: christoph.quix@fit.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_11

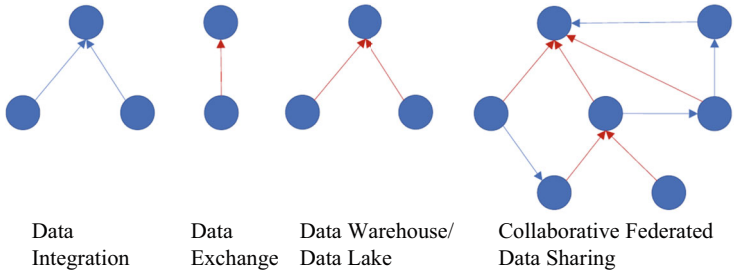


Fig. 11.1 Data integration scenarios (adapted from [1])

remains a key challenge in data science which still requires about 80% of all data analytics work.

Maurizio Lenzerini, arguably the internationally best-known data integration researcher, has recently proposed a useful classification of data integration scenarios that have shaped requirements in the field (cf. Fig. 11.1).

In the figure, the blue nodes represent datasets. They can play the role of data sources or integration results in the data integration tasks indicated by the arrows; arrows are formally expressed as query specifications, also called *view definitions*. The blue arrows indicate *virtual data integration* by query specifications that selectively produce integrated data only when actually activated by a user or application program. In contrast, red arrows indicate *materialized data integration*, i.e., the complete set of specific query results from the view definition is pre-computed and stored as a physical dataset and needs to be maintained when source data change. Briefly, materialized data integration moves the data from the source to the query site, whereas virtual data integration moves the query algorithm partially to the sources. Thus, materialization increases query performance by pre-computation of views, whereas the specific user query in virtual data integration can reduce data transfer, thus reducing communication network load and increasing data sovereignty.

Applying these observations to the four scenarios in Fig. 11.1, early *Data Integration* research laid the formal foundations for query answering among the growing number of distributed transactional, usually relational, databases.

In the 1990s, the novel concept of decision support by data analytics required a separation of data warehouses (DW) [2] from transactional databases: firstly, the need to keep historical data over long period of times for, e.g., time series analyses; secondly, problems in transaction management concurrency control caused by interference of short update transactions and analytics queries spanning large datasets; and thirdly, the need for more user-friendly multidimensional online-analytical processing (OLAP) query facilities in so-called data marts. In practice, the view materialization in DW is usually supported by a so-called *ETL workflow* which coordinates various commercial or open-source tools for *Extraction* from data sources, *Transformations* for cleaning the extracted data and integrating data from

different sources into a uniform DW data format, and finally *Loading* the transformed into the DW storage.

While the DW approach resulted in high-quality integrated data, the ETL effort turned out to be a very high upfront investment, often well before or even without a return on investment which came only with successful analytic queries.

Reacting to the growing need for real-time decision support based on high-frequency stream data inputs – first in finance, later on in web shops, social media, and recently technical and science applications – this upfront effort appeared to be no longer doable. In addition, novel algorithms such as MapReduce enabled highly parallel initial loading and structuring of incoming data and analytics queries on the sources, and the concurrency control issue was partially resolved by novel main memory techniques such as the column store of SAP/HANA. As a consequence, data lakes introduced since the mid-2010s changed the workflow structure from ETL to an ELT format: after Extraction, the data is loaded to the data lake in its original form; Transformations are applied only to the data in the lake if required for a specific application, to avoid the complexity of ETL process design for big data. Service-oriented architectures enable not only data integration but also the integration of complex workflows in which several application systems and services such as Hadoop, Apache Spark, or KAFKA are involved.

On the data lake usage side, data scientists explore various datasets to find rules or functions that confirm and formalize correlations. In these settings, each scenario requires a different set of data sources to be integrated, with different requirements for granularity and data quality. Thus, materialized data integration as applied in data warehouses is less applicable for big data use cases; instead, rich structural and semantic metadata are typically created on top of the raw data, to permit multiple topical perspectives on the data lake. Further details on data lake functions and metadata are elaborated in a recent survey paper [3].

However, data management does not only include data sources within an organization; increasingly, external data sources have to be considered. These external sources come in different forms, e.g., datasets provided by some marketplace. Therefore, data exchange across organizations and usage of data provided by external partners, as studied by the International Data Space Association, is becoming highly relevant [4]. As we shall discuss more deeply in the remainder of this chapter, such workflow solutions also dominate practice in the tasks that are typical for the *Data Space* scenarios of individual *Data Exchange* and trusted *Collaborative Data Sharing* in federations of partners within a Data Space. They do not just combine virtual and materialized data integration and querying, but also include additional Data Space management tasks such as data source discovery, negotiated access, and usage control components in data integration workflow.

In the following section, we use a mobility data space scenario to illustrate a solution which supports a view materialization workflow in the International Data Space setting. Such a setting has been implemented in similar form in several IDS use cases described in later chapters of this book.

Less attention has been paid so far to transferring the virtual, query-driven data integration strategy to the data exchange and federated data sharing of Data Spaces.

In Sect. 11.3, we review such technologies in a more formal, logic-based framework and present a perspective how logic-based query optimization could further improve federated data integration and exchange in Data Spaces.

11.2 Federated Data Integration Workflows in Data Spaces

In this section, we first sketch the challenges of federated data integration in a case study setting of mobility engineering, and then use this case study to present extensions to data integration workflows for a data space support beyond similar efforts in data warehouses and data lakes.

11.2.1 A Simple Demonstrator Scenario

Although data exchange scenarios have also been considered in data integration research [5], constraints in using external data have not been considered. For example, there might be usage constraints on the data, i.e., data may not be stored in local repositories and can be processed only once. Another requirement might be to protect the privacy of data providers, i.e., data can only be accessed to perform aggregate computations which do not reveal detailed information. To illustrate these requirements, Fig. 11.2 shows a mobility scenario extracted from a large collaborative project with the automotive and communication industries on Car2X communication [7].

In the scenario, cars equipped with Car2X communication devices [8] expose detailed vehicle data to a data aggregation service (e.g., a specific platform of the car manufacturer such as BMW ConnectedDrive). This data is made available in a mobility data marketplace (e.g., Mobilitätsdatenmarktplatz (MDM) in Germany, https://www.mdm-portal.de/). Furthermore, additional external data sources could be made available in this marketplace, such as weather data, information about construction sites, or information about local events. This information could be used by service providers that offer information about the current traffic state, compute fastest routes, or give real-time warnings about road hazards.

However, these service providers should not have full access to detailed information of data providers, e.g., their locations, speed, etc. For example, detailed location data might reveal sensitive information as it has been shown with taxi data from New York City, even though the data has been anonymized.[1] To address this challenge, a data exchange platform should enable data providers to specify usage policies for their data. For example, a usage policy could state that data may not be

[1] https://www.fastcompany.com/3036573/nyc-taxi-data-blunder-reveals-which-celebs-dont-tip-and-who-frequents-strip-clubs

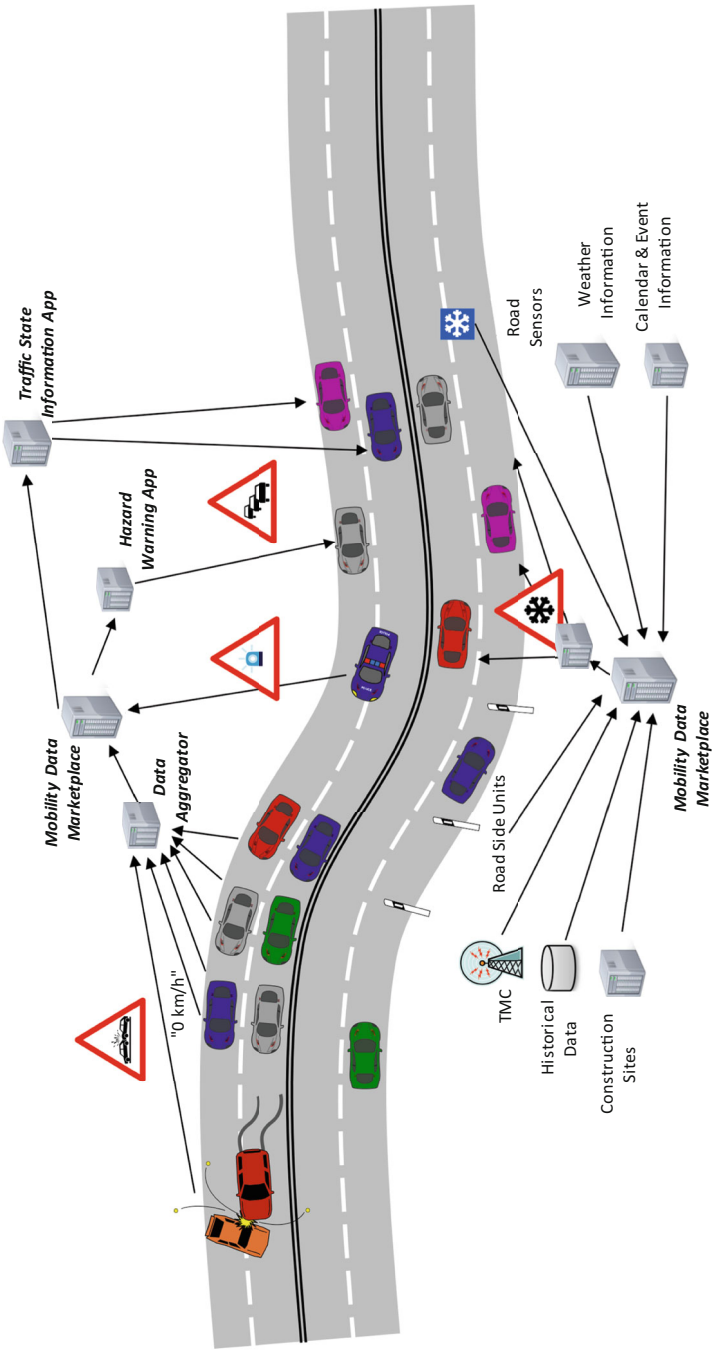


Fig. 11.2 Data exchange in a mobility scenario (adapted from [6])

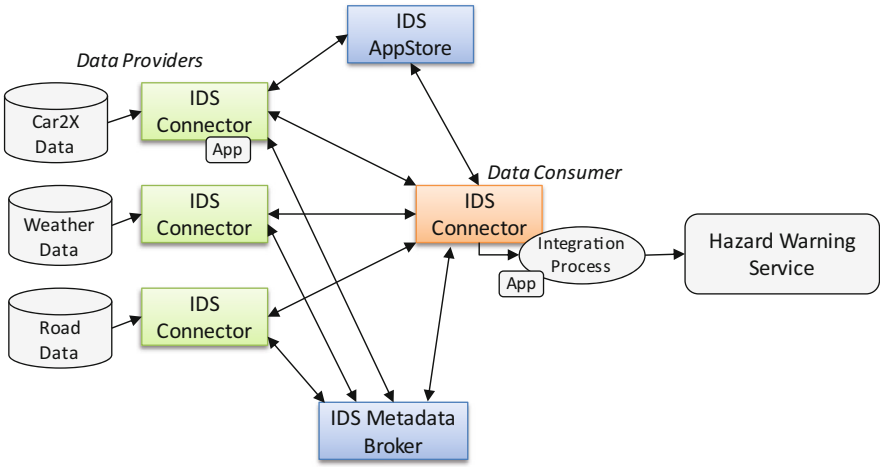


Fig. 11.3 IDS integration architecture for traffic scenario

stored and can be processed only once as in data stream management systems [9], or data can only be processed if it is anonymized and aggregated with other data items.

11.2.2 A Data Integration Workflow Solution for Data Spaces

Current IDS implementations follow a pragmatic, workflow-oriented approach that has been applied also successfully in data warehousing and many other integration approaches. In the IDS, data integration is done within a connector by using integration workflows that are implemented as routes in the Apache Camel framework. Figure 11.3 illustrates an integration architecture for the traffic scenario from Fig. 11.2.

In the architecture, three connectors are used to provide data about vehicles, their location, speed, etc. weather data, and data about roads. These connectors could be run by three different data providers that aim at selling their data in the IDS. To advertise the data and to make it available in the IDS, they need to publish descriptions of their datasets, including technical information of their connectors (e.g., URLs and protocols to query the data), at the IDS Metadata Broker. The IDS Metadata Broker is a central component in the IDS architecture [10] that manages metadata according to the IDS Information Model [11]. The IDS Information Model uses the Web Ontology Language (OWL) to define the basic vocabulary in which data resources in the IDS can be described.

Another central component of the IDS architecture is the App Store. It can provide "Data Apps," i.e., applications that can be deployed inside a connector to perform certain data operations. For example, a data app could transform data items between different standard formats (e.g., XML to JSON), join datasets from different

sources, and perform a complex aggregation or machine learning task. Data apps can be part of the integration processes that are defined within a connector.

Within such a process, a data app might have an input, an output, or both. In case the app has only input or output, it is considered to be a system adapter, i.e., it has an interface to a backend system in which data is written or from which data is read. However, this backend system is not part of the IDS infrastructure; in the viewpoint of the IDS connector, the data app might have only input or output.

As stated, data apps might perform any kind of data operation. In the case of the connector for Car2X data, the data app is responsible for aggregating, anonymizing, and noising the data in order to guarantee the privacy of the original data providers, e.g., car drivers.

To integrate data from different connectors, as required in this scenario, the connector of the data consumer has to employ an integration process. An integration process could be implemented in many different ways, e.g., manually implemented in any programming language or using one of the many (open-source) data integration frameworks. In the IDS, some connector implementations have decided to use Apache Camel. Apache Camel is an integration framework that allows to define data integration and transformation as *routes*. A route is a sequence of operations that are applied to datasets. Several routes can be combined by using a join-like operation. A benefit of Apache Camel is its broad support for different source systems and operations. It can also be extended with customized operations. Apache Camel uses a domain-specific language for the specification of routes that can be defined in various languages (e.g., XML, Java).

In the scenario of Fig. 11.3, we can use a Camel route to integrate the data from the three data providers. A data app could implement more intelligent behavior than just joining and aggregating data, e.g., analyze the data stream with current traffic information and apply a machine learning model to predict traffic states in the near future or to detect road hazard (e.g., icy road, queue end) [7].

The additional functionality to guarantee data sovereignty can be easily integrated into the Camel routes by using the extensibility features of Camel. When data is exchanged between two IDS connectors, the route needs to be enriched with an interceptor pattern. The interceptor will check the usage policies which are defined for the current dataset, and enforce the corresponding usage restrictions. Interceptors can be defined in different ways and depend on the language that is used to define usage policies. MY DATA[2] and LUCON [12] are two examples for usage control systems which implement the interceptor pattern.

[2] https://www.dataspaces.fraunhofer.de/de/software/usage-control/mydata.html

11.3 Toward Formalisms for Virtual Data Space Integration

The materialization-oriented workflow satisfies the requirements but does have a few limitations. If the user asks a specific query, the integrated answer will usually be much smaller than the sum of the data needed to produce it. This will result in many more data to be transported to the integration site than necessary, creating network overload and unnecessary access risks and related IDS controls for data that perhaps would not even need to be exported by the data provider if the query could be adequately optimized.

However, there is an additional issue that we did not mention before. In the Introduction, we interpreted data as evidence about reality; this would argue for an understanding that data sources are not a "ground truth" but themselves limited-quality views on a reality that could be structured according to different integrative worldview (or ontology) by the provider and consumer organizations. In database jargon, this perspective is called "local-as-view" (LAV). In contrast, most materialization-based workflows like the one shown in Sect. 11.2 do consider the data sources as ground truth and the global integrated data as an integrative view on these sources ("global as view – GAV"). The same holds where data are only accessible through parameterized reusable software services [13] or apps. Thus, it is impossible to make statements, e.g., about the completeness of data sources with respect to a broader ontology and about the credibility of conflicting evidences among different sources.

Modern logic-based theories of data integration therefore adopt a combined GLAV ("global-local-as-view") perspective, cf. the combination of red and blue arrows in the federated collaborative data sharing scenario of Fig. 11.1.

In the remainder of this section, we demonstrate the necessary logic formalisms, reusing the example of Fig. 11.2, and discuss how the recent approaches for data integration in data lakes could be extended to the Data Space setting.

11.3.1 Logical Foundations for Data Integration

To illustrate the concepts of data integration in this section, Fig. 11.4 extracts a simplified heterogeneous schema integration setting from the scenario in Fig. 11.2. An XML web service *S* provides information about vehicles and their location. A relational data source *R* delivers information about roads, their locations, and other geographical data. A service provider needs integration of these two sources into a web service T, providing current traffic state information in a JSON format which is frequently used in web or mobile applications.

Research in data integration has developed several key technologies and fundamental concepts to address such data integration problems.

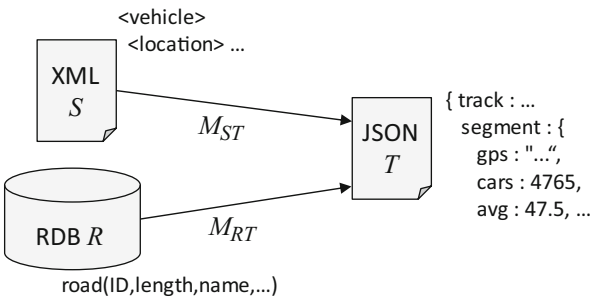


Fig. 11.4 Simplified heterogeneous integration scenario for traffic data

Model and Mapping Management Early approaches to data integration mainly focused on efficient data transformation and scheduling the integration and cleaning jobs in such a way that the operational systems were not affected [14, 15]. Management of schemas or formal specifications of the transformation procedures was a side activity in optimizing the ETL workflows. Only the vision of model management was to consider schemas, mappings, and other metadata of information systems as "first-class citizens" of data integration systems [16]. The goal was to develop an algebra with basic operations to match, integrate, and compose models and mappings. Although the vision of an algebra has not been achieved completely, the idea of model management initiated research on these fundamental operations for data integration [17], such as schema matching and mappings, as discussed in the following paragraphs.

Heterogeneity An immanent problem of data integration is heterogeneity. An obvious challenge in data integration is the wide variety of data formats that are used in data sources. Once the data formats have been identified, appropriate transformation functions can be applied to deal with this problem. A more challenging problem is heterogeneity at the model level and at the semantic level. In the example of Fig. 11.4, three different modeling languages are used. A modeling language is a formalism that is used to describe the data structures, e.g., XML Schema, "CREATE TABLE" statements in SQL, or JSON Schema. In order to map and link the models in these different languages, one has to translate these models into a common formalism that uses a generic representation [18]. More sophisticated is the problem of semantic heterogeneity. In the example, different terminology is used (e.g., "track" vs. "road", "vehicle" vs. "car"). In order to link data sources, a data integration engineer has to understand the different entities that are encoded in the data sources. While advanced algorithms for schema matching have been developed in the last two decades (see paragraph on matching below), a fully automatic approach that can resolve all heterogeneities is most likely not possible. Human intervention will be always required to verify and to complete schema alignments.

Schema Matching Schema matching is the task of identifying correspondences in two schemas. This field has been very active in the last two decades and can be subdivided into the following categories [19, 20]:

- *Algorithms:* Various algorithms detect similarities in schemas, based on labels, structure, or auxiliary information. Usually, not a single approach is sufficient; thus, a combination of several algorithms is important, for which different strategies for combining individual results can be developed [21].
- *Schema representation:* for schema matching, it is often sufficient to represent schemas as directed graphs. Although this representation loses some details about schema constraints, algorithms for graph alignment and matching can be applied in this case. Other approaches transform the schemas into a generic representation with detailed representations for constraints and datatypes, which can be leveraged to find correspondences of schema elements [22].
- *Semantic matching:* especially in the field of ontology alignment [20], additional sources with semantic information are exploited. This includes, for example, dictionaries and thesauri to identify terms with similar meaning, or repositories with existing alignments. In addition, reasoning about the relationships and rules expressed in the ontologies can be used to detect related elements. More information on the IDS metalevel ontologies and linked metadata structures is presented in a later chapter of this book.
- *Matching systems:* As stated above, algorithms have to be combined in order to achieve a satisfying matching result. Schema matching systems have focused on the integration of different algorithms, such that the output of one algorithm can be used as input for another algorithm to compute an extended alignment [23, 24]. Another important point in the development of matching systems is performance. A complexity of $O(n^2)$ seems to be unavoidable as all elements of two given schemas should be compared, but this would not scale to large schemas. Thus, schema matching systems have used different optimization techniques to reduce complexity and to compute alignments efficiently, e.g., by exploiting hashing or parallel computations.

11.3.2 Data Integration Tool Extensions for Data Spaces

Some of the technologies have been integrated into commercial data integration products, but due to the complexity of matching algorithms, these technologies are limited to simple label similarity (or even equality) and structural comparisons.

However, there exist algorithms and large-scale prototypes in scientific and industrial research that are promising for extension to the Data Space setting:

Schema Integration Given the information about similarities of two schemas, a new schema should be derived that integrates the information of both schemas [25]. In the example of Fig. 11.4, the schema of the JSON document *T* that integrates *R* and *S* needs to be defined. In this scenario, application requirements for the traffic state information will mainly determine the outline of the *T*; however, in more information-centric applications, e.g., data warehouse systems, the integrated

schema will be defined as the "union" of the source schemas, taking into account computed alignments. While intuitively the problem of schema integration is clear, the difficulty is to formalize the schema integration process: which requirements should be satisfied by the integrated schema, how can we verify that the integrated schema is correct, and what is the meaning of "correctness" and "completeness" in schema integration? One approach could be to follow a query-oriented approach: queries that can be answered on the sources should deliver the same result on the integrated schema [26].

Mappings The alignments that have been computed by schema matching algorithms only state that there is a similarity between two schema elements. However, to create integrated datasets, one needs to transform the data instances from the sources into the desired target structure. In the example from Fig. 11.4, we need to have an XQuery and an SQL query that extract the required information from the XML document and the relational database. Furthermore, we need to have a function that creates the required JSON document. All these queries and transformations need to be transformed within the same data transformation tool, such that the data can be integrated and aggregated into one dataset. In the example, we need to aggregate the number of cars per road and compute their average speed, which requires to combine the data from both sources. Mapping frameworks for heterogeneous data integration scenarios as in this example have been proposed in research [27, 28] several open-source projects have adopted these approaches and provide similar functionality, e.g., Apache Spark and Apache Drill.

Related Dataset Discovery In today's huge distributed data lake or Data Space environment with hundreds or thousands of possible data sources, consumers often face the problem that they do not even know which data source (and data provider) contains useful relevant information for their intended query. In the example of Fig. 11.3, this has simply been delegated to a separate IDS Agent called the Information Broker, but in reality, where no fully integrated schemas exist, also this "Related Dataset Discovery" problem needs to be addressed in a query-specific manner to maintain relevance of the answer and avoid unnecessary network traffic. A promising recent approach pursued, e.g., in the Aurum [29], JOSIE [30], and D3L [31] projects, is that the user specifies the format of their desired answer table together with some relationship aspects that are important to them (value overlaps, numerical data distributions, textual attribute semantics, etc.), and fast similarity algorithms propose possible extensions of the query table in the sense of "what else is relevant beyond what you have explicitly asked for"; the algorithm then adds the corresponding data sources for those extensions considered relevant by the user to the query rewriting process.

Query Rewriting The benefit of the aforementioned open-source projects is especially the ability of query rewriting (or query translation): queries are defined in a generic query language (e.g., Spark SQL in the case of Apache Spark) and translated into the query language of the data sources. Without such mappings, even the simplest, but often extremely efficient, data optimization heuristic such as "execute

selection before join" cannot be propagated backwards from the query to the data sources. However, the query translation is usually done only with simple one-to-one mappings between an integrated schema and the source schemas, which usually implies that the source data has first to be transformed into the required structure.

More complex query rewritings, which involve join and union operations or distributed data and heterogeneous polystore source databases, are still subject to research [32–34]. An additional demand in query rewriting for sovereign data exchange in Data Spaces is the inclusion of privacy, access, and usage control rules into the query evaluation. Fortunately, we already know since Mike Stonebraker's 1975 Ph.D. thesis on how to do this.

In summary, even though operational solutions for federated data integration as the basis for sovereign data exchange and collaborative data sharing in Data Spaces exist, there remains ample room for further research and promising industrial uptake in the field.

Acknowledgments This research was supported by the German BMBF (IndaSpacePlus, Fkz 01IS17031), by DFG (national excellence cluster EXC-2023 "Internet of Production"—390621612), and by the Fraunhofer cluster of excellence "Cognitive Internet Technologies".

References

1. Lenzerini, M. (2019). Direct and reverse rewriting in data interoperability. In *Proceedings CAiSE* (pp. 3–13). Springer.
2. Jarke, M., Lenzerini, M., Vassiliou, Y., & Vassiliadis, P. (2003). *Foundations of data warehouses* (2nd ed.). Springer.
3. Hai, R., Quix, C., & Jarke, M. (2021). *Data lake concept and systems: A survey*. CoRR abs/2106.09592.
4. Otto, B., & Jarke, M. (2019). Designing a multi-sided data platform: Findings from the international data spaces case. *Electronic Markets, 29*(4), 561–580.
5. Fagin, R., Kolaitis, P., Miller, R. J., & Popa, L. (2005). Data exchange: Semantics and query answering. *Theoretical Computer Science, 336*, 89–124.
6. Jarke, M., Jeusfeld, M., & Quix, C. (2014). Data-centric intelligent information integration - from concepts to automation. *Journal of Intelligent Information Systems, 43*(3), 437–462.
7. Geisler, S., Quix, C., Schiffer, S., & Jarke, M. (2012). An evaluation framework for traffic information systems based on data streams. *Transportation Research Part C, 23*, 29–55.
8. Fuchs, H., Hofmann, F., Löhr, H., & Schaaf, G. (2015). Car-2-X. In H. Winner, S. Hakuli, F. Lotz, & C. Singer (Eds.), *Handbuch Fahrerassistenzsysteme* (p. 28). Springer Vieweg. https://doi.org/10.1007/978-3-658-05734-3
9. Geisler, S. (2013). Data stream management systems. In P. G. Kolaitis, M. Lenzerini, & N. Schweikardt (Eds.), *Data exchange, integration, and streams* (Vol. 5, pp. 275–304). Dagstuhl follow-Ups. https://doi.org/10.4230/DFU.Vol5.10452.275
10. Otto, B., & et al. (2019). *Reference architecture model*. IDSA. https://www.internationaldataspaces.org/ressource-hub/publications-ids/, version 3.0.
11. Bader, S. R., Pullmann, J., Mader, C., Tramp, S., Quix, C., Müller, A. W., Akyürek, H., Böckmann, M., Imbusch, B. T., Lipp, J., Geisler, S., & Lange, C. (2020). The international data spaces information model - an ontology for sovereign exchange of digital content. In

Proceedings of International Semantic Web Conference (ISWC), Athens, Greece, 176–192 (part II), LNCS 12507. Springer.
12. Schütte, J., & Brost, G. S. (2018). LUCON: Data flow control for message-based IoT systems. In *17th IEEE International Conference on trust, security and privacy in computing and communications/12th IEEE International Conference on big data science and engineering* (pp. 289–299). https://doi.org/10.1109/TrustCom/BigDataSE.2018.00052.
13. Chen, P. S., Hennicker, R., & Jarke, M. (1993). On the retrieval of reusable software components. In *Proceedings of 2nd International Workshop Software Reusability, Lucca Italy* (pp. 99–108). IEEE.
14. Simitsis, A., & Vassiliadis, P. (2018). Extraction, transformation, and loading. In *Encyclopedia of database systems* (2nd ed.). Springer.
15. Inmon, W. H. (1996). *Building the data warehouse* (2nd ed.). John Wiley & Sons.
16. Bernstein, P. A., Halevy, A. Y., & Pottinger, R. (2000). A vision of management of complex models. *SIGMOD Record, 29*(4), 55–63.
17. Bernstein, P. A., & Melnik, S. (2007). Model management 2.0: Manipulating richer mappings. In L. Zhou, T. W. Ling, & B. C. Ooi (Eds.), *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 1–12). ACM Press.
18. Kensche, D., Quix, C., Chatti, M. A., & Jarke, M. (2007). GeRoMe: A generic role based metamodel for model management. *Journal on Data Semantics, 8*, 82–117., Springer, LNCS 4380.
19. Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal, 10*(4), 334–350.
20. Shvaiko, P., & Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Trans. Knowledge and Data Engineering, 25*(1), 158–176.
21. Marie, A., & Gal, A. (2008). Boosting schema matchers. In R. Meersman & Z. Tari (Eds.), *On the move to meaningful internet systems, (OTM 2008). LNCS 5331*. Springer. https://doi.org/10.1007/978-3-540-88871-0_20
22. Quix, C., Kensche, D., & Li, X. (2007). Matching of Ontologies with XML Schema using a Generic Metamodel. In *Proceedings of the 6th International Conference Ontologies, DataBases, and Applications of Semantics (ODBASE), Vilamoura, Portugal* (pp. 1081–1098). Springer.
23. Aumüller, D., Do, H. H., Massmann, S., & Rahm, E. (2005). Schema and ontology matching with COMA++. In *Proceedings SIGMOD Conference* (pp. 906–908). ACM Press.
24. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I. F., & Couto, F. M. (2013). The AgreementMakerLight ontology matching system. In R. Meersman et al. (Eds.), *On the move to meaningful internet systems (OTM 2013). LNCS* (Vol. 8185). Springer. https://doi.org/10.1007/978-3-642-41030-7_38
25. Batini, C., Lenzerini, M., & Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys, 18*(4), 323–364.
26. Li, X., Quix, C. (2011). Merging relational views: A minimization approach. In *Proceedings 30th International Conference Conceptual Modeling (ER 2011),* Brussels, Belgium.
27. Haas, L. M., Hernández, M. A., Ho, H., Popa, L., & Roth, M. (2005). Clio grows up: From research prototype to industrial tool. In F. Özcan (Ed.), *Proceedings ACM SIGMOD Conference* (pp. 805–810). ACM.
28. Kensche, D., Quix, C., Li, X., Li, Y., & Jarke, M. (2009). Generic schema mappings for composition and query answering. *Data & Knowledge Engineering, 68*(7), 599–621.
29. Fernandez, R. C., Abedjan, Z., Koko, F., Yuan, G., Madden, S., & Stonebraker, M. (2018). Aurum: A data discovery system. In *ICDE 2018* (pp. 1001–1012). IEEE.
30. Zhu, E., Deng, D., Nargesian, F., & Miller, R. J. (2019). JOSIE: Overlap set similarity search for finding joinable tables in Data Lakes. In *Proceedings ACM-SIGMOD* (pp. 847–864). ACM.

31. Bogatu, A., Fernandes, A. A. A., Paton, N. W., & Konstantinou, N. (2020). Dataset discovery in data lakes. In *ICDE 2020* (pp. 709–720). IEEE.
32. Halevy, A. Y., Rajaraman, A., & Ordille, J. J. (2006). Data integration: The teenage years. *Proceedings VLDB, 5*(2), 9–16.
33. Hai, R., Quix, C., & Zhou, C. (2018). Query rewriting for heterogeneous data lakes. In *ADBIS 2018* (pp. 35–49). Springer.
34. Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings ACM-PODS* (pp. 233–246). ACM.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 12
Semantic Integration and Interoperability



Sören Auer

Abstract A key aspect of establishing data spaces is to develop a common understanding of the data to be shared in the data space. Semantic standards and technologies were developed for this purpose since over two decades. In this article, we will discuss the history and importance of semantic integration for data spaces. We will introduce the base concepts of semantic integration—including the global identifiers for data and the W3C standards RDF, RDF-Schema, and OWL. As a result these standards and technologies can be used to devise versatile Knowledge Graphs capturing domain conceptualizations and concrete data representations. We explain how data interoperability can be achieved by linking and mapping between different data and knowledge representations. Finally we will showcase their use with an example for data integration in supply chains with the ScorVoc vocabulary.

12.1 Introduction

If we look back at the history of computing, we can see that IT technologies were initially very much bound to the capabilities of the hardware, but became more intuitive and "human" over the past decades. In the very beginning of computing, programmers had to physically interact with the technology via pushing and pulling registers or punch cards. In the 1970s and 1980s, assembler programming became more prevalent, where you still interacted relatively close with the physical hardware, but at least it was not a physical exercise anymore. Computer scientists then discovered that there are more intuitive ways to interact with computers and got inspired by cooking recipes: functional/procedural programming of the 1980s and 1990s (e.g., using programming languages such as PASCAL or C) basically resembled what cookbooks have done for centuries: describing the ingredients and a sequence of steps to realize a certain outcome. Even more intuitive was

S. Auer (✉)
TIB Leibniz Information Centre for Science and Technology and Leibniz University of Hannover, Hannover, Germany
e-mail: auer@tib.eu

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_12

object-oriented programming dominating the next two decades 1990s and 2000s, where large amounts of source code were not expressed in lengthy spaghetti-code but organized intuitively in objects and methods. This programming paradigm can already be seen as being inspired how our brain sees the real world—we learned concepts of abstract objects (abstract entities such as cars, trees, or buildings) and see their realization (or instantiation) in reality. Also, objects have certain characteristics (size, color, shape) and functions, which correspond to data and methods associated with the objects.

This, however, is not the end of the development. The problem with object-oriented programming is that functions and methods are more dominant and the data is often deeply hidden in the code or in data silos where the structure of the data is only known to a few experts. We currently see that there is increasing attention to data (e.g., in the context of big data, smart data, data science)—data is becoming more and more a first class citizen of computing. Still many challenges are ahead of us to realize the vision of cognitive data. We need to find and use more intuitive representations of data, which capture their structure and semantics in machine and human comprehensible ways, so that we develop a common understanding of the data along use cases, organizations, applications, value chains, or domains. Knowledge graphs, linked data, and semantic technologies (see e.g., also [1–5]) are good candidates in this regard and discussed in this article as a basis for realizing data spaces.

12.2 The Neglected Variety Dimension

The three classic dimensions of Big Data are *volume*, *velocity*, and *variety*. While there has been much focus on addressing the volume and velocity dimensions, the variety dimension was rather neglected for some time (or tackled independently). However, meanwhile most use cases, where large amounts of data are available in a single well-structured data format, are already exploited. The music plays now, where we have to aggregate and integrate large amounts of heterogeneous data from different sources—this is exactly the variety dimension. The Linked Data principles emphasizing the holistic identification, representation, and linking allow us to address the variety dimension. As a result, similarly as we have with the Web a vast global information system, we can build with the Linked Data principles a vast global distributed data space and efficiently integrate enterprise data. This is not only a vision, but has started and gains more and more traction as can be seen with the schema.org initiative, Europeana, or the International Data Spaces.

12.2.1 From Big Data to Cognitive Data

While there has been much focus on addressing the volume and velocity dimensions of Big Data, e.g., with distributed data processing frameworks such as Hadoop,

Spark, and Flink, the variety dimension was rather neglected for some time. We have not only a variety of data formats—e.g., XML, CSV, JSON, relational data, graph data, etc.—but also data distributed in large value chains, in different departments inside a company, under different governance regimes, data models, etc. Often the data is distributed across dozens, hundreds, or in some use cases even thousands of information systems.

An analysis by SpaceMachine[1] demonstrates that the breakthroughs in AI are mainly related to the data—while algorithms were devised early and are relatively old, only once suitable (training) datasets became available, we are able to exploit these AI algorithms. Another important factor of course is computing power, which, thanks to Moore's law, allows us to efficiently process after every 4–5 years data being a magnitude larger than before..

In order to deal with the variety dimension of Big Data and to establish a common understanding in data spaces, we need a lingua franca *for data moderation*, which allows to:

- Uniquely *identify* small data elements without a central identifier authority. This sounds like a small issue, but identifier clashes are probably the biggest challenge for data integration.
- *Map* from and to a large variety of data models, since there are and always will be a vast number of different specialized data representation and storage mechanisms (relational, graph, XML, JSON, etc.).
- Distribute modular data *schema definition* and incremental schema *refinement*. The power of agility and collaboration became meanwhile widely acknowledged, but we need to apply this for data and schema creation and evolution.
- Deal with *schema and data* in an integrated way, because what is a schema from one perspective turns out to be data from another one (think of a car product model—it's an instance for the engineering department, but the schema for manufacturing).
- Generate *different perspectives* on data, because data is often represented in a way suitable for a particular use case. If we want to exchange and aggregate data more widely, data needs to be represented more independently and flexibly, thus abstracting from a particular use case.

The Linked Data principles[2] (coined by Tim Berners-Lee) allow us to exactly deal with these requirements:

1. *Use Universal Resource Identifiers (URI) to identify the "things" in your data*—URIs are almost the same as the URLs we use to identify and locate Web pages and allow us to retrieve and link the global Web information space. We also do not need a central authority for coining the identifiers, but everyone can create his own URIs simply by using a domain name or Webspace under his control as

[1] http://www.spacemachine.net/views/2016/3/datasets-over-algorithms

[2] http://www.w3.org/DesignIssues/LinkedData.html

prefix. “Things” refers here to any physical entity or abstract concept (e.g., products, organizations, locations and their properties/attributes, etc.)
2. *Use http://URIs so people and machines can look them up on the web (or an intra/extranet)*—an important aspect is that we can use the identifiers also for retrieving information about them. A nice side effect of this is that we can actually verify the provenance of information by retrieving the information about a particular resource from its original location. This helps to establish trust in the distributed global data space.
3. *When a URI is looked up, return a description of the thing in the W3C Resource Description Format (RDF)*—as we have a unified information representation technique on the Web with HTML, we need a similar mechanism for data. RDF is relatively simple and allows to represent data in a semantic way and to moderate between many different other data models.
4. *Include links to related things*—as we can link between web pages located on different servers or even different ends of the world, we can reuse and link to data items. This is a crucial aspect to reuse data and definitions instead of recreating them over and over again and thus establish a *culture of data collaboration*.

As a result, similarly as we have with the Web a vast global information system, we can build with these principles a vast global distributed data management system, where we can represent and link data across different information systems. This is not just a vision, but currently already started to happen; some large-scale examples include:

- The *schema.org initiative*[3] of the major search engines and Web commerce companies, which defined a vast vocabulary for structuring data on the Web (and is used already on a large and growing share of Web pages) and uses GitHub for collaboration on the vocabulary.
- Initiatives in the cultural heritage domain such as *Europeana,*[4] where many thousands of memory organizations (libraries, archives, museums) integrate and link data describing the artifacts.
- The *International Data Spaces Initiative,*[5] aiming to facilitate the distributed data exchange in enterprise value networks, thus establishing *data sovereignty* for enterprises.
- The *National Research Data Infrastructure*[6] aiming to build a research data space comprising data repositories, ontologies, as well as exploration and visualization infrastructure for all major research areas.

Further similar initiatives started in other areas such as Open Government Data, Life-Science, or Geo-Spatial Data.

[3] http://schema.org
[4] https://www.europeana.eu
[5] https://www.internationaldataspaces.org
[6] https://www.nfdi.de/

12.3 Representing Knowledge in Semantic Graphs

For representing knowledge in graphs, we need two ingredients, unique identifiers and a mechanism to link and connect information from different sources. Universal Resource Identifiers (URIs) and subject-predicate-object statements according to the W3C RDF standard allow exactly this. As we can build long texts out of small sentences, we can build large and complex knowledge graphs from relatively simple RDF statements. As a result, knowledge graphs can capture the semantics and meaning of data and thus lay the foundation for data spaces between different organizations or a data innovation architecture within an organization.

12.3.1 Representing Data Semantically

Let us look how we can represent data semantically, so it can capture meaning, represent a common understanding between different stakeholders, and allow to interlink data stored in different systems.

Identifying Things The basis of semantic data representation is URIs—Universal Resource Identifiers. Similarly as every Web page has its URL (which you can see in the location bar of your browser), URIs can identify every thing, concept, data item, or resource. Here are some examples of some URIs:

- http://dbpedia.org/resource/LinkedIn—identifier for the entity LinkedIn (as described on the Wikipedia page),
- http://schema.org/Organization—identifier for the concept organization from the schema.org vocabulary,
- http://xmlns.com/foaf/spec/#term_lastName—identifier for the last name of a person from the FOAF vocabulary.

Everyone (who has a domain name or webspace) can coin its own URIs, so you do not rely on a central authority (e.g., GS1, ISBN) as with other identifier systems. Since every URI contains a domain name, information about the provenance and the authority coining the URI is built in the identifier. It is important to note that these URI identifiers can point to any concept, thing, entity, and relationship, be it physical/real or abstract/conceptual.

Representing Knowledge Once we have a way to identify things, we need a way to connect information. The W3C standard RDF follows simple linguistic principles: the key elements of natural language (e.g., English) sentences are subject, predicate, and object. The following subject-predicate-object triple, for example, encodes the sentence “InfAI institute organizes the Leipzig Semantic Web Day 2021”:

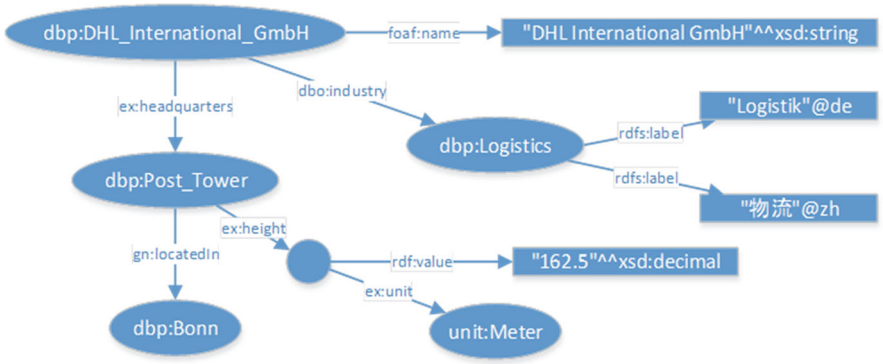


As you can see, we use an identifier http://infai.org for "InfAI institute," http://conf-vocab.org/organizes for the predicate "organizes," and http://infai.org/LSWT2021 as identifier for the object of the sentence "Leipzig Semantic Web Day 2021." As we connect sentences in natural language by using the object of a sentence as subject of further sentence, we can add more triples describing LSWT2021, for example, in more detail, by adding the start day and the location:



As you can see in this example, a small knowledge graph starts to emerge, where we describe and interlink entities represented as nodes. You can also see that we can mix and mesh identifiers from different knowledge bases and vocabularies, e.g., here the predicates from a conference vocabulary and the location referring to the DBpedia resource Leipzig. The start date of the event is here not represented as a resource (having an identifier), but as a Literal—the RDF term of a data value, which can have various data types (e.g., string, date, numbers).

Knowledge Graphs Build on these simple ingredients; we can build arbitrary large and complex knowledge graphs. Here is an example graph describing a company:



A knowledge graph [6] now is a fabric of concept, class, property, relationships, and entity descriptions, which uses a knowledge representation formalism (typically according to the W3C standards RDF, RDF-Schema, OWL) and comprises holistic knowledge covering multiple domains, sources, and varying granularity. In particular:

- *Instance data* (ground truth), which can be open (e.g., DBpedia, WikiData), private (e.g., supply chain data), or closed data (product models) and derived, aggregated data.
- *Schema data* (vocabularies, ontologies) and meta-data (e.g., provenance, versioning, documentation licensing) as well as comprehensive taxonomies to categorize entities.
- *Links* between internal and external data and *mappings* to data stored in other systems and databases.

Meanwhile a growing number of companies and organizations (including Google, Thompson Reuters, Uber, AirBnB, UK Parliament) are building their knowledge graphs to connect the variety of their data and information sources and build an data innovation ecosystem [7].

12.4 RDF a Holistic Data Representation for Schema, Data, and Metadata

A major advantage of RDF-based knowledge graphs is that they can comprise data, schema, and metadata using the same triple paradigm. Data integration is thus already built in into RDF, and knowledge graphs can capture information from heterogeneous distributed sources. As a result, RDF-based knowledge graphs are a perfect basis for establishing a data innovation layer in the enterprise for mastering digitalization and laying the foundation for new data-based business models.

RDF is a holistic data and knowledge representation technique, which allows to represent not only data but also its structure, the schema (and metadata) in a unified way. This is important, since what is schema from one perspective is data from another and vice versa. Let's look again at our company knowledge graph from the section. We can represent the entities DHL and PostTower in the graph as RDF triples consisting of subject, predicate, and object (the simple text syntax is called RDF Turtle):

```
DHL        rdf:type     Company
DHL        fullName     "DHL Int. GmbH"
DHL        inIndustry   Logistics
DHL        headquarter  PostTower
```

```
PostTower  rdf:type     Building
PostTower  locatedIn    dbpedia:Bonn
PostTower  height       "162.5"^^meter
```

From the triple representation, you can already see why RDF is superior for data integration, since it is already built in into the data model. In order to integrate RDF

data from different sources, we just have to merge the respective triples (i.e., throw them into a larger graph). Imagine if you have a multitude of different XML or relational data schemata, integrating those requires dramatically more effort.

The entities are now also assigned to classes: *DHL* is a *company* and *PostTower* a *building* (indicated by the *rdf:type* property). We can define these classes, by listing the properties, which can be used in conjunction with them and the respective expected types:

Class: Company

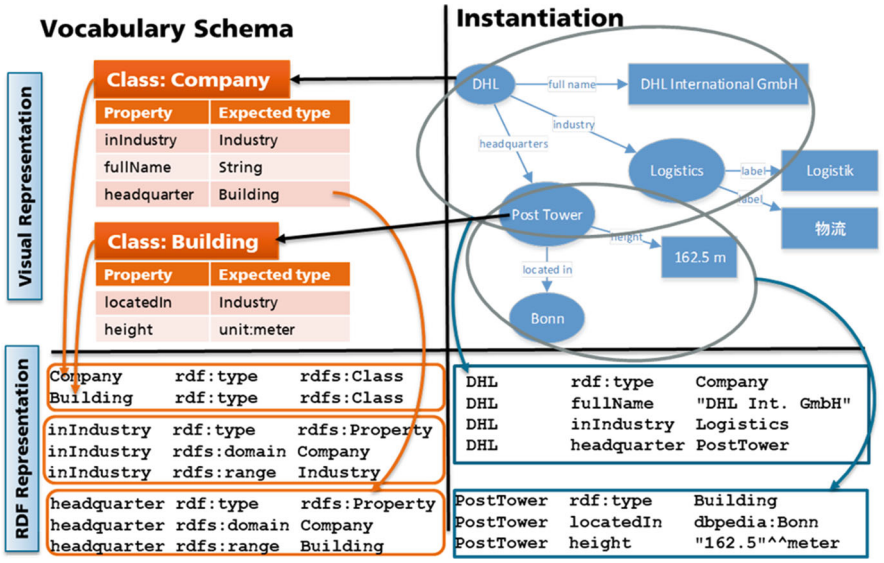
Property	Expected type
inIndustry	Industry
fullName	String
headquarter	Building

Class: Building

Property	Expected type
locatedIn	Industry
height	unit:meter

This means that the properties *inIndustry*, *fullName*, and *headquarter* are supposed to be used with instances of the class company and *inIndustry* should point to an instance of another class *Industry*, while the values assigned to the *fullName* property should be strings. The *schema.org initiative*[7] of the major search engines defines a large number of classes and associated properties, and meanwhile a large percentage of Web pages are annotated using the schema.org vocabulary. The following figure illustrates how schema and data can be both represented in triples:

[7] https://schema.org/



Although it sounds like a small thing, the integrated representation of schema and data in a simple, flexible data model is a key requisite for heterogeneous data integration. As such, RDF follows a bit the schema-last-paradigm—you do not have to define a comprehensive data model upfront, but can easily add new classes, attributes, and properties simply by using them. Also the URIs as unique identifiers and the fine-grained elements (triples) allow representing data, information, and knowledge in an integrated way while keeping things flexible.

12.5 Establishing Interoperability by Linking and Mapping between Different Data and Knowledge Representations

After *computer scientists were looking for the holy grail of data representation* for decades (remember the logic, ER/relational, XML, graph, NoSQL waves), it is now meanwhile widely accepted that *no single data representation scheme fits all* [8]. Instead we have a vast variety of data models, structures, systems, and management techniques, which all have their right to exist, since they are good serving one particular requirement (e.g., data representation or query expressiveness, scalability, or simplicity). As a result of this plurality, there is a *paramount importance for a systematic approach for linking and integration*. RDF fulfills exactly this requirement: it can moderate between different data models and evolve incrementally, as the original data sources and schemes change.

A unique and key feature of RDF is that it is perfectly suited to link and moderate between different data schemas, conceptual/data granularity levels, or data

management systems. Let me illustrate this with three examples how relational and taxonomic/tree data as well as logical/axiomatic can be represented in RDF.

The following diagram shows schematically how a relational table can be transformed into RDF triples. The rationale is that a (primary) key is used for generating URI identifiers (here the Id column) and columns are mapped to RDF properties (here "Title" to *rdfs:label* and "Screen" to *hasScreenSize*) and rows to RDF instances. The following example shows how the first database table row can be represented in RDF triples:

Electronics

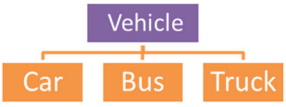
Id	Title	Screen
5624	SmartTV	104cm
5627	Tablet	21cm

Tabular/Relational Data

```
Prod:5624  rdf:type        Electronics
Prod:5624  rdfs:label      "SmartTV"
Prod:5624  hasScreenSize   "104"^^unit:cm
...
```

Of course, in large-scale applications, it is not required to actually physically transform the data to RDF (i.e., materialize the RDF view). Instead data can be transformed on demand when RDF links are accessed or RDF queries (e.g., in the SPARQL query language) executed. The W3C R2RML standard[8] provides a comprehensive mapping language for mapping relational data to RDF, which is meanwhile integrated into major DBMS (e.g., Oracle) and vendor-independent stand-alone mapping systems (e.g., eccenca's CMEM[9]).

The next example illustrates how taxonomic or tree data can be represented in RDF. Here the idea is to simply express each sub-taxon relationship with a respective RDF triple.



Taxonomic/Tree Data in RDF

```
Vehicle  rdf:type          owl:Thing .
Car      rdfs:subClassOf   Vehicle .
Bus      rdfs:subClassOf   Vehicle .
...
```

This example also demonstrates that RDF is perfectly suited as a data integration lingua franca, since its triples are small-grained building blocks, which can be combined in very different ways.

Another common type of information, which needs to be represented, is schema, constraint or logical information. Here is an example how this works and can be even augmented with logical axioms:

[8] https://www.w3.org/TR/r2rml/

[9] https://www.eccenca.com/en/products/eccenca-corporate-memory.html

$$\forall x : Human(x) \Leftrightarrow Male(x) \vee Female(x)$$

$$\nexists x : Male(x) \wedge Female(x)$$

Logical Axioms / Schema in RDF

```
Male     rdfs:subClassOf    Human .
Female   rdfs:subClassOf    Human .
Male     owl:disjointWith   Female .
...
```

First "Male" and "Female" are defined as subclasses of "Human," and the logical axiom of the OWL Web Ontology Language in addition states that both subclasses must be disjoint. This example illustrates that we can start initially with relatively simple representations and then iteratively add more structure and semantics as we go. For example, we can start expressing the data in RDF, then add schema (e.g., class and property definitions), then add more constraints and axioms, and so on—thus following the schema or here (logic/constraints) last paradigm.

These were only three examples; there are many more meanwhile standardized ways how to map, link, and integrate other data representation formalisms with RDF such as:

- *JSON Linked Data (JSON-LD)* for JSON.[10]
- *RDF in Attributes (RDFa)* for embedding/integrating RDF with HTML.[11]
- *CSV on the Web* for Tabular CSV data.[12]

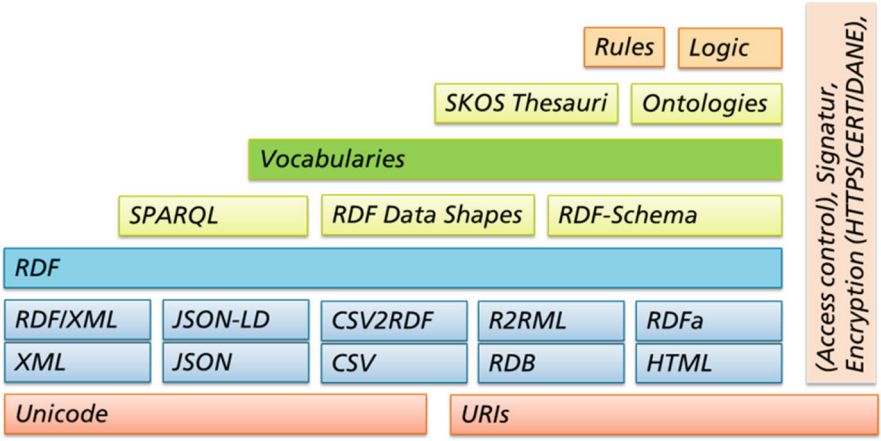
After the disgrace of its late birth (after XML) and the initial misconceptions of positioning/aligning RDF too much with XML (encoding the triple data model in XML trees) and heavyweight ontologies, meanwhile with Linked Data, JSON-LD, and knowledge graphs, a much more pragmatic and developer as well as application-friendly positioning of RDF is gaining traction. This is illustrated in an adaptation of the original Semantic Web layer cake from 2001,[13] where we now can see that RDF integrates well with many different data models and technology ecosystems:

[10] https://www.w3.org/2018/jsonld-cg-reports/json-ld/

[11] https://www.w3.org/TR/rdfa-primer/

[12] https://www.w3.org/TR/tabular-data-primer/

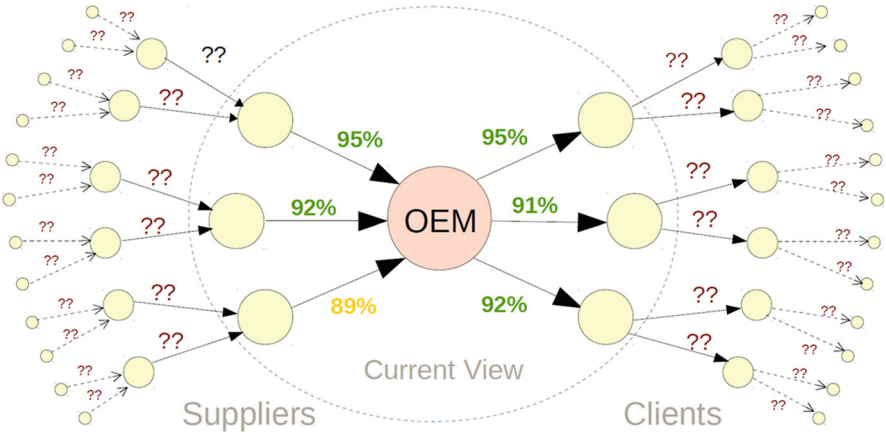
[13] http://www.w3.org/2001/10/03-sww-1/slide7-0.html



12.6 Exemplary Data Integration in Supply Chains with ScorVoc

Supply Chain Management aims at optimizing the flow of goods and services from the producer to the consumer. Closely interconnected enterprises that align their production, logistics, and procurement with one another thus enjoy a competitive advantage in the market. To achieve a close alignment, an instant, robust, and efficient information flow along the supply chain between and within enterprises is required. However, still much less efficient human communication is often used instead of automatic systems because of the great diversity of enterprise information systems, data governance schemes, and data models.

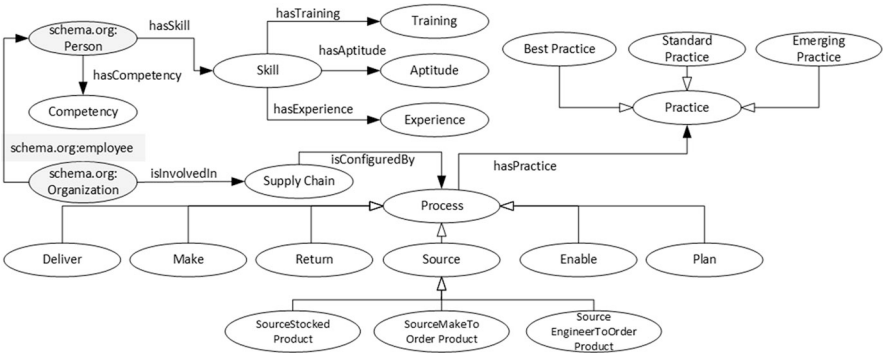
Automatic communication and analysis among various enterprises requires a common process model as a basis. The industry-agnostic Supply Chain Operation Reference (SCOR) [9], backed up by many global players (including IBM, HP, and SAP), precisely aims at tackling this challenging task. By providing 201 different standardized processes and 286 metrics, it offers a well-defined basis that allows describing supply chains within and between enterprises. A metric represents a KPI that is used to measure processes. The applicability of SCOR, however, is still limited, since the standard stays on the conceptual and terminological level and major effort is required for implementing the standard in existing systems and processes. The following figure represents a typical supply chain workflow.



Each node represents an enterprise and each arrow a connection. The values besides the connection can have many dimensions: the reliability of a delivery, the costs involved, and the time it takes to deliver from one place to another.

A semantic, knowledge graph-based representation of supply chain data exchanged in such a network enables the configuration of individual supply chains together with the execution of industry-accepted performance metrics. Employing a machine-processable supply chain data model such as the SCORVoc RDF vocabulary implementing the SCOR standard and W3C standardized protocols such as SPARQL, such an approach represents an alternative to closed software systems, which lack support for inter-organizational supply chain analysis.

SCORVoc [10] is an RDFS vocabulary that formalizes the SCOR reference model. It contains definitions for the processes and KPIs ("metrics") in the form of SPARQL[14] queries. A process is a basic business activity. The following figure gives an overview on the SCORVoc vocabulary.



[14] http://www.w3.org/TR/2010/REC-sparql11-query-20130321/

As an example, there are multiple delivery processes: *scor:SourceStockedProduct*, *scor:SourceMakeToOrderProduct*, and *scor:SourceEngineerToOrderProduct* depending on whether a delivery is unloaded into the stock, used in the production lines, or used for special engineered products. Each time such an activity takes place, all data that is needed to evaluate how well this process performed is captured such as whether the delivery was on time, whether it was delivered in full, or if all documents were included. Each process contains at least one metric definition. Due to its depth, we chose SCORVoc as our common data basis. The vocabulary is available[15] including definitions and localization for each concept.

Once partners in a supply network represent supply chain data according to the SCORVoc vocabulary, this data can be seamlessly exchanged and integrated along the supply chain. Due to the standardized information model, it is easy to integrate new suppliers in the network. Also, the knowledge graph representation allows to integrate supply chain data easily with other data (e.g., engineering, product, marketing data) in the enterprise. Due to the flexibility and extensibility of the RDF data model and vocabularies, such auxiliary data can also be easily added to the RDF-based data exchange in the supply network.

12.7 Conclusions

In this chapter, we have given an overview on the foundations for establishing semantic interoperability using established semantic technology standards by the World Wide Web Consortium, such as RDF, RDF-Schema, OWL, and SPARQL. These standards can be used to establish a data interoperability layer within organizations or between organizations, e.g., in supply networks. Using vocabularies and ontologies, participating departments (inside organizations) or companies (between organizations) can establish a common understanding of relevant data, by defining concepts as well as their properties and relationships. Establishing such a semantic layer and a common understanding of the data is crucial for realizing the vision of data spaces. Meanwhile, a number of data spaces are emerging. In addition to enterprise data spaces, this also includes data spaces in the cultural domain (e.g., the Europeana data space), in the government and public administration area, or in the research domain with the European Open Science Cloud or National Research Data Initiatives.

A deeper and wider penetration of semantic technologies in enterprises is still required in order to fully realize the potential of digitalization and artificial intelligence. While the semantic technology standards were already developed often more than a decade ago and the vision to leverage these for enterprise data integration was long discussed (cf. e.g., [11, 12]), mature and enterprise grade software platforms

[15] https://github.com/vocol/scor/blob/master/scor.ttl

(e.g., eccenca Corporate Memory[16]) started only to emerge in the last years. More work needs to be done to broaden the interoperability of semantic technologies with other enterprise technology ecosystems, such as property graphs or big data lakes. Also, traditional industry standardization methodologies need to shift to more agile interoperability paradigms following the Schema.org example of leveraging GitHub as an agile collaboration infrastructure.

References

1. Al-Safi, O., Mader, C., Lytra, I., Galkin, M., Endris, K. M., Vidal, M.-E., & Auer, S. (2018). Shipping knowledge graph management capabilities to data providers and consumers. In *12th IEEE International Conference on Semantic Computing, ICSC 2018, Laguna Hills, CA, USA, January 31–February 2, 2018* (pp. 9–16). IEEE.
2. Attard, J., Orlandi, F., & Auer, S. (2017). Exploiting the Value of Data through Data Value Networks. In *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2017, New Delhi, India, March 07–09, 2017* (pp. 475–484).
3. Grangel-González, I., Halilaj, L., Vidal, M.-E., Lohmann, S., Auer, S., & Müller, A. W. (2018). Seamless integration of cyber-physical systems in knowledge graphs. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC 2018, Pau, France, April 09-13, 2018* (pp. 2000–2003). ACM.
4. Grangel-González, I., Baptista, P., Halilaj, L., Lohmann, S., Vidal, M.-E., Mader, C., & Auer, S. (2017). The industry 4.0 standards landscape from a semantic integration perspective. In *22nd IEEE International Conference on Emerging Technologies and Factory Automation ETFA* (pp. 1–8). IEEE.
5. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J.E.L., Kirrane, S., Neumaier, S., Polleres, A., & Navigli, R. (2020). Knowledge graphs. arXiv preprint arXiv:2003.02320.
6. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). *A survey on knowledge graphs: Representation, acquisition, and applications*. IEEE Transactions on Neural Networks and Learning Systems.
7. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: Lessons and challenges. *Communications of the ACM, 62*(8), 36–43.
8. Mami, M. N., Graux, D., Thakkar, H., Scerri, S., Sören Auer, Jens Lehmann (2019). The query translation landscape: A survey. *CoRR* abs/1910.03118.
9. Supply Chain Council. Supply chain operations reference model (SCOR). Version 11. 2012.
10. Petersen, N., Grangel-Gonzalez, I., Coskun, G., Auer, S., Frommhold, M., Tramp, S., Lefrancois, M., & Zimmermann, A. (2016). SCORVoc: Vocabulary-based information integration and exchange in supply networks. In *10th IEEE International Conference on Semantic Computing (ICSC)*. IEEE.
11. Stephens, S. (2007). The enterprise semantic web. In *The semantic web* (pp. 17–37). Springer.
12. Song, F., Zacharewicz, G., & Chen, D. (2013). An ontology-driven framework towards building enterprise semantic information layer. *Advanced Engineering Informatics, 27*(1), 38–50.

[16] https://eccenca.com/products/enterprise-knowledge-graph-platform-corporate-memory

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 13
Data Ecosystems: A New Dimension of Value Creation Using AI and Machine Learning



Dirk Hecker, Angelika Voss, and Stefan Wrobel

Abstract Machine learning and artificial intelligence have become crucial factors for the competitiveness of individual companies and entire economies. Yet their successful deployment requires access to a large volume of training data often not even available to the largest corporations. The rise of trustworthy federated digital ecosystems will significantly improve data availability for all participants and thus will allow a quantum leap for the widespread adoption of artificial intelligence at all scales of companies and in all sectors of the economy. In this chapter, we will explain how AI systems are built with data science and machine learning principles and describe how this leads to AI platforms. We will detail the principles of distributed learning which represents a perfect match with the principles of distributed data ecosystems and discuss how trust, as a central value proposition of modern ecosystems, carries over to creating trustworthy AI systems.

D. Hecker (✉)
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Schloss Birlinghoven, Sankt Augustin, Germany

ML2R Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr, Schloss Birlinghoven, Sankt Augustin, Germany
e-mail: dirk.hecker@iais.fraunhofer.de

A. Voss
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Schloss Birlinghoven, Sankt Augustin, Germany

Kompetenzplattform KI.NRW, Schloss Birlinghoven, Sankt Augustin, Germany
e-mail: angelika.voss@iais.fraunhofer.de

S. Wrobel
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Schloss Birlinghoven, Sankt Augustin, Germany

ML2R Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr, Schloss Birlinghoven, Sankt Augustin, Germany

Institut für Informatik III, Rheinische-Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
e-mail: stefan.wrobel@iais.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_13

13.1 Introduction

In recent years, the effective use of machine learning and artificial intelligence (AI) has become a decisive factor for the competitiveness of individual companies and entire economies. Due to fundamental advances in available algorithms and computing power, given enough data, tasks that previously would have required human intelligence can now be automated, dramatically increasing efficiency and enabling completely new products and services. The availability of data, however, has proven to be a major roadblock for all but the largest and most digitally affine corporations. Typically, AI applications have used within-company data resources only, severely limiting the quality and feasibility of AI deployment in many cases.

The availability of a digitally sovereign, technically secure, and economically viable architecture for shared data spaces and digital ecosystems across different partners therefore is a fundamental game changer for the use of machine learning and artificial intelligence almost everywhere. Using a data space along the value chain or with complementary partners, new data sources can be brought into the construction of machine learning applications in a well-organized and, most importantly of all, self-determined fashion. Since availability and accessibility of data are always governed by the participants in the ecosystem, companies remain in control, thus enabling significantly more data to be made available to others.

Secondly, due to the semantic interoperability and unified data modeling of the data space, a sensible integration of different data sources, which is especially crucial for machine learning, can be performed with orders of magnitude less effort than in traditional project-wise approaches. With more and more complementary data sources becoming available, the quality and scope of machine learning results increase dramatically.

Furthermore, due to the inherently distributed nature of a data space, companies can now leverage the full potential of novel, distributed machine learning approaches. With these approaches it is no longer even necessary to combine all data at a central place; instead, they can be processed locally, at the edge of the data space nodes. This not only guarantees confidentiality of data but also affords considerable savings and potential for scalability.

Embedding machine learning and AI applications into data ecosystems using data space architectures thus not only makes AI applications accessible to organizations outside of the classical data-intensive digital sector, but also brings about significant benefits for those who have already deployed AI by allowing them to naturally move towards a distributed perspective on AI systems [1].

In the following sections, we will first (Sect. 13.2) give a brief overview of how artificial intelligence and machine learning applications are built today in the development cycle that centers on the use of the right data combined in the right way. We will then widen our perspective in Sect. 13.3, moving away from the technical development cycle and focusing instead on the platforms that are used to put these applications into deployment. In Sect. 13.4, we will then take a deep dive into one particular machine learning technology, distributed machine learning, that is

especially suited for AI systems and federated ecosystems and promises to deliver all the advantages of powerful machine learning modeling without centralized data storage. From thereon, in Sect. 13.5, we will generalize and discuss a more general architecture for machine learning in (distributed) digital ecosystems. Since trust is a core value proposition of these ecosystems, in Sect. 13.6 we will conclude the chapter with a brief discussion of the aspects of trustworthiness and AI applications.

13.2 Big Data, Machine Learning, and Artificial Intelligence

Big data became a hot topic in Europe around 2013 and was soon followed by data science, machine learning, deep learning, and artificial intelligences as trending topics. *Artificial intelligence* (AI) is a subfield of computer science, and machine learning and deep learning are now its most successful areas. M*achine learning* produces knowledge in the form of statistical models. There are different types of models, each equipped with a learning algorithm that optimizes the model automatically from training data. The most popular *models* are decision trees for classification tasks, regression curves for quantitative prediction tasks, clusters of similar data for pattern recognition, and artificial neural networks from deep learning [2, 3].

The recent success of machine learning is due to the volumes of data, and especially unstructured data, that can be stored in big data architectures: text, speech and audio, images and video, and streams of sensor data. *Deep learning* takes such data and trains artificial neural networks. They can make predictions and give recommendations. They can develop strategies for actions in games or to control a robot. They can generate images, texts, and language [4].

Data science combines methods from mathematics, statistics, and computer science for the discovery of knowledge in data. *Knowledge discovery* is also called *data mining*. Already in 1999 a process model, called CRISP-DM, was published as a step-by-step data mining guide. CRISP-DM stands for "Cross-Industry Standard Process for Data-Mining" and became an industry standard. The six phases of the process, shown in Fig. 13.1, are interlaced because data mining is an explorative process. Problem statements and the hypotheses initially formulated during *business and data understanding* may have to be adjusted when it turns out that the available data do not yield good enough models. Therefore it is important to identify many sources of high-quality data. Apart from data catalogs, data understanding can be supported by tools for descriptive statistics and visual analytics to find outliers, gaps, missing data, or underrepresented cases.

Data preparation involves data cleaning and transformations: generation of missing or derived data and removal of outliers, possibly also annotations and semantic enrichments. The latter transform the data into a standard vocabulary and guarantee seamless and ambiguous exchange between companies [5]. Data



Fig. 13.1 The CRISP-DM process

preparation is still the most time-consuming phase and may take up to 70 % of the entire process [6].

Modeling and evaluation are supported by dedicated machine learning tools. For training, many *learning algorithms* require data which is labelled with the correct results or with other semantic annotations. This often drives the cost of data preparation. Machine learning is a statistical approach, and its models only return approximate results. Therefore *evaluating* a model is mandatory. It is done by running the model on extra data, which has not been used for training, and comparing predicted and correct results. Data which is not representative and contain errors or prejudices lead to deficient models.

Not for all learning tasks there is enough data to train good machine learning models. Therefore, researchers are now investigating algorithms that can learn causal relations and supplementary models that can represent facts and symbolic knowledge [7, 8].

A trained and evaluated model is not yet an AI application or intelligent solution. Usually, the model is placed into a workflow and linked to components for data fetching, data preprocessing, and model invocation, and the workflow is embedded in coded components that act upon the model's results. For instance, in an email

filter, a model may classify an email as junk or no-junk, but ordinary code must be written to move the email into the corresponding folder. A trained model embedded in traditional code constitutes the *deployable* application or solution.

13.3 An Open Platform for Developing AI Applications

In the preceding section, we have been focusing on the abstract data science process. When developing and deploying applications, this process is typically carried out with the help of software toolkits with the view towards deployment platforms which we will have a look at in the following paragraphs.

In the marketplace, there are several toolkits and platforms for machine learning and data mining, both open source and free to use as well as commercial. They provide the discussed means to analyze and preprocess data, to generate and test models, and to wrap them with workflows for deployment in smart applications. But to industrialize machine learning, such platforms need to be combined with data sharing platforms like the IDS on the one hand and digital business platforms with AI services on the other hand. All big ICT enterprises provide such integrated environments. But each of them supports a particular language and a set of compatible libraries, such as SciKit Learn, TensorFlow, H2O, and RCloud. This imposes specific standards and interfaces which lock the users into the provider's particular ecosystem.

In this section, we want to focus on ACUMOS AI [9], a platform that avoids the lock-in effect. It is an open-source development by the Linux foundation, was adopted by the European flagship project AI4EU [10], and thus will set a standard in the way AI applications are governed in Europe. The key idea of the AI4EU ACUMOS platform is that many models are reusable. Models pre-trained with big data sets often can be transferred to specific tasks by post-training with additional, task-specific data. In artificial neural networks, for instance, this can be achieved by retraining an emptied last layer, which is responsible for the final results.

To maximize the reusability of machine learning models, ACUMOS AI separates the work of the machine learning specialists from that of the application developers as shown in Fig. 13.2.

Machine learning specialists who have explored, trained, and evaluated a model in their preferred machine learning toolkit can *onboard* it to ACUMOS AI (1 in Fig. 13.2). That means, the model is packaged into combinable micro-services and described in a catalog. The catalog also contains components for data access, data transformation, and complementary software. ACUMOS provides a design studio, where application developers can graphically connect components from the catalog without any coding and without knowledge of the components' interna. Thus, they can build training workflows to retrain models into so-called predictors and put them into application workflows (2). Workflows can also be *published* in the catalog where others can rate them or give more specific feedback (3). Application workflows, packaged into a docker image, can be *deployed* in an execution

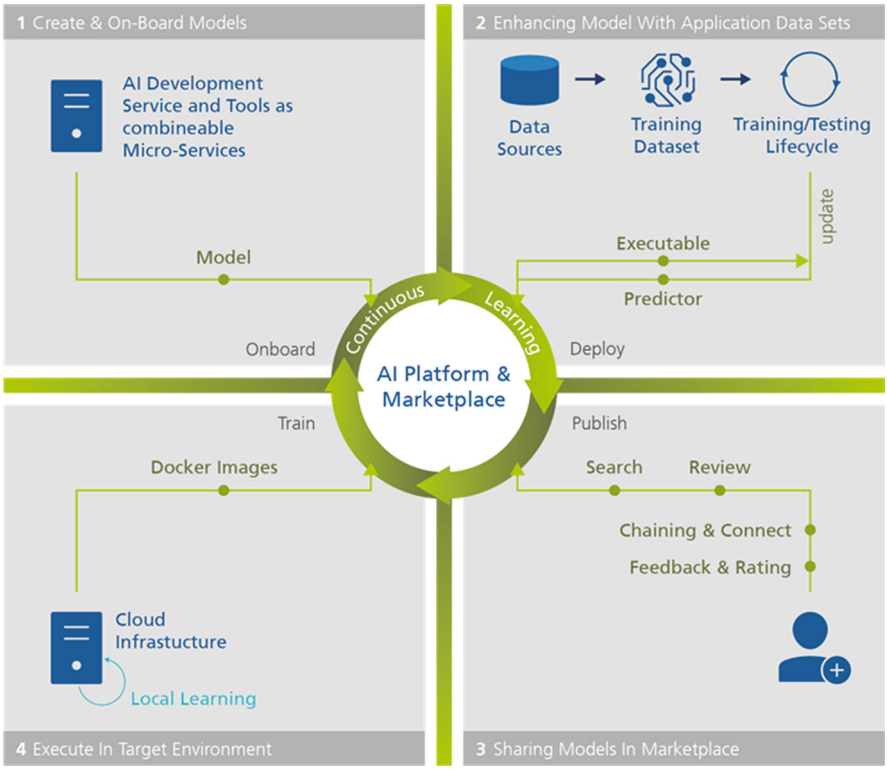


Fig. 13.2 Application building in ACUMOS AI

environment such as Azure, AWS, other popular cloud services, or any corporate data center or any real-time environment (4).

Within AI4EU, ACUMOS AI was fed with various AI models. Application developers can find them in the catalog and combine them into hybrid models [11]. ACUMOS AI is a federated platform, which means that users can access catalogs from different ACUMOS instances. The German node of the AI4EU-platform will be hosted by KI.NRW at Fraunhofer IAIS. KI.NRW acts as an umbrella for the transfer of AI from science into companies in North Rhine Westphalia [12].

In summary, ACUMOS AI facilitates collaboration of data scientists with different competences and roles: Experts who experimentally develop AI models with dedicated toolkits and application developers who possibly retrain them and graphically compose them into deployable AI applications. The data needed for machine learning models is identified, selected, and preprocessed by data managers. The IDS with its data connectors, data transformation apps, and semantic vocabularies provides an environment where this can be done in a controlled way.

13.4 Machine Learning at the Edge

Having discussed that today, machine learning algorithms will typically operate on data residing in distributed and federated platforms, a natural question is whether then the classical way of using machine learning by first centralizing all data can be improved by using the distributed data for machine learning where they are. Indeed this is possible and has been proven to work extremely well even for complex models [13], so let us have a brief look at how this works and what the benefits are.

Deep artificial neural networks can cope with complex learning tasks because they have many parameters, namely, a weight at every link (or synapsis) between two nodes at connected layers. Optimizing many weights requires lots of data [14]. Therefore, learning complex models usually takes place on a central server in the cloud, meaning that all data have to be transferred to this server. Usually, even the trained model remains in the cloud so that all application-specific data is moved to a central server. A popular example are speech assistants and translators.

However, there is a range of applications where such an approach is infeasible for legal reasons, to protect business secrets, or due to technical restrictions. An example for legal restrictions is the protection of personal data, specifically in the health sector. An example for business secrets are production data from machines, which the producing company does not want to disclose to a central service that is operated by the machine manufacturer for predictive maintenance or quality control. An example for technical restrictions are autonomous vehicles or driving assistants, which cannot transfer all data from the various cameras and sensors into a cloud in order to obtain piloting and navigation instructions.

Distributed machine learning solves this problem. A key idea is to learn at each of the distributed local data sources, which means at the edge of the cloud. Of course, a local node does not see much data; therefore, the second idea is to transmit the local model, rather than the local data, to the central server. Here the models are aggregated and redistributed to the local nodes. Thus all nodes indirectly profit from all data without ever exchanging it. For artificial neural networks, model aggregation means calculating the average of every parameter, which is a simple matrix operation. Figure 13.3 illustrates the communication between the central and local nodes.

All nodes continue training their model with new local data. For all nodes to continuously profit from their respective improvements, the local nodes need to exchange their updated models repeatedly. This can be done at fixed intervals or dynamically. The dynamic approach is more ambitious because here the nodes must somehow signal their progress. Distributed learning terminates when there is no more significant progress.

In 2016 researchers at Google were the first to successfully train a deep network with fixed synchronization intervals [15]. A team at Fraunhofer IAIS and Volkswagen elaborated this for dynamic synchronization and could show that the quality of the models suffered marginally while the communication effort could be reduced considerably [13].

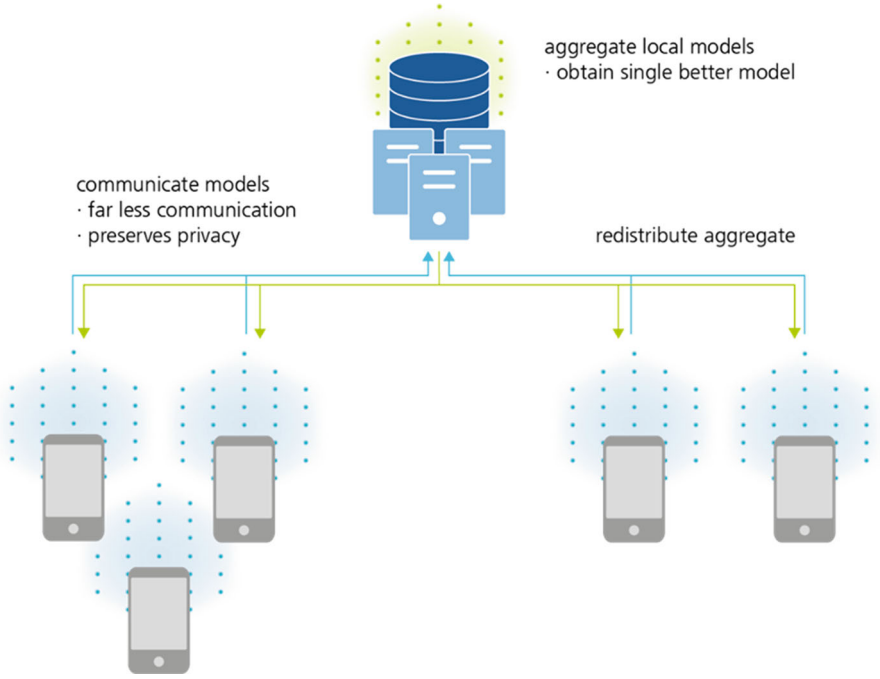


Fig. 13.3 Communication of models in distributed machine learning

13.5 Machine Learning in Digital Ecosystems

Having powerful algorithms for distributed and non-distributed machine learning and the right platforms for deploying their applications as discussed in the preceding sections, let us now zoom out again and look at the overall structure of digital ecosystem that results when these technologies are used and explain the different "spaces" in which digital value is then created.

Digital business ecosystems, in practically all domains such as mobility, healthcare, industrial production, logistics, and finance, will evolve around a shared data space and profit from AI to create value from the data by improving processes or products and by generating new business models.

In digital business ecosystems, there will be many data owners with similar kinds of data, like different hospitals with medical data records or different manufacturers of intelligent cars with data from the cars' cameras and other sensors. Probably, the data will come in more or less different formats. To benefit from the ecosystem means to pass one's data to the same applications. Therefore, the IDS provides vocabularies and data transformation apps so that data with the same semantics also ends up in the same format [16].

This is an ideal situation for machine learning, because data in standardized form will dramatically reduce the data preparation effort, which may take most of the

effort in first-of-a-kind projects. In a shared data space, machine learning can be industrialized. Not only are there input data in a standardized format; a model, which was originally trained on shared data, can also be reused and transferred. Data providers, who in principle are competitors, may collaborate in training a shared basic model and tune it with their individual data to their individual context and deploy it in their distinguished smart application. Such an industrialized way of machine learning calls for a good separation of work between the different kinds of data scientists: the machine learning and other AI experts who deliver models, experts for model adaptation and application building who deliver smart applications, and data managers and data engineers who find and transform the data into a ready-to-use form.

In Fig. 13.4, the realm of the data engineers is the data space. Data engineers use the data broker to find suitable data and the app store to find data transformation apps for the target vocabularies. The data transformation apps can be applied in connectors that expose data in a preprocessed standard format.

AI development environments are the realm of machine learning experts and application builders work. In Fig. 13.4 they operate in the development space. It sits on top of the data space layer because it can receive training data via connectors from the data space. The smart applications developed in the AI development space can be deployed in domain-specific digital business ecosystems. They reside at the top of Fig. 13.4 in the solution space. Smart applications can also be published as smart apps in the app store of underlying shared data space.

Machine learning at the edge is a special case, treated on the right-hand side of Fig. 13.4. Since here, learning is a continuous activity, smart applications must be "smart learning applications" which can learn in the execution environment by aggregating local models and redistributing them. The "edge space" in Fig. 13.4 is a specialized data space that extends down to special sensors, like IoT sensors for the internet of things, and energy-efficient hardware. Data must not leave the edge space. The edge space provides smart local learning apps and connectors to exchange models between the local learning apps and the smart learning application in the solution space.

13.6 Trustworthy AI Solutions

With descriptions of the basic machine learning technologies, the data science process, the platforms, and digital ecosystems in place in the preceding sections, let us now return to one of the core value propositions of modern federated data ecosystems, and in particular of the industrial data space (IDS) architecture [5] and GAIA-X [16]: Trust. While the ecosystem architecture focuses on trust in the data providers, the data consumers, brokers, and transport, here we want to focus on what it takes to make the AI applications built on top of these ecosystems trustworthy.

Modern machine learning methods are extremely powerful, but due to their data-driven nature and the extremely high dimensionality of their models, establishing

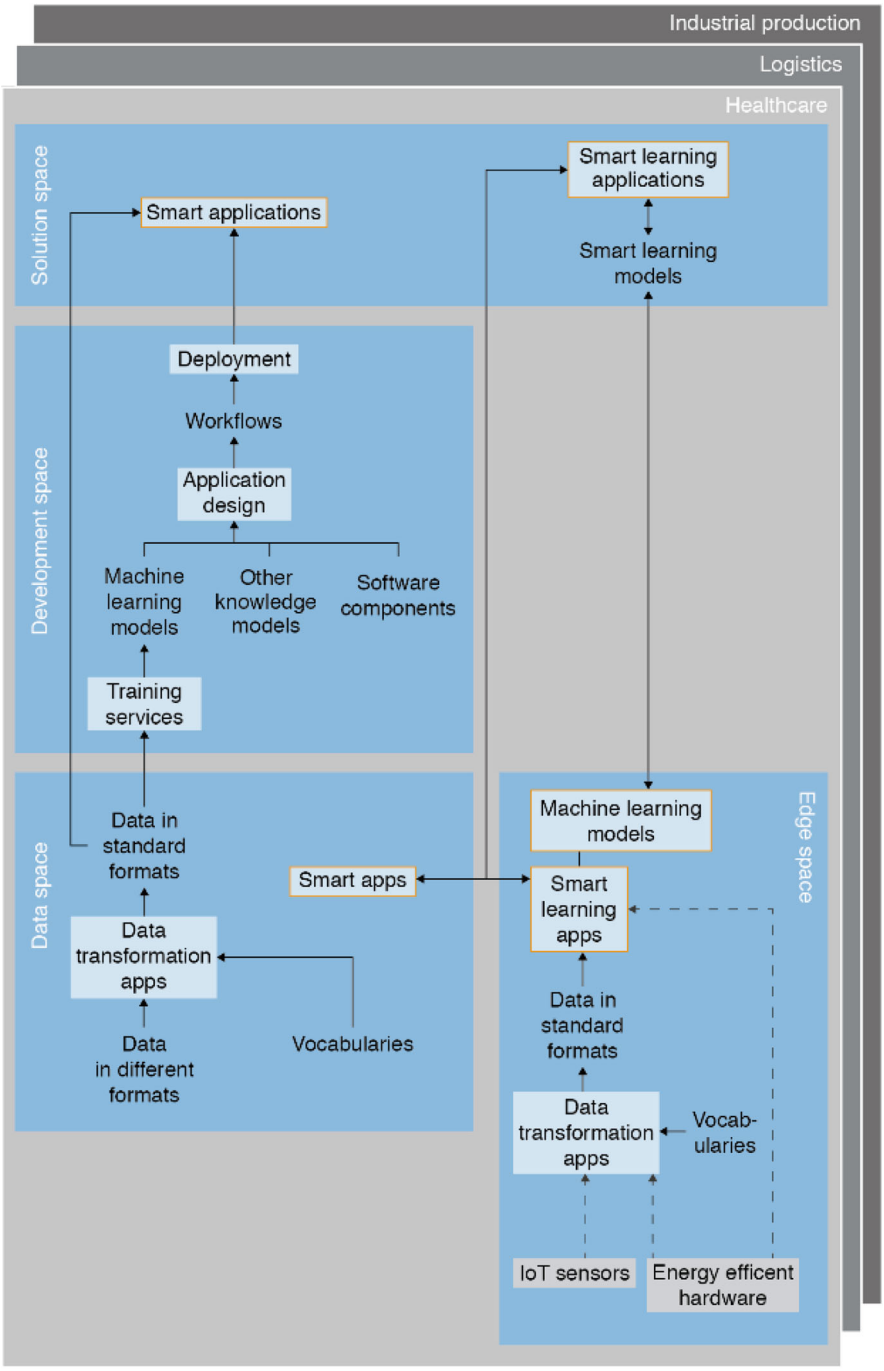


Fig. 13.4 Ai application, machine learning, and shared data spaces

trust in their results presents particular challenges. Artificial neural networks are a particularly good example of this. These networks are "black boxes" because they do not contain any code nor rules that humans could easily inspect. Moreover, all machine learning models return results that are not perfectly true or completely false, but more or less correct. Therefore, many models can also be made to output a confidence value which reflects their uncertainty. Finally, AI applications can be built to *learn continuously*, where the model is updated from new data or feedback during operation—with possibly unforeseen effects. Edge machine learning is only one type of continuous learning.

So it is difficult to argue that an application with machine learning inside will behave as intended. Figure 13.5 gives an overview of important principles that a *trustworthy application* of AI should incorporate [17].

By choosing and configuring the type of model, the data scientist tries to achieve accurate predictions. If the model is not robust against small changes, it can be improved by adding noise and other systematic transformations of the input data. Correcter and more robust models are more *reliable*. A black-box model can be supplemented with more explicatory models to facilitate its interpretation or defend individual results. Both increase the model's *transparency*. Other methods prevent a model from exposing any *private* data which may be encoded in the millions or even billions of weights.

Quality and quantity of the training data have a huge impact on the *reliability*, *privacy,* and *fairness* of the model because training data may be unrepresentative in general and contain wrong features or examples of low quality. This must be investigated in particular with respect to gender, religion, ethnicity, disability, and age to improve the *fairness* of the model. Of course, all data must comply with the data protection regulations and laws.

A model cannot sense, behave, or communicate. It is conventionally coded software wrapped around and controlling the model's invocation that determines the functionality and appeal of the application at the user interface. This software must be designed according to the desired level of control. The user's *agency and oversight* is high in smart assistant tools, can be decreased while keeping the user in the loop or on the loop, and is minimized in a fully autonomous system. Applications

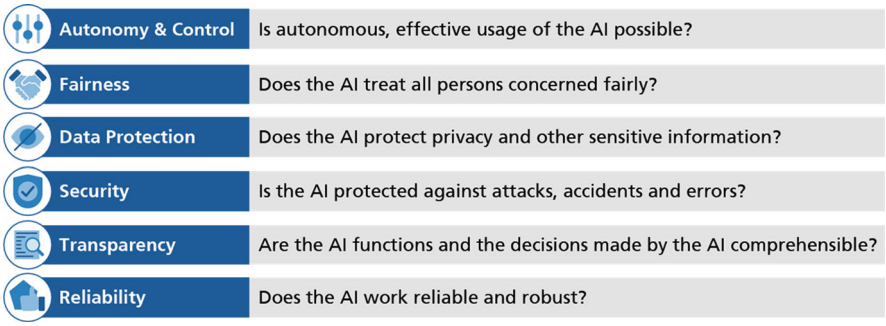


Fig. 13.5 Values to be respected by trustworthy AI applications

can be designed so convenient that users over-rely on them and unlearn important competences. Intelligent devices and robots, in particular, can be made to look and feel so human-like that users get overly attached and dependent. The embedding software also contributes to *reliability*. It must invoke the model only when the application context fits to the training data, and it must override the model's output when its confidence is low. For the worst cases, fail-safe procedures must be invoked. A final job of the embedding code is data logging so that failures of the application are documented and can be investigated, thus contributing to *transparency*.

Unfortunately, the requirements on trustworthy AI may be conflicting. Especially reliability may suffer when transparency, privacy, and fairness are improved. Moreover, creating a trustworthy AI application will be costly. Therefore, for each principle of trustworthiness, the risks of ignoring it must be assessed. The effectiveness of improvement measures should correlate to the risk, with low risks requiring no measures at all.

A European standard for auditing trustworthy AI applications could be a competitive advantage for European providers of AI software. The standard would have to balance costs against risks so that an AI certificate would be a competitive advantage and promote innovative trustworthy solutions, without raising to high barriers for market entry [18]. North Rhine Westphalia is supporting such an endeavor by Fraunhofer and partners. The so-called Bonner Katalog [19] will provide framework for certification based on the principles from Fig. 13.5 that elaborates the recommendations of the European High Level Expert Group on AI [20].

13.7 Summary

In this chapter, we have described how the arrival of modern federated data ecosystems acts as a driver that pushes forwards the use of artificial intelligence and machine learning technologies across all application areas. By making larger volumes of data from multiple partners available to all participants in such ecosystems in a trustworthy fashion, more and more companies will be capable of developing and/or deploying successful artificial intelligence systems. Moreover, as we have described in this chapter, recent developments in particular in distributed machine learning are a particularly good match for the environment that is provided by federated data ecosystems. Thus, we can expect that in the future, AI and machine learning will be a core part of any digital ecosystem in the manner that we have discussed above. This opens up the exciting prospect that federated data ecosystems will be the basis for a thriving economy that is characterized by fairness, competition, market orientation, and, thus, best possible value creation for enterprises and citizens alike.

References

1. Hecker, D., Koch, D. J., Heydecke, J., & Werkmeister, C. (2017b). Big-Data-Geschäftsmodelle – die drei Seiten der Medaille. *Wirtschaftsinformatik & Management, 8*(6), 20–30.
2. Döbel, I., Lies, M., Vogelsang, M. M., Neustroev, D., Petzka, H., Riemer, A., Rüping, S., Voss, A., Wegele, M., Welz, J. (2018). *Maschinelles Lernen* [online]. *Eine Analyse zu Kompetenzen, Forschung und Anwendung*. Sankt Augustin, Fraunhofer Gesellschaft. Accessed November 11, 2020, from https://www.bigdata.fraunhofer.de/de/big-data/kuenstliche-intelligenz-und-maschinelles-lernen/ml-studie.html
3. Hecker, D., Döbel, I., Rüping, S., & Schmitz, V. (2017a). Künstliche Intelligenz und die Potenziale des maschinellen Lernens für die Industrie. *Wirtschaftsinformatik & Management, 9*(5), 26–35.
4. Paass, G., & Hecker, D. (2021). *Künstliche Intelligenz - Was steckt hinter der Technologie der Zukunft?* Springer. ISBN ist: 978-3-658-30210-8.
5. Bader, S. & Oevermann, J. (2017). *Semantic annotation of heterogeneous data sources: Towards an integrated information framework for service technicians*. Proceedings of the 13th International Conference on Semantic Systems (pp. 73–80), SEMANTICS 2017, Amsterdam, The Netherlands, September 11–14.
6. Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing, 5*(4), 13–22.
7. ML2R. (2020). *ML2R*: *The competence center machine learning Rhine-Ruhr* [online]. Sankt Augustin: Fraunhofer IAIS. Accessed November 11, 2020, from https://www.ml2r.de/en/
8. von Rüden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., & Schuecker, J. (2020). Informed machine learning: A taxonomy and survey of integrating knowledge into learning systems. *arXiv, 1903*, 12394. arXiv preprint.
9. AT&T and The Linux Foundation. (2018). *Acumos*: *An open source AI machine learning platform* [online]. The Linux foundation. Accessed November 11, 2020, from https://www.acumos.org/wp-content/uploads/sites/61/2018/03/acumos_open_source_ai_platform_032518.pdf
10. AI4EU. (2020). *AI4EU*: *A European AI on demand platform and ecosystem*. THALES SIX GTS FRANCE SAS: Gennevilliers, France [online]. Accessed November 11, 2020, from https://www.ai4eu.eu/
11. Rehm, G., Bontcheva, K., Choukri, K., Hajic, J., Piperidis, S., and Vasiljevs A. (Eds.). (2020). *Towards an interoperable ecosystem of AI and LT platforms: A roadmap for the implementation of different levels of interoperability*. Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020) (pp. 96–107). International Workshop on Language Technology Platforms (IWLTP-2020) Marseille, France, 16.05.2020.
12. KI.NRW. (2020). *Kompetenzplattform KI.NRW* [online]: *Die zentrale Anlaufstelle für Künstliche Intelligenz in Nordrhein-Westfalen*. Sankt Augustin: Fraunhofer IAIS. Accessed November 11, 2020, from https://www.ki.nrw/
13. Kamp, M., Adilova, L., Sicking, J., Hüger, F., Schlicht, P., Wirtz, T., & Wrobel, S. (2018). *Efficient decentralized deep learning by dynamic model averaging*. Proceedings of ECML-PKDD, Dublin, Ireland, 14–16 September 2018.
14. Hecker, D., & Paaß, G. (2020). *Künstliche Intelligenz. Was steckt hinter der Technologie der Zukunft?* Springer. ISBN: 978-3-658-30211-5.
15. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Agüera y Arcas, B. (2017). *Communication-efficient learning of deep networks from decentralized data*. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017 (pp. 1273–1282), Fort Lauderdale, Florida, USA. JMLR: W&CP 54.

16. Bader, S., Pullmann, J., Mader, C., Tramp, S., Quix, C., Müller, A. W., . . . & Geisler, S. (2020). The international data spaces information model–An ontology for sovereign exchange of digital content. In *International Semantic Web Conference* (pp. 176–192). Cham: Springer.
17. Cremers, A., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., Rosenzweig, J., Rosstalski, F., Sicking, J., Volmer, J., Voosholz, J., Voss, A., & Wrobel, S. (2019). *Trustworthy use of artificial intelligence* [online]. Sankt Augustin, Fraunhofer IAIS. Accessed November 11, 2020, from https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf
18. Heesen, et al., (2020). https://www.plattform-lernende-systeme.de/aktuelles-newsreader/certification-of-ai-systems-plattform-lernende-systeme-names-challenges.html
19. Poretschkin, M., et al. (2021). *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz - KI-Prüfkatalog*. Fraunhofer IAIS. https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html
20. HLEG on AI. (2019). *Ethics guidelines for trustworthy AI* [online]. Brussels: European Commission. Accessed November 11, 2020, from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 14
IDS as a Foundation for Open Data Ecosystems



Fabian Kirstein and Vincent Bohlen

Abstract Open data is a popular and flourishing concept. The availability of open and structured data is the foundation of new business models, citizen engagement, and scientific research. However, open data still faces many issues to unfold its full potential, including usability, quality, legal, privacy, strategic, and technical barriers. In addition, the public sector remains its main provider, while industry stakeholders are still reluctant to participate in open data ecosystems. In this article, we present an architecture to overcome these drawbacks by utilizing the concepts, specifications, and technologies provided by International Data Spaces. We developed a prototype to demonstrate and evaluate the practical adoption of our architecture. Our work shows that IDS can act a vital foundation for open data ecosystems. The presented solution is available as open source software.

14.1 Introduction

Open data is a popular and continuously evolving concept, focusing on the open provision and reuse of structured datasets. Established publishers are public bodies, such as public administrations, research institutes, and nonprofit organizations. In addition, the relevance for private companies and industries as publishers of open data is increasing. Typical domains are traffic, weather, geographical, statistical data, and research data [1]. Open data creates transparency, supports innovations, and contributes to participation processes. The data should meet a set of key characteristics; in particular it should be complete, as raw as possible, up-to-date, accessible, free of charge, and machine-readable. Open data is mostly made available via web

F. Kirstein (✉)
Fraunhofer FOKUS, Berlin, Germany

Weizenbaum Institute for the Networked Society, Berlin, Germany
e-mail: fabian.kirstein@fokus.fraunhofer.de

V. Bohlen
Fraunhofer FOKUS, Berlin, Germany
e-mail: vincent.bohlen@fokus.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_14

portals, provided by data publishers [2]. Currently more than 2600 individual portals exist worldwide [3]. An integral element of open data is the aggregation of available data into centralized platforms to create single access points and harmonized views. An example for open government data is the European Data Portal (EDP), aiming to make available public sector information from across Europe [4]. A similar platform is operated by the European project OpenAIRE, gathering research data and publications from a plethora of distributed repositories [5]. However, open data is still facing numerous barriers on technical, legal, organizational, strategic, and usability levels [6]. In this article, we argue that the technologies, specifications, and methodologies of the International Data Spaces (IDS) can act as a powerful foundation to build and maintain open data ecosystems and improve and remedy existing challenges and issues. The core objective of our work is to lift open data into a unified and common data space. Therefore, we developed an IDS-based architecture, allowing the publication and dissemination of open data in a decentralized and timely fashion. We evaluated our work through a working prototype, thus demonstrating its practical feasibility. Finally, we discuss our findings and indicate directions for future developments. The main contributions of our work are:

1. An IDS-based software architecture and toolkit to build and operate open data ecosystems and to foster the adoption of open data in industry-driven scenarios.
2. A new paradigm for disseminating open data according to its often decentralized origins and provenance, facilitating more timely and sovereign access.

14.2 Barriers of Open Data

Barriers and issues in open data are well researched and constitute the principal motivation of our work. Beno et al. [6] collected extensive data about these barriers based on existing literature and an online survey. The authors distinguish between the users and the publishers of open data. The most relevant issues for users are the limited availability and functionality of machine-readable representations and interfaces for the data, outdated and obsolete data, and incomplete metadata. In addition, users reported challenges in searching and browsing, poor performance of the portals, and lacking information on data quality. Publishers are more concerned with legal aspects, ownership of the data, and potential loss of control. In addition, they seek resource-efficient methods to publish their data. In general, low data quality, the lack of a definite standard for describing metadata, and poor interoperability between data providers impede open data to meet its full potential [7].

A crucial reason for the current state of open data is the strong fragmentation and decentralization of data providers, although a single and central point of access and a unified data space are desired. This leads to a plethora of individual solutions, various implementations, and differing publication schemes. The current mitigation strategy is to set up centralized platforms to aggregate and unify open data, such as through Google Dataset Search [8]. However, these services may lead to undesired vendor lock-in and loss of data sovereignty.

14.3 Related Work

The technical implementation of open data is highly related to standardized data formats, open-source and propriety platform solutions, Semantic Web standards, and decentralization approaches.

An important foundation is the Data Catalog Vocabulary (DCAT). DCAT aims to describe meta-datasets meaningfully and seeks to increase interoperability between data catalogs. It is based on the Resource Description Framework (RDF) and is agnostic with regard to the actual data. Its core comprises three main classes: catalog, dataset, and distribution. One catalog consists of multiple datasets, and a dataset is referencing to one or more distributions, constituting a representation of the actual data [9]. The DCAT Application Profile for data portals in Europe (DCAT-AP) is an extension of DCAT, designed to describe public sector datasets. It is designed as a unifying standard for the publication of open data in Europe and extends DCAT with additional properties and mandatory controlled vocabularies. For instance, properties like language, spatial information, or file format can be aligned with pre-defined and centrally managed vocabularies that are formalized with the Simple Knowledge Organization System (SKOS) [10]. DCAT-AP is widely adopted in the European Open Data landscape with country-specific extensions emerging as well.

Many solutions exist for publishing open data on the Web and for making it accessible for both humans and machines. The open-source solution CKAN has grown to be the de facto standard in the public sector for creating open data catalogs, but is increasingly adopted by private companies too. CKAN offers many functionalities for managing the entire publication process and offers a wide range of plug-ins. Its API is well documented and provides an extensive way to access the data programmatically [11]. Another open source solution is uData, which puts a focus on social interaction and customization. It offers specific features for data reuse and community contributions, distinguishing it from other solutions [12]. A proprietary and closed source solution for creating open data platforms is OpenDataSoft. It is widely adopted and emphasizes interactive feature, for example, visualizations and the generation of extensive APIs for structured data [13].

With the increasing relevance and dissemination of open data throughout the Web, aggregation and federation services are becoming more crucial. Therefore, there is an intrinsic interest of the open data movement to implement methods and systems to deliver a unified and central point of access. Besides national and pan-national portals, like the EDP, the Google Dataset Search aims to make all open data, published worldwide on the Web, available in a central place. It employs structured metadata, embedded on web sites that are publishing open data. The Google search index is used to extract this metadata, process it, and aggregate it into a convenient, harmonized view [8]. Yet, it remains proprietary and does neither offer a machine-readable interface nor transparency for data publishers.

Besides straightforward harvesting approaches of the datasets, approaches that are more elaborate have evolved. One popular interface standard applied in open data is SPARQL, offering the possibility for federated queries. This allows the

retrieval of data from multiple endpoints simultaneously. Many other implementations and methods are available, but the solutions are not yet ready for production use. Rakhmawati et al. [14] compiled a comprehensive overview of existing research in this field. More recent research was conducted in the combination of peer-to-peer mechanisms and open data, especially regarding the rise of blockchain and distributed ledger technologies. Truong et al. [15] proposed a solution based on Hyperledger Fabric and the InterPlanetary FileSystem (IPFS) to improve the provision and integrity of open data. The tool Regerator follows the methodology of CKAN and effectively enables the creation of decentralized registries for open data, by employing smart contracts on the Ethereum blockchain [16]. Garcìa-Barriocanal et al. [17] sketched an architecture based on Ethereum, IPFS, and the decentralized database BigchainDB to facilitate a secure and trustworthy metadata repository.

Finally, the Dat project constitutes a vivid initiative, a range of open source tools, and a custom peer-to-peer protocol to handle decentralized publication and archiving of data across multiple organizations. Its core objective is to shift the provision of data from central and commercial services to self-controlled consortium networks. Currently it primarily targets the research community but can be applied for any kind of data [18].

However, these partly experimental decentralization approaches are highly complex. In addition, many of them do not take the existing specifications for open data into account. To the best of our knowledge, no work pursues the middle ground between centralized aggregation and decentralization to publish open data, like the methodology and architecture of IDS can offer. Additionally, existing approaches are not considering any connection to industry data processing and exchange.

14.4 International Data Spaces and Open Data

In principle, open data and International Data Spaces address different data sharing problems. While open data aims to create transparency, the central objective of IDS is to establish trust between companies and to ensure the secure handling of data. However, these two approaches are not mutually exclusive. For instance, companies utilizing the IDS to share confidential data could also consume open data to support data analyses or even publish non-confidential data under an open license. While the IDS offers various approaches to ensure data sovereignty of data providers, the underlying objective is to enable participants to exchange data. From a purely technical perspective, exchanging open data is no different from exchanging industry data. Therefore, IDS can be used to distribute both closed and open data.

In the following, we will present our IDS-based open data architecture and illustrate the mutual benefits, arising from this combination.

14.4.1 IDS as an Open Data Technology

The underlying concepts and technologies of open data and IDS are very similar. Both initiatives rely on metadata repositories to share information about the availability and accessibility of data. Such repositories store knowledge about participating data publishers, available data offers, and possible access restrictions, without the need to transfer actual data into central data hubs. Therefore, both concepts follow principles of decentralization and transfer metadata to and from central information access points, i.e., metadata brokers and open data portals follow the same basic conception. The actual data remains under control of the data publisher's infrastructure until a potential data user issues a request for it. Both open data portals and IDS metadata brokers act as gateways to their respective ecosystems, providing (comprehensive) information about the available data.

As introduced in our related work, open data uses Semantic Web specifications, like DCAT. Likewise, the IDS information model is based on Linked Data principles and DCAT. This makes the two systems easily compatible and ensures straightforward interoperability.

14.4.2 IDS Components in an Open Data Environment

The architectural similarities of open data and International Data Spaces allow a straightforward matching of artifacts. To build open data ecosystems based on the IDS specifications, the corresponding components and actors must be practically aligned. IDS connectors provide the means to publish and expose data, matching the role of a data provider in the domain of open data. IDS metadata brokers collect and provide the metadata about available data offers, taking the role of open data portals. An IDS open data ecosystem therefore consists of at least one IDS connector and at least one IDS metadata broker:

- **Open Data Connector**

  The IDS connector takes the role of an open data provider and becomes an open data connector. Each data publishing entity in an open data ecosystem applies an instance of the connector to announce availability and grant access to data resources. Data consumers request said data from the connector, which then responds by serving the actual data from internal data management systems. In contrast to other IDS connectors, usage policies or access restrictions are not necessary requirements for exchanging data under open licenses. Therefore, the open data connector's usage control features can be reduced to a minimum, allowing for easier configuration and handling.

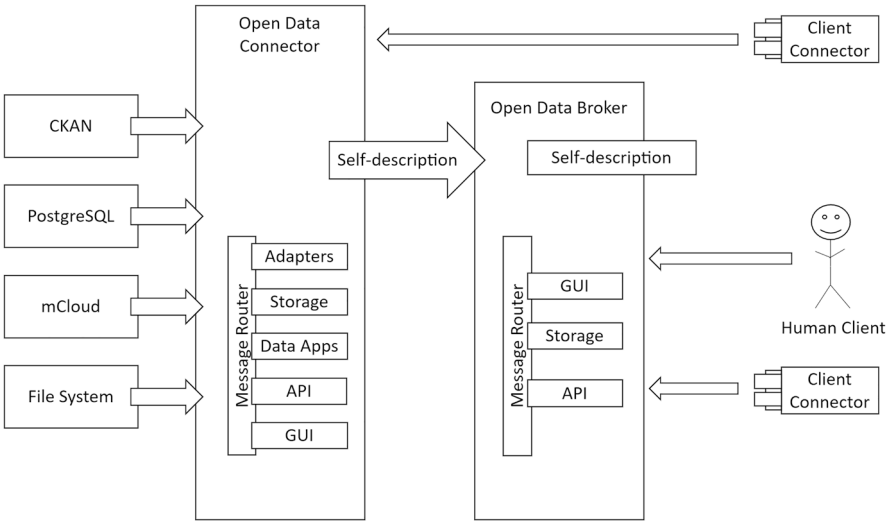


Fig. 14.1 Component overview and (meta)data flow in the IDS open data ecosystem

- **Open Data Broker**

  The IDS metadata broker fulfills a very similar role as an open data portal and becomes an open data broker. It represents a central entity, distributing information about what kind of data is available from which participant and which conditions apply for using the data. Data consumers use the metadata acquired from the broker to find and select their desired data and request the data directly from the publishing connector. As in the domain of open data, a broad variety of open data brokers can exist. Municipalities, e.g., on regional, country, or continent level, or private companies may operate their own open data broker. Different metadata brokers can also assemble data offers from specific domains, such as transport or energy.

Connectors and brokers communicate via well-defined IDS communication means, such as IDS messages, defined by the IDS information model, or the IDS Communication Protocol (IDSCP). This mechanism permits the IDS connector to pick the specific IDS metadata brokers to publish the metadata. It also provides the connector with the possibility to revoke the publishing of data while instantly informing the metadata broker about the changes made to the data offer.

An architecture overview of the described components is provided in Fig. 14.1. The figure depicts the (meta)data flow in the open data ecosystem. Additionally, an identity provider can be applied to ensure that the offered data is served from the expected publisher, tackling issues concerning the authenticity and integrity of data.

14.4.3 Benefits

Open data faces issues with data availability and metadata quality. Especially in open data portals that are harvesting large numbers of data sources, accessibility and overall usability depend on the metadata supplied by the original data providers. Standards like DCAT provide a common ground, but it remains within the responsibility of the data provider to follow these standards. IDS provides stricter specifications, not only covering metadata, but the entire communication process. This allows a much more harmonized communication and improved interoperability. Hence, the number of datasets published with insufficient metadata can be reduced significantly, and unified data exchanges between publishers and portals can be established.

In traditional open data environments, availability of datasets can be a problem if data publishers cannot ensure the accessibility of said data. As a result, open data portals must deal with unavailable data, dead links, or server limitations on the provider's side. To counteract these problems, portals periodically harvest the publisher's data catalogs and perform availability checks to confirm the data is still reachable. In the IDS open data ecosystem, the responsibility to keep the information at the portal/broker up to date is reversed. The open data connector informs the open data broker about currently available or updated datasets. Whenever changes occur to the actual data or the metadata, the broker is informed about the changes and adapts the information published accordingly. The "pull" approach of responsibility in traditional open data environments is inverted to a "push" approach in the IDS open data ecosystem. This approach does not only shift the responsibility of housekeeping the data offers from the portal/broker to the data publisher but presents the data publisher with new possibilities to control the initial placement of said data offers. Traditionally, publishers make their datasets available for download on their infrastructure, and the data is harvested by different open data portals. With IDS, the publisher chooses which metadata broker to register to, giving the data publisher additional data sovereignty. Obviously, this control ends when the data is downloaded—as there are no usage controls applied in the open data domain. For the data consumers, this approach can lead to improvements regarding the findability of timely data and reduces fragmentation. A fully developed IDS open data ecosystem will, thus, act as one virtual data space and removes the need to search for data in different places.

An exemplary comparison of the two dissemination approaches is illustrated in Fig. 14.2, modelling the differences between data flows. While open data and industry data are conceptually distinct in various aspects, the core concept is equally relevant for both domains. Consequently, an alignment of the two types of data on a technological level is highly beneficial. New possibilities for integration and reuse emerge if open data can be acquired, handled, and processed with the same tools and applications that are already applied in industry. The entire process of using, analyzing, and creating value from open data can be made more efficient and lower the entry barriers to using and publishing open data. On the other hand, it

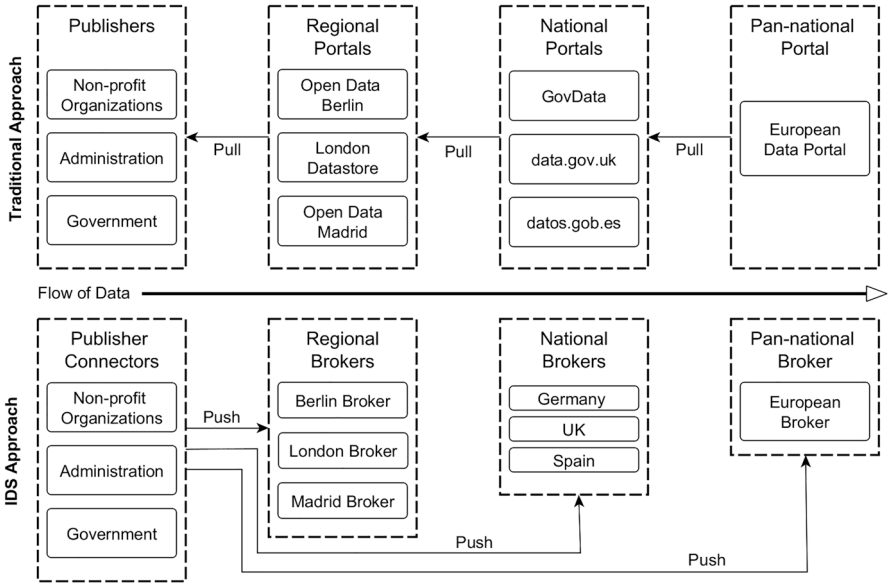


Fig. 14.2 Comparison of different data flows

can lower the barrier for industries to enter the IDS in the first place and act as a catalyst for adoption. Since legal hurdles and complex usage control are negligible when dealing with open data, it can also encourage interested parties to start sharing data via IDS. This may be followed by industry use cases. The following table provides a summary of how IDS can mitigate typical open data barriers:

Open data barrier	IDS solution
Limited availability of machine-readable representations and interfaces	Metadata exchange between data publishers and open data portals via the highly standardized IDS information model
Challenges in searching and finding open data	The virtual data space can lower fragmentation and supports the creation of a single point of truth, thus simplifying the search for suitable data
Potential loss of control	Data sovereignty through the ability to freely choose the brokers for publication and revocation of data offers
Poor interoperability between data providers	Data sources (open or closed) operating under the same standardized IDS conditions. Support for direct industry contributions
Central and timely point of access	The IDS "push" approach allows more control and immediate updates to central access points

14.5 The Public Data Space

The Public Data Space is a practical implementation and prototype of the aforementioned IDS open data ecosystem. It consists of an open data connector used to publish offers of open data to the ecosystem and an open data broker to collect and present metadata about available data offers. The Public Data Space successfully establishes a link between IDS and the Web by publicly exposing the open data offers. In addition, it targets human users through a full-fledged open data frontend for convenient browsing, searching, and querying. The components are published under the open source Apache 2.0 license, and the source code is available on GitHub.[1]

The following section will introduce the components in more detail and present an open government data use case as an example.

14.5.1 The Open Data Connector

The Open Data Connector is an implementation of the IDS connector specification with the aim of providing data owners an easy to use, out-of-the-box solution for publishing their data as open data. The connector is built to seamlessly interact with the Open Data Broker and offers functionality to communicate within the realm of International Data Spaces while simultaneously offering technology-agnostic ways to retrieve the published data.

The Open Data Connector is extensible and able to connect to any standard or proprietary data storage solution via data source adapters. Currently, paradigmatic implementations of adapters for SQL queries to PostgreSQL databases, the CKAN API metadata schema, and the mCLOUD[2] API metadata schema are available. Data is not replicated inside the connector but is retrieved on demand from the registered data sources. API guidelines for the development of data source adapters are provided as an OpenAPI[3] specification and can be easily applied to provide access to the desired data source solution.

An overview of the Connector's internal component architecture is presented in Fig. 14.3. The Open Data Connector offers two ways of communication: (1) it accepts IDS multipart messages and is able to respond to incoming requests for artifacts and self-descriptions and (2) additionally, the connector offers HTTP GET endpoints to retrieve the offered datasets without the technical barrier of having to set up a consuming IDS connector. This design decision was made to reflect the open nature of the offered resources and avoid imposing additional technological barriers on the retrieval of data.

[1] https://github.com/public-data-space

[2] https://www.mcloud.de/

[3] https://www.openapis.org/

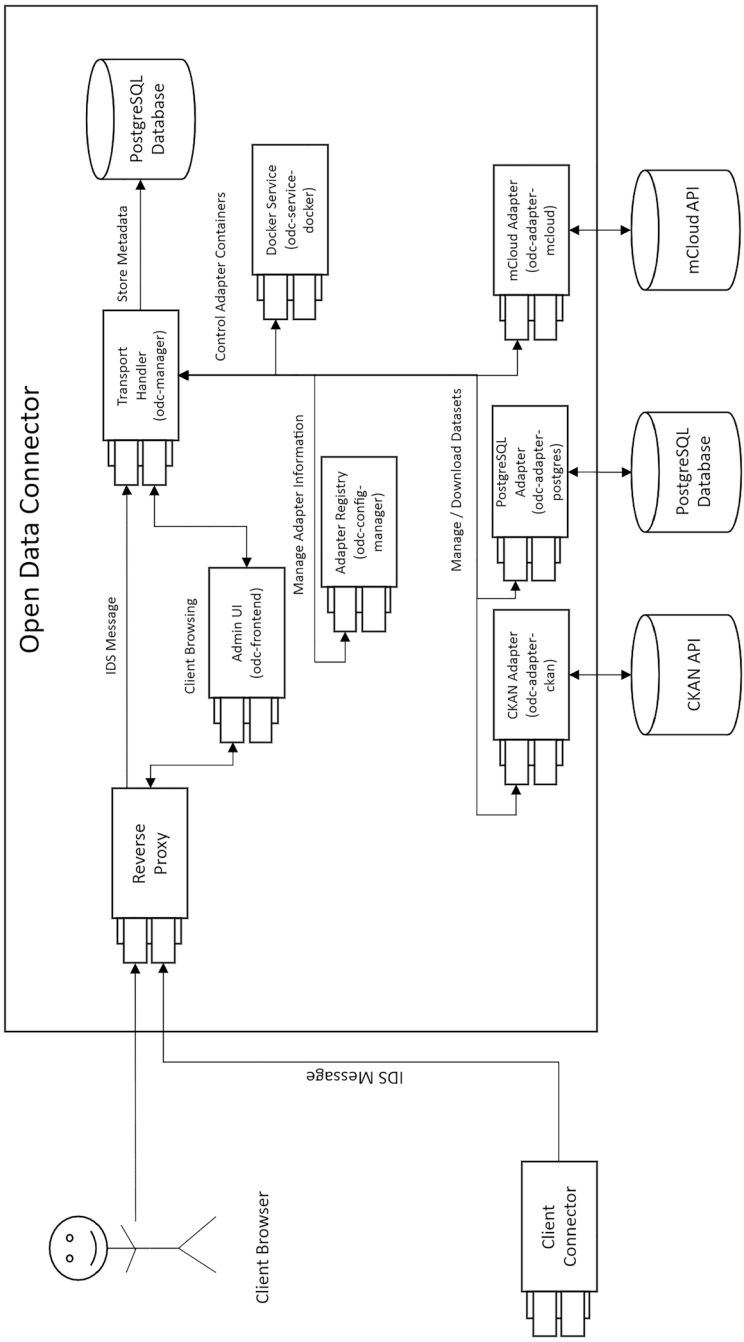


Fig. 14.3 Component architecture of the Open Data Connector

There exist no particular technical differences in consuming open data compared to any other type of data published via IDS. Therefore, the Open Data Connector is not intended to consume data and does not contain such functionalities. Data from Open Data Connectors should be consumed, using the same connectors already set up to retrieve commercial data for a given use case.

14.5.2 The Open Data Broker

The Open Data Broker is an implementation of the IDS metadata broker specification. In contrast to existing IDS metadata broker implementations, the Open Data Broker is not only offering endpoints for IDS components to consume and navigate the data offerings, but is equipped with the functionality of Open Data portals. Human users can search and browse the registered data using a modern web frontend without the need to deal directly with the IDS interfaces, used in the backend of the portal. Therefore, the HTTP GET endpoints of the Open Data Connector are used. This allows users of the Broker's portal to directly download data from the connectors through the user interface. The Broker's open data portal functionality is based on the open data management platform piveau [19]. Piveau offers comprehensive functionalities to store and search metadata based on Linked Data, as well as a modern portal web frontend.

In addition, the Open Data Broker offers the IDS endpoints specified in the IDS metadata broker specification. Via the */infrastructure* endpoint, the Broker accepts IDS multipart messages and is able to respond to incoming requests to register or unregister a connector or dataset and handles self-description requests. Via the */data* endpoint, the Broker accepts IDS query messages containing SPARQL payloads. Figure 14.4 presents an overview of the Broker's internal component architecture.

The Open Data Broker differs from other IDS metadata brokers in that it provides a web frontend specific to the needs of open data portals and by offering optimized interoperability with the Open Data Connector. Furthermore, the IDS information model was extended with additional metadata fields from the DCAT specification. Both components make use of this additional metadata and allow an improved user experience and more fine-grained information displayed at the Broker.

14.5.3 Use Case: Publishing Open Government Data

The Public Data Space is employed as a demo showcase and publicly available since December 2020.

The domain of open data is largely defined by an innumerable amount of different data publishers and data portals. The showcase has been set up to replicate this environment and to demonstrate how the Public Data Space solution runs in a production environment. The demo showcase consists of ten Open Data Connectors

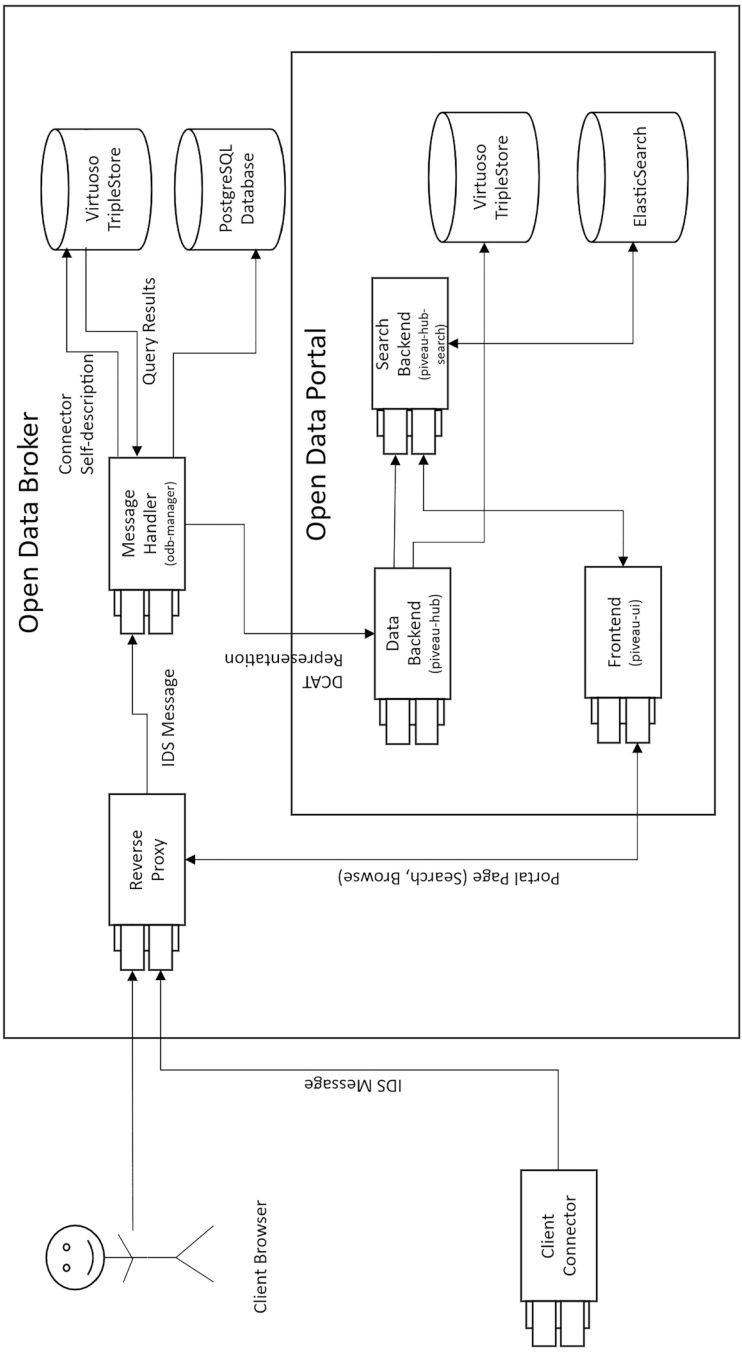


Fig. 14.4 Open Data Broker component architecture

and one Open Data Broker. The connectors have been set up to simulate one specific municipal data owner publishing their data. The connectors expose existing open data offers, by connecting to established open data portals, such as Berlin Open Data, Deutsche Bahn Open Data, and Open.NRW Köln. They are registered to an instance of the Open Data Broker representing a central open data portal for open government data. As such, the portal created in this showcase resembles existing portals such as GovData or the European Data Portal in the way they are combining data offers from a variety of different publishers and smaller portals. Each Open Data Connector publishes 30–50 exemplary datasets totaling 318 datasets registered at the Open Data Broker. Figure 14.5 shows the web frontend of the Open Data Broker in the open government data demo showcase. With this use case, we demonstrate that the Public Data Space solution can replicate the existing open data domain of data publishers and open data portals. Furthermore, the use case shows that communication between open data publishers and open data portals can be achieved by utilizing IDS. The scenario confirms the Open Data Connector's and Open Data Broker's ability to exchange IDS messages, allowing the Connector to exert full control over the metadata registered at the Broker and enabling the benefits presented in Sect. 14.4. In addition, the Open Data Connector showed to be fully functional in providing data

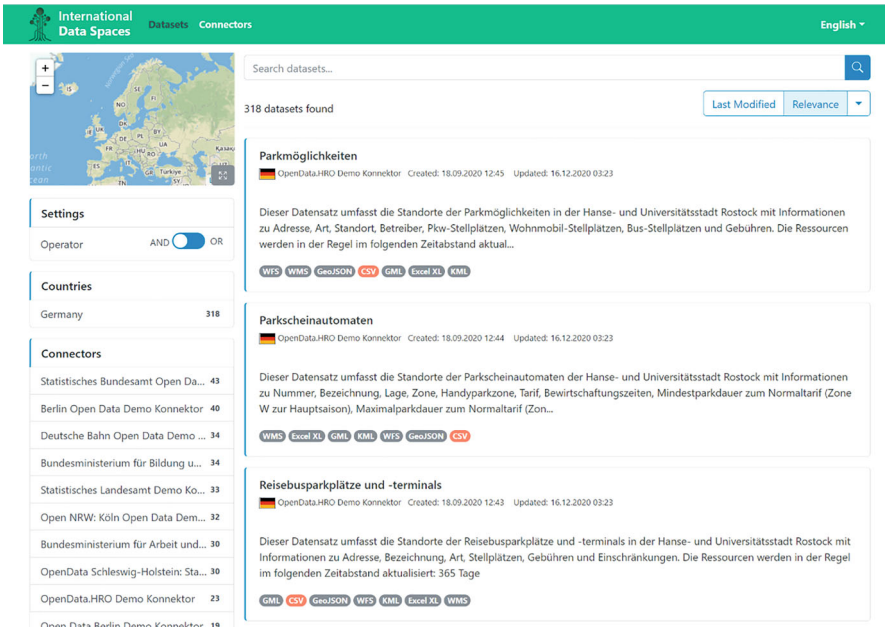


Fig. 14.5 Open Data Broker with published open government data

to a consuming instance of the Enterprise Integration Connector[4] and an instance of the DataSpace Connector[5] during the IDSA Plugfest activities in August 2020.

14.6 Discussion and Conclusion

In this paper, we have investigated and demonstrated how the specifications and artifacts of the Industrial Data Spaces can act as a powerful foundation for building and fostering open data ecosystems. We have designed a comprehensive IDS-based architecture and a working prototype to create a Public Data Space, where both public and private organizations can disseminate open data in decentralized, timely, and self-determined fashion.

Open data is currently mostly provided by actors from the public domain and rarely by private companies. Yet, it holds great economic and social potential. A variety of open specifications and software solutions exists for publishing open data. However, it still faces many barriers, mostly regarding data quality and availability. The IDS architecture can be a powerful building block to mitigate these barriers and enable the industry to become an open data user and provider. Many IDS use cases can benefit from using open data, e.g., using traffic data in a digital supply chain. In addition, the engagement with open data in the context of IDS can create awareness about it, leading to the creation of industrial open data.

IDS offer a selection of specifications and artifacts for trustworthy data sharing. Although many features relate to confidential data sharing, the very core mechanisms are suitable matches for creating open data ecosystems. This includes the well-defined information model, the principle of decentralization, and the standardized communication protocols. In addition, the IDS Connector-Broker architecture allows implementing a novel approach for making open data accessible. The data providers actively push their data offering to a central broker platform while maintaining sovereignty. This increases availability and timeliness in comparison to pull or harvesting mechanisms, applied by central services, like the Google Dataset Search or the European Data Portal. Hence, it maps the decentralized nature of open data to a technical solution while avoiding the complexity of fully decentralized approaches, like peer-to-peer methods. In addition, both IDS and open data foster the same metadata standards (DCAT).

Our Public Data Space prototype applies and consolidates the IDS artifacts information model, connector, broker, and data source adapters to an out-of-the-box solution for IDS-conform data publishing. It currently can integrate and publish data from four sources: CKAN, PostgreSQL, file system, and the open data platform mCloud. In addition, the entire solution is available as open source and acts as a

[4] https://www.dataspaces.fraunhofer.de/en/software/connector/ei_connector.html

[5] https://www.dataspace-connector.io/

transparent implementation of IDS artifacts, representing a starting point for potential IDS open data providers or other IDS use cases.

A successful broader adoption of our solution will require an adoption by actual providers of open data from the public and private domain. Our solution can be integrated on top of existing platforms, in principle allowing a smooth migration. Yet, this process requires substantial efforts and the support of the current actors of open data. We believe that with the rising popularity of International Data Spaces, the demand for integration and adoption in domains beyond industry data sharing will increase.

In future work, we aspire a Base Security Profile certification of our connector, adding an additional layer of trust to open data. In addition, we plan to extend the application domain towards the publication and sharing of open research Data. In addition, the Open Data Broker's interoperability with consuming IDS connectors will be demonstrated further in upcoming IDSA Plugfest activities. Finally, we will provide an extended data model to offer a richer, semantic description of the available data.

References

1. Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management, 29*(4), 258–268. https://doi.org/10.1080/10580530.2012.716740
2. Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., et al. (2018). The open data landscape. In Y. Charalabidis, A. Zuiderwijk, C. Alexopoulos, et al. (Eds.), *The world of open data: Concepts, methods, tools and experiences* (pp. 1–9). Springer International Publishing.
3. Opendatasoft. (2021a). Open data inception - A comprehensive list of 2600+ open data portals in the world. In *Open Data Inception*. Accessed January 25, 2021, from https://opendatainception.io/
4. Kirstein, F., Dittwald, B., Dutkowski, S., et al. (2019). Linked data in the European data portal: A comprehensive platform for applying DCAT-AP. In I. Lindgren, M. Janssen, H. Lee, et al. (Eds.), *Electronic government* (pp. 192–204). Springer International Publishing.
5. Manghi, P., Manola, N., Horstmann, W., & Peters, D. (2010). An infrastructure for managing EC funded research output: The OpenAIRE project. *The Grey Journal (TGJ): An International Journal on Grey Literature, 6*(1), 31–40.
6. Beno, M., Figl, K., Umbrich, J., & Polleres, A. (2017). Perception of key barriers in using and publishing open data. *JeDEM - eJournal of eDemocracy and Open Government, 9*, 134–165. https://doi.org/10.29379/jedem.v9i2.465
7. Neumaier, S., Thurnay, L., Lampoltshammer, T. J., & Knap, T. (2018). *Search, filter, fork, and link open data: The ADEQUATe platform: Data- and community-driven quality improvements*. In Companion Proceedings of the Web Conference 2018 (pp 1523–1526). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.
8. Noy, N., Burgess, M., & Brickley, D. (2019). *Google dataset search: Building a search engine for datasets in an open Web ecosystem*. In 28th Web Conference (Web Conf 2019).
9. W3C. (2020). *Data Catalog Vocabulary (DCAT) - Version 2*. Accessed March 10, 2019, from https://www.w3.org/TR/vocab-dcat-2/

10. European Commission. (2021). About DCAT application profile for data portals in Europe. *Joinup*. Accessed February 10, 2019, from https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/about
11. CKAN Association. (2021) *CKAN*. Accessed March 9, 2019, from https://ckan.org/
12. Etalab. (2021). *uData Documentation*. Accessed February 5, 2021, from https://udata.readthedocs.io/en/latest/
13. OpenDataSoft. (2021b) OpenDataSoft - Turn your data into actions. In *OpenDataSoft*. Accessed March 17, 2019, from https://www.opendatasoft.com/
14. Rakhmawati, N. A., Umbrich, J., & Karnstedt, M., et al., (2013). Querying over Federated SPARQL Endpoints: A state of the art survey. *arXiv*, 13061723 [cs].
15. Truong, D.-D., Nguyen-Van, T., Nguyen, Q.-B., et al. (2019). Blockchain-based open data: An approach for resolving data integrity and transparency. In T. K. Dang, J. Küng, M. Takizawa, & S. H. Bui (Eds.), *Future data and security engineering* (pp. 526–541). Springer International Publishing.
16. Tran, A. B., Xu, X., Weber, I., et al. (2017). Regerator: A registry generator for Blockchain. *CEUR Workshop Proceedings, 1848*, 81–88.
17. García-Barriocanal, E., Sánchez-Alonso, S., & Sicilia, M.-A. (2017). Deploying metadata on Blockchain technologies. In E. Garoufallou, S. Virkus, R. Siatri, & D. Koutsomiha (Eds.), *Metadata and semantic research* (pp. 38–49). Springer International Publishing.
18. Robinson, D. C., Hand, J. A., Madsen, M. B., & McKelvey, K. R. (2018). The Dat project, an open and decentralized research data tool. *Scientific Data, 5, 180221*. https://doi.org/10.1038/sdata.2018.221
19. Kirstein, F., Stefanidis, K., Dittwald, B., et al. (2020). Piveau: A large-scale open data management platform based on semantic web technologies. In A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, et al. (Eds.), *The semantic web* (pp. 648–664). Springer International Publishing.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 15
Defining Platform Research Infrastructure as a Service (PRIaaS) for Future Scientific Data Infrastructure



Yuri Demchenko, Cees de Laat, Wouter Los, and Leon Gommans

Abstract Modern science increasingly works with large amount of data, which are heterogeneous, are distributed, and require special infrastructure for data collection, storage, processing, and visualization. Science digitalization, likewise industry digitalization, is facilitated by the explosive development of digital technologies and cloud-based infrastructure technologies and services. This paper attempts to understand impact and new requirements to the future Scientific Data Infrastructure imposed by growing science digitalization. The paper presents two lines of analysis: one is a retrospective analysis related to the European Research Infrastructure (RI) development stages and timeline from centralized to distributed and current Federated Interoperable; another line provided analysis of digital technology trends and identified what technologies will impact the future Scientific Data Infrastructure (SDI). Based on this analysis, the paper proposes a vision for the future RI Platform as a Service (PRIaaS) that incorporates recent digital technologies and enables platform and ecosystem model for future science. Notably the proposed PRIaaS adopts TMForum Digital Platform Reference Architecture (DPRA) that will simplify building and federating domain-specific RIs while focusing on the domain-specific data value chain with data protection and policy-based management by design.

15.1 Introduction

Modern science is becoming more and more data driven and works with a large amount of data, which are heterogeneous, are distributed, and require special infrastructure for data collection, storage, processing, and visualization. Science

Y. Demchenko (✉) · C. de Laat · W. Los
Complex Cyber Infrastructure Group, University of Amsterdam, Amsterdam, The Netherlands
e-mail: y.demchenko@uva.nl; C.T.A.M.deLaat@uva.nl; W.Los@uva.nl

L. Gommans
Air France – KLM, Amsterdam, The Netherlands
e-mail: Leon.Gommans@klm.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_15

digitalization, likewise industry digitalization, is facilitated by the explosive development of digital technologies as well as infrastructure technologies and services.

New large-scale scientific problems such as climate, global warming, genome, and fast response to pandemics require using the modern Big Data, Cloud Computing, and Artificial Intelligence technologies. However further research platform advancement requires new approaches to infrastructure services provisioning and management that could facilitate the essential research process and minimize overhead of infrastructure provisioning and management.

Future digital science opens new possibilities of cross-domain/cross-sector integration and consolidation of resources and capacities. It will require new type of infrastructure that would provide extended functionality to collect, store, distribute, process, exchange, and preserve research data to support common knowledge growth and exchange [1–3]: We will refer to this new infrastructure as Future Scientific Data Infrastructure (FutureSDI or FutureRI).

Recent European initiatives and projects such as the European Open Science Cloud (EOSC) [4] and Research Data Alliance (RDA) [5] facilitated implementation of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles [6] that allow for effective data exchange and integration across scientific domains, making scientific data a valuable resource and a growth factor for the whole digital economy and society. To uncover the potential of the future digital and data-driven science, the FutureSDI must provide a platform for effective use of scientific data by allowing creating specialized consistent ecosystems supporting full cycle of the value creation from data collection to model creation and knowledge acquisition and exchange. Shift of the focus from infrastructure operation to value creation will require new FutureSDI design approach, operation, and evolution to respond to changing requirements and evolving technologies. Growing infrastructure complexity will require automation of the infrastructure provisioning and operation, allowing researchers to focus on problem-solving.

This paper attempts to analyze current technology that can advance SDI development and support future digital science. Based on this analysis, the paper proposes a vision for the future RI Platform as a Service (PRIaaS) that incorporates recent digital technologies and enables platform and ecosystem model for future science.

The proposed analysis and PRIaaS architecture are based on the authors' long-time involvement in numerous EU and national projects on RI development, studies, and initiatives, including current ongoing projects GEANT4 [7], FAIRsFAIR [8], and SLICES-DS [9] dealing with different modern Research Infrastructure and e-Infrastructure developments. The paper refers to the previous authors' works on defining the Big Data Architecture Framework (BDAF) [10] and Scientific Data Infrastructure requirements [11] and developing practical aspects of the cloud services network infrastructure [12] that provide a strong foundation for current research.

The paper is organized as follows. Section 15.2 provides a short reference to recent regulations, initiatives, and projects in the European Research Area that drive future SDI and RI development. Section 15.3 provides an overview of the key technology development that may facilitate FutureSDI development. Section 15.4

describes the main features of the future digital science, analyzes the timeline of the European RI development, and proposes a vision for the key technologies that can shape the FutureSDI marked as EOSC-2. Sections 15.5 summarizes the general requirements to FutureSDI and describes the proposed PRIaaS architecture and its operation. Section 15.6 discusses important aspects of the research data management to support FutureSDI requirements and PRIaaS functionality. Section 15.7 presents a conclusion and refers to ongoing and future developments.

15.2 European Research Area

15.2.1 European Research Infrastructures and ESFRI Roadmap

European Research Area (ERA) is an important area of the European policy development and funding to support European science and ensure its competitiveness while facilitating European cooperation and integration. The Research Infrastructures (RI) is one of the pillars of ERA designated to connect research, higher education, and innovation [13].

A European Research Infrastructure (RI) is a facility or (virtual) platform that provides the scientific community with resources and services to conduct top-level research in their respective fields. The research infrastructures can be single-sited or distributed or an e-infrastructure and can be part of a national or international network of facilities, or of interconnected scientific instrument networks.

Important instrument in defining European RI development and evolution is the ESFRI (European Strategy Forum on Research Infrastructures) Roadmap [14]. The new ESFRI Roadmap 2021 defines the important priorities that include consolidating the landscape of European RIs, opening, interconnecting and integrating RIs to develop the full potential of data generated by RIs and increase the innovation potential of ERA/European science in its cooperation with industry [15]. Research Infrastructures constitute a powerful resource for industry, a prerequisite for collaboration between industry and academia.

To facilitate RI and science digitalization, the new ESFRI Roadmap includes a new DIGIT area whose focus is to support research on digital technologies.

e-Infrastructure is another area of the European policy and funding that is designated to support ESFRI and constitute the essential building block for ERA. e-Infrastructures address the needs of European researchers for digital services in terms of networking, computing, and data management. e-Infrastructures provide digital-based services and tools for data- and computing-intensive research in virtual and collaborative environments.

e-Infrastructures are key in the future development of research infrastructures, as activities go increasingly online and produce vast amounts of data. Current European e-Infrastructure capacity includes such Trans-European operational infrastructures

as GÉANT, the high-capacity and high-performance communication network [16], and PRACE, European HPC services for European research [17].

15.2.2 European Open Science Cloud (EOSC)

The European Open Science Cloud (EOSC), started in 2016, is the part of the "European Cloud Initiative—Building a competitive data and knowledge economy in Europe" [18, 19] that is targeted to capitalize on the data revolution. Under this initiative, EOSC federates existing and emerging e-Infrastructures to provide European science, industry, and public authorities with world-class data infrastructure connected to high-performance computers (HPC).

The EOSC goal is to enable the Open Science Commons [20]. At the present time, the EOSC projects created the foundation for research data interoperability and integration for European IRs. The Minimum Viable EOSC (MVE) achieved by the end of 2021 will create a starting point for future EOSC development [21].

MVE defines EOSC Core that is designed to provide a federated data exchange environment for research projects and communities where data comply FAIR principles. EOSC Core includes the following components/functionalities:

- Shared Open Science policy framework.
- Authentication and Authorization Interoperability framework.
- Data Access framework.
- Service Management and Access framework.
- A minimum legal metadata framework.
- An open metrics framework.
- PID framework and service.
- Portal providing web access to the EOSC services and offering Catalog and Marketplace services.

The further EOSC development based on MVE (which we can refer to as EOSC-1) will require designing a new type of infrastructure that can benefit from existing and emerging digital and infrastructure technologies.

15.3 Technology-Driven Science Transformation

15.3.1 Science Digitalization and Industry 4.0

Science digitalization is a demand of time and advised by the OECD report [1]. Science and industry digitalization make easier exchange of technologies, solutions, and application and also adopting recent industry trends such as Industry 4.0 [22] and platform-based ecosystems.

Industry 4.0 will bring tremendous changes to both business models and the way future factories will operate. The key Industry 4.0 elements that both empower new data economy and will be facilitated by the new business and consumer models include Cyber-physical systems; Internet of Things; Internet of services; Smart factories; Mobile technologies and highspeed access networks; Cloud Computing and distributed data processing; Big Data; Artificial Intelligence and Machine Learning; and Automation, Robotics, and Digital Twins.

The digital nature of ongoing economy transformation opens opportunity for faster technologies and solution exchange with science and research. Science can benefit from massive investments into industrial digital and data-driven technologies that can be directly used in digital science, in particular, experimental research automation and following data processing and management. The scientific community should follow the development and be open to wider use of technologies that are advanced by industry; actually all technologies powering Industry 4.0 can be effectively used both in the Future SDI and domain-specific scientific applications.

15.3.2 Transformational Role of Artificial Intelligence

Similar to Industry 4.0, Artificial Intelligence will have a strong transformative effect on future science [23]. Benefits that AI can bring to scientific research and SDI include but not limited to:

- Extending possibilities of research when working with big data.
- Automating data preparation, processing, and analysis.
- Smart infrastructure and tool operation and management.
- AI-driven and Machine Learning-powered scientific discovery and decision support, digital model creation (Digital Twins).
- AI-powered self-learning assistant to a researcher/scientist capable of creating domain-related intelligence; many research questions will be pursued semi-automatically [24].
- Role of data will change: the learned model will replace data; theory becomes data for next-generation AI [24].

It is recognized that an effective work of AI and ML technologies is critically dependent on the quality of data and their availability at all stages of the AI lifecycle [25]. This will impose the specific requirement to the FutureSDI, including general compute and storage, distributed federated ML algorithms, edge computing, and highspeed access network.

Consistent data management including FAIR compliance, quality assurance, data lineage, and privacy protection are general preconditions for successful AI implementation [26].

15.3.3 Promises of 5G Technologies

5G technologies promise to solve not only high-speed mobile communication for smart(phone) applications but also e2e land/terrestrial network communication. 5G architecture defines three main future use cases (or usage scenarios) that can be adopted by the FutureSDI [27]:

- Enhanced Mobile Broadband (eMBB): this also covers IoT, robotics, and sensor network.
- Massive Machine Type Communications (mMTC) to support HPC and large-scale distributed data processing.
- Ultra Reliable and Low Latency Communications (URLLC): industry automation, process control, and real-time applications.

To address these use cases and corresponding requirements, 5G architecture offers e2e network slicing technology that allows proving isolated virtual overlay networks using Network Functions Virtualization (NFV) and cloud native services deployment model and mechanisms. In addition to slices isolation, the 5G architecture is also offering a consistent security model that enables Trusted Execution Environment (TEE) [28] for running secure and trusted services by using the hardware Root of Trust (whose idea is originated from the Trusted Computing Platform architecture [29]).

15.3.4 Adopting Platform and Ecosystems Business Model for Future SDI

The platform economy [30, 31] and digital ecosystems [32] are the two trends shaping ongoing transformation of the modern economy facilitated by digitalization. The wide adoption of the platform business and operational model (as an alternative to the pipeline model) facilitates the creation of the value chain between producers and consumers when using (composable) platform services powered by extended data collection and availability from the platform providers. This allows creating consistent business-oriented digital ecosystems as loose associations of stakeholders and capabilities instantiated on the platform provider facilities. An ecosystem has members that interact in the context of a defined set of services and offerings.

TeleManagement Forum (TMF) defines the Open Digital Architecture (ODA) [33] and the Digital Platform Reference Architecture (DPRA) [34], where the infrastructure provisioning component is defined as the Actualization Platform whose architecture is illustrated in Fig. 15.1.

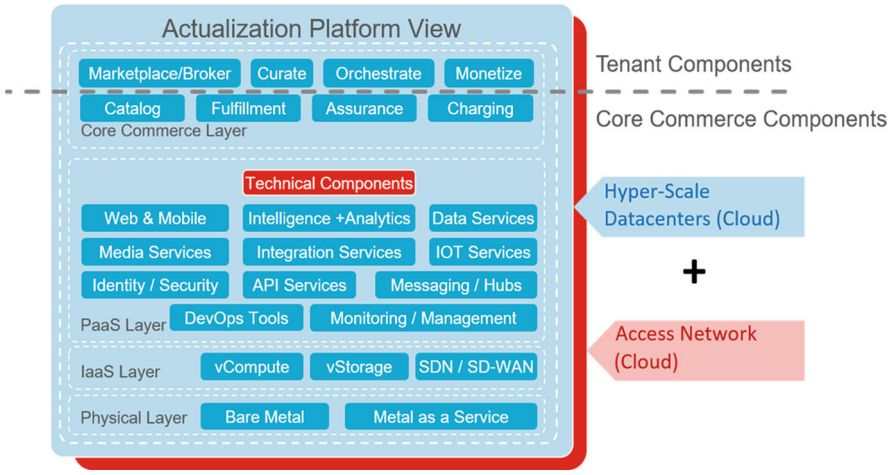


Fig. 15.1 Main functional components of the TMForum Actualization Platform as a core part of DPRA (adopted from [34]). © TM Forum 2020

The Actualization Platform includes the following essential (group of) components:

- Common infrastructure and platform services.
- Data and digital content (media) services.
- Integration and Lifecycle Management.
- Integration, orchestration, and DevOps.
- Security and Identity Management.
- Core commerce services including Catalog, Accounting and Billing, Fulfillment Platform components, and customer/tenant facing services.

The Fulfillment Platform defined in DPRA "allows for user/service configuration and activation data to be sent for each individual component service, and also for fully composed product offers (of the customizable templates or design patterns). It allows a product creator to configure (fulfill) a service that is being composed into an e2e offer—this could involve adding an end-user (authorization credentials, establishing an account), or any other actions required for configuration management" [34].

ODA and DPRA are adopted by many telecom providers, and we can benefit from adopting it for FutureSDI that could serve to create instant virtualized RI and ecosystems for specific user communities.

15.3.5 Other Infrastructure Technologies and Trends

The following are recent technologies that can be adopted to build the Future SDI:

- Cloud-based federated hyperconverged infrastructure allowing for provisioning on-demand secure private infrastructure [35]

- IDSA architecture and IDS Trusted Connector enabled data exchange infrastructure [36, 37]
- Infrastructure automation technologies and tools (virtualization, microservices, composability, containerization, code libraries, API).
- DevOps and CI/CD that trends to become integrated into the change management process to ensure the continuous evolution of the target system [38]
- Data-centric models DataOps/MLOps (whose examples are services offered by Azure cloud platform) [39, 40]
- Semantic Data Lakes as integrated data storage and data analytics platform (whose example is Azure Data Lake gen2 that offers storage for heterogeneous data and provides integrated data analytics) [41, 42]
- Permissioned blockchain technologies that allow for traceable and policy enforceable data sharing and lineage [43]
- Infrastructure-related security technologies that propose solutions for trust bootstrapping and creating secure trusted virtual execution environment for data processing (such as Confidential Computing or secure enclave computing [44, 45]).

15.4 Defining Future Scientific Data Infrastructure

15.4.1 Paradigm Change in Modern Data-Driven/Digital Science

Ongoing Science digitalization is powered by the rapid development of Cloud Computing, Big Data, Artificial Intelligence, and DevOps-based infrastructure automation technologies.

The FutureSDI should consolidate existing and future RIs focusing on specific scientific domains and minimize costs and efforts of creating specialized RI for different scientific communities. Achieving MVE/EOSC-1 will create a platform for FAIR data interoperability and sharing, a key step in the future digital transformation of science.

Here we summarize the main characteristics of the (future) digital science powered by recent advancement in data-driven technologies and AI (also refer to our previous analysis [10]):

- Availability of Pan-European Research Infrastructure Platform as a Service (later defined as PRIaaS) that uses cloud-native technologies (S/P/IaaS) for on-demand provisioning of the fully operational infrastructure for end-to-end scientific research (both experiments and data processing) by using composable infrastructure and application design templates, supported by DevOps tools.
- Automation of scientific experiments and all data handling processes, including data collection, storing, classification, pre-processing and curation, and provenance.

- Adopting and leveraging DevOps and DataOps/MLOps technologies found rapid adoption in the industry and supported with a variety of tools available with cloud-based infrastructure platforms such as AWS, Azure, Google Cloud Platform, and from multiple vendors.
- Digitizing existing artifacts and creating their digital twins, AI-assisted documenting and cataloging, building subject/domain knowledge base using self-learning algorithms.
- The full adoption of the FAIR data principles, both prospective and retrospective, to ensure reusability of available data/datasets in the cross-domain and secondary research.
- Adopting STREAM data properties and corresponding infrastructure to enable trusted multipurpose data sharing and exchange, including data trading as economic goods and enabling different economic models for data sharing.
- Availability of new algorithms for distributed secure data processing such as federated machine learning, or blockchain-enabled policy-aware distributed data processing.
- Global data availability and access over the network for cooperative group of researchers, including wide public access to scientific data, however subject for the data sharing and access policies, in particular GDPR.
- Advanced security, access control, and identity management technologies that ensure the secure operation of the complex research infrastructures and scientific instruments and allow creating a trusted secure environment for cooperating groups and individual researchers.

The future SDI should support the whole data lifecycle and explore the benefit of data aggregation and provenance at a large scale and during a long/unlimited period of time.

Data security is not limited by a secure and trusted storage but also requires a secure and trusted data processing environment that would allow data processing using proprietary algorithms. Demand for RI trustworthiness and security is increasing to address both personal data protection and the trustworthiness of the research process itself. Data infrastructure must ensure data security (integrity, confidentiality, availability, and accountability), trustworthiness, and, at the same time, data sovereignty that include both data ownership protection and control of data sharing and processing by data owners. There should be a possibility to enforce data/dataset policy (sharing, processing, derivative/secondary data) in the distributed data storage, sharing, and processing environment.

15.4.2 Timeline of the European RI Development/Evolution

In our research on the technologies for FutureSDI, we analyzed the development and evolution of the European RIs. Figure 15.2 below illustrates the timeline of the European RI evolution (based on the authors' expertise and wide community

SLICES

RI Type (evolution stage)	Centralised 1994-1996	Interconnected 2004-2006	Distributed 2011-2012	Federated 2016-2018	EOSC-1 2020-2022	EOSC-2 (future)
Definition	Institutions based, centralised facility	Multi-institutions, interconnected	Large distributed facilities, domain or experiment oriented	Federated RIs supporting inter-domain cooperation and data exchange	Interoperable (European) RI, FAIR RI	Virtualised Pan-European RI platform as a Service and ecosystem (PRIaaS)
Network & Compute	Mainframe, variety of protocols, Advent of Internet, web, email	Interconnected data centers and experimental facilities, Internet TCP/IP as common protocol, remote access	Distributed interconnected computing facilities, SOA and webservices, Grid as cooperative and distributed computing	Cloud adoption, infrastructure services on-demand Federated facilities and network access, Federated access and Identity management, 3G->4G	Distributed scalable computing, cloud based Big Data technologies, high performance networks, 5G technologies, wireless access, IoT sensor networks	Composable virtualized RI provisioning on demand, common federated computing and networking platform/environment, Cloud, DevOps and AI enabled, Digital Twins
Data	Proprietary formats, system or experiment specific	Standard format for data exchange, proprietary metadata	Domain/RI based data/metadata interoperability, custom data models, distributed storage, directories	Interoperable data, domain based metadata	FAIR data, Data Factories, Metadata registries, Interoperable/common Data Management model	Fully adopted FAIR principles, Semantically enabled scientific data lakes, secure/trusted data exchange, full data value chain
Infrastructure Management Technologies	Local management	Local management, management information exchange	Common Management Model, Distributed management, 3G Roaming	OSS/BSS, Automated deployment, adaptation, monitoring	Integrated Operation and Automation, Automated identity provisioning	Fully automated RI and services provisioning, management and operation, optimisation

Fig. 15.2 Timeline of RI evolution and SLICES positioning

Table 15.1 Details of the technologies used in current EOSC-1 and future EOSC-2

RI type	EOSC-1 (2016–2021)	EOSC-2 (future 2022–2025+)
Definition	Interoperable federated RI, FAIR RI	Virtualized Pan-European RI platform as a Service (PRIaaS)
Network and compute	• Distributed scalable computing. • Cloud-based Big Data technologies. • High performance networks. • 5G technology readiness • IoT sensor networks. • Portal. • Services Catalog and API repository. • Industry standards and IDSA adoption.	• Composable virtualized RI provisioning on demand (including for services integration). • Common federated computing and networking platform/environment, enabling virtual RIs. • Cloud based and cloud enabled. • 5G adoption: wireless access network, e2e network slicing • DevOps- and AI-enabled services. • Digital twins. • Interoperability and integration with industry infrastructure (e.g., IDSA+, industrial internet).
Data infrastructure	• FAIR data management and exchange. • Metadata registries. • PID and data factories. • Interoperable/common data management model.	• Fully adopted FAIR principles, extended to ontologies. • Semantically enabled scientific data lakes, common vocabulariesSecure/trusted data exchange (data markets). • Full data value chain supported (cross-domain).
Security	• Federated identity management, federated access control. • Automated identity provisioning.	• Federated identity management, federated access control. • Automated identity provisioning. • Zero trust security, trust bootstrapping. • Homomorphic encryption and data processing. • Quantum-ready encryption, quantum-enabled key management.
Infrastructure management technology	• Integrated operation and automation. • Automated identity provisioning.	• Fully automated RI and services provisioning, management, and operation. • AI-enabled optimization of infrastructure and operation. • DevOps and reusable design patterns.

Evolution from EOSC-1 to EOSC-2 means that new advanced EOSC-2 services are built on the available and operational EOSC-1 services

discussions aligned with technology evolution and trends) that covers past stages: Centralized, Interconnected, Distributed, and Federated, where the current stage is labeled as EOSC-I (actually implementing EOSC Core) and foreseeing future stage labeled as EOSC-II. Table 15.1 provides extended details about technologies that are suggested to drive the transition from EOSC-I to EOSC-II.

Past stages (before EOSC) delivered Federated Research Infrastructures supporting inter-organizational and interdomain cooperation and data sharing using well-defined metadata ensuring data interoperability, however in many cases limited to a science domain. Examples of such RIs are EGI, EUDAT, GEANT,

PRACE, and other landmark RIs as reviewed in the ESFRI Roadmap 2018 [46]. The European Open Science Cloud (EOSC) provides a basis for European RI integration and interoperability based on adoption of the FAIR principles both for data and for RIs themselves. H2020 EOSC-hub project established and operates EOSC Portal offering services Catalog and Marketplace that enables services and data findability, interoperability, and reusability based on published APIs [47].

Future progression and adoption of modern technologies such as Cloud and Edge Computing, Big Data, AI, IoT, and Digital Twins will enable fully virtualized Pan-European RI platform as a Services (PRIaaS) that will allow virtualized RI provisioning on demand for specific scientific domain and community; advanced data management and processing technologies will allow full FAIR principle implementation and trusted data exchange, supporting whole data lifecycle and value chain with the necessary infrastructure services. Adoption of the 5G technologies is expected to start a preparatory stage at the EOSC-I stage in some individual projects and testbeds and will become the main enabling technology for virtualizing/slicing network and RI in the future, combining with the Virtual Private Cloud (VPC) [48] technologies supported by modern cloud platforms.

The envisioned PRIaaS definition leverages the TMForum DPRA concepts and principles that define the provider actualization platform as a way to enable provisioning customer-tailored services platform/ecosystem on demand.

Recently started the SLICES-DS project [9] intends to bridge the current EOSC-I stage and future EOSC-II stage by advancing infrastructure technologies to fully virtualized customized domain-specific RI provisioning on-demand. Many modern advanced and emerging technologies need to be tested, adopted, and prototyped to make them easily usable by different RIs and embedded into the PRIaaS platform (see Sect. 15.5 for PRIaaS architecture).

15.4.3 General Requirements to Future Data-Driven Research Infrastructures

From the overview above, we can specify the following general infrastructure requirements to the future Scientific Data Infrastructure:

- Cloud-based platform for provisioning (on-demand) instant RIs, fully configured and functional including Virtual Organization for user management
- Support of virtual scientist communities, addressing dynamic user group creation and management, federated identity management—to enable cooperation and support scientific workflows
- Support FAIR data principles by providing necessary metadata services and data sharing facilities
- Secure trusted data infrastructure, ensuring data sovereignty and trustworthiness, supporting STREAM data properties for effective and value-added data exchange [49]

- Support long-running experiments and large data volumes generated at high speed
- Trusted environment for data storage and processing
- Support for data integrity, confidentiality, accountability, provenance, sovereignty
- Mechanisms for policy binding to data to protect privacy, confidentiality, and IPR that ensure the policy is attached to data during the whole data lifecycle; mechanisms for policy provisioning and roaming as part of the provisioned infrastructure to ensure policy enforcement by design in a diverse heterogeneous environment.

15.5 Proposed PRIaaS Architecture Model

We propose the PRIaaS Architecture for FutureSDI as illustrated in Fig. 15.3. This model contains the three generalized layers:

Virtualized Resources (VR): Virtualized general compute, storage and network resources that are composed to create infrastructure components and are used by other services and applications.

Actualization Platform: This is the main component and layer that enables provisioning, monitoring, and operating fully functional instant Virtual RIs for specific scientific domains, projects, or communities.

Virtual (Private) RI (VirtRI): Virtual RI provisioned on demand that contains a full set of services, resources, and policies needed to serve the target scientific community and create full value change of data handling. VirtRI is operated by the specific community and uses services provided by the Actualization platform, including the possibility of cross-platform data sharing.

Users and external resources include researchers, developers and operators, and external datasets.

Federation Access Infrastructure and Tenants Management (FAI&TM) layer serves as interface layer enabling communication between distributed Actualization Platform resources and services and generally distributed and multiorganizational VirtRI. FAI&TM is also the place where VirtRI and Actualization Platform policy are enforced and managed.

15.5.1 Actualization Platform Components

The PRIaaS Actualization Platform includes the following groups of services required to develop, deploy, manage, and operate the Virtual RI during its whole lifecycle, including resources and users that can be grouped into Virtual Organizations.

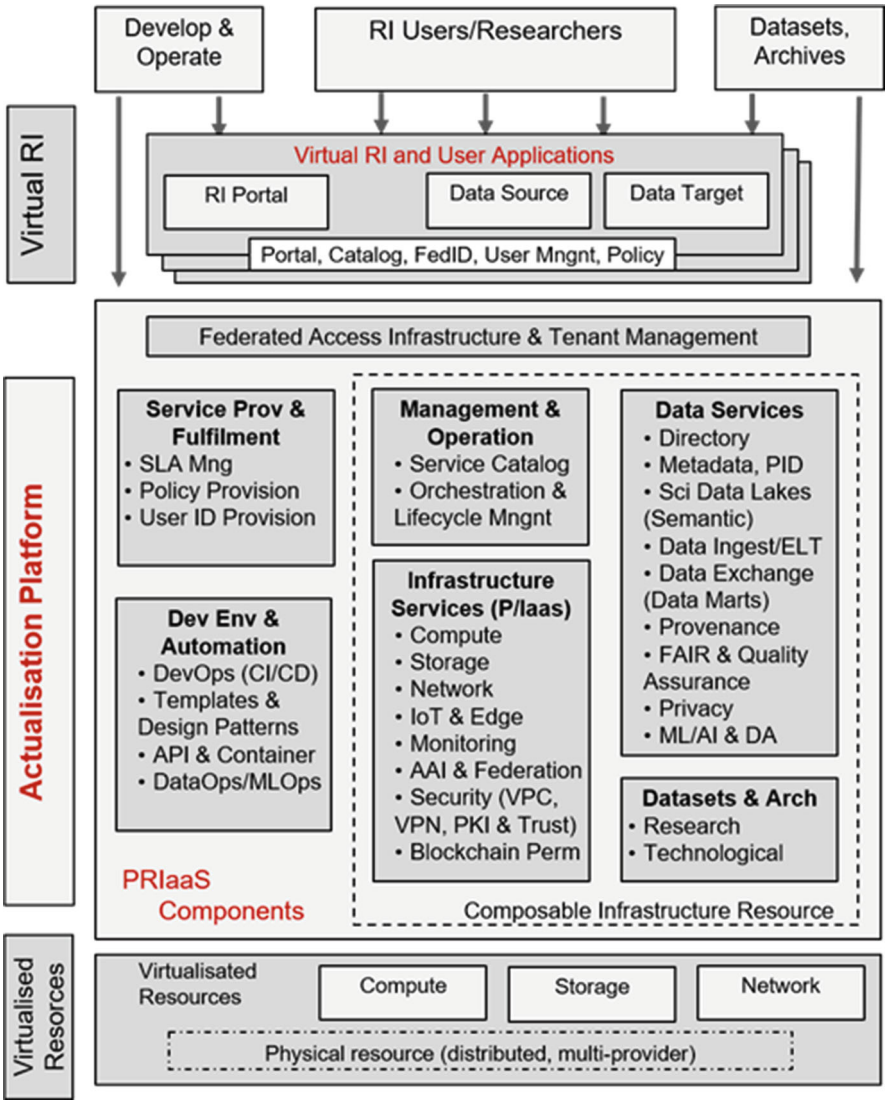


Fig. 15.3 The proposed PRIaaS architecture

- Core Infrastructure Services (IaaS & PaaS) including compute, storage, network, IoT&Edge, blockchain, Access Control and Federated Identity management, infrastructure security.
- Data Services including directory, metadata/PID, lineage/provenance, FAIR &QA, semantic data lakes, data analytics, and AI tools.
- Management and Operation including Service Catalog and Lifecycle Management, orchestration, and management.

- Service provisioning and fulfillment including user provisioning, SLA management and policy provisioning.
- Development Environment and Tools that support DevOps process related to platform and VirtRI development, provisioning, and operation; this group also maintains the repository of API, containers, and design templates that can facilitate VirtRI design and provisioning.

VirtRI provisioning process is based on well-known and commonly used DevOps tools and is supported by the Management and Operation functions. As the PRIaaS platform will progress, the repository of the design patterns, templates, and containerized applications and functions will grow. A starting point for such a repository can be the EOSC Catalog [47] that already contains information about API for applications and services offered by existing RIs and service providers.

The policy provisioning, management, and enforcement are important functions of the Actualization Platform that can be attributed to the Fulfillment function. The policy that is defined by the target community is provisioned as a part of VirtRI provisioning. Policy management and enforcement infrastructure should support policy roaming and combination for the multi-domain distributed resources and tenants.

15.6 Research Data Management in the Future SDI

15.6.1 European-Wide and International Initiatives and Projects

The importance of data and research information sharing has been central in a number of European-wide initiatives and projects, such as Open Access, Open Data, Open Science, and Open Commons. The Research Data Alliance (RDA) that was created in 2012 jointly by the National Science Foundation of USA (NSF) and European Commission became a key community coordination body to exchange and develop best practices in research data management. One of the important RDA developments became Persistent Identifiers (PID for data objects to enable data interoperability and findability) [50].

To facilitate research data sharing and implementation of the FAIR principles, European Commission started Open Research Data (ORD) Pilot [51], and currently all EU-funded projects are required to develop and implement the Data Management Plan (DMP) at the initial stage of the project. Data produced in the project must be stored in the open available but secure repositories (operated by the project or using national or European data archive services). Metadata must be published and quality of data ensured, in particular, compliance with the FAIR principles.

15.6.2 From FAIR Data Principles to STREAM Data Properties

FAIR data principles are important for creating trusted research-friendly environment for data sharing. FAIR data is a key element/layer of the EOSC core. However, the data exchange infrastructure requires additional data properties that would allow trusted and economical data exchange, also supporting data value chain creation.

Data exchange and data trading/market have been long-time interest area by/from the industry where data represent also companies' intellectual property, and companies want to remain in control of their data which is defined as data sovereignty.

Data Sovereignty is a key principle of the industrial data exchange as defined by the International Data Spaces Association (IDSA) Reference Architecture Model (RAM) [36].

Data involved in industrial processes and business relations are becoming a part of the economic relations and added value creation process. However, data as economic goods are in many aspects different from the traditional economic goods and commodities. We refer to our research on data properties as economic goods as part of the RDA Interest Group on Data Economics (IG-DE) [52].

Emerging data-driven economy and modern Big Data technologies facilitate interest in making data a new economic value (data commoditization) and consequently the identification of new properties of data as economic goods. The STREAM data properties for industrial and business data have been proposed by the authors in [49]. To become an economic goods and bring business value to data producers and data consumers, data must be [**S**] sovereign, [**T**] trusted, [**R**] reusable, [**E**] exchangeable, [**A**] actionable, and [**M**] measurable.

Other data properties important to enabling data commoditization and allowing data trading and exchange for goods include quality, value, auditability/trackability, branding, authenticity, as well as original FAI(R) properties: findability, accessibility, interoperability, and reusability. Special features that must be managed in all data transfers and transformations are data ownership, IPR, and privacy. The data property originated from its digital form of existence defined as not-Rivalry, on one hand, makes data exchange (copying, distribution) easy, but on the other hand, it creates a problem when protecting proprietary, private, or sensitive data or IPR.

15.7 Future Research and Development

In this paper, we presented analysis of current trends in digital technologies that can be used to build Future Scientific Data Infrastructure and in particular can be used to progress the current EOSC infrastructure, also proposing a common platform for future European RI integration. Further research will require a closer analysis of the typical use cases in ESFRI and EOSC projects. The presented research and proposed PRIaaS are based on the authors long-time experience in infrastructure research and

developing/implementing practical solutions in a number of national EU-funded projects such as EGEE, GEANT, and GEYSERS, as well as standardization activity in such bodies as IETF, OGF, NIST, and CEN.

The proposed PRIaaS architecture and DPRA-inspired operational model require a variety of technologies to work together realizing data-centric data exchange and transformation to enable data-based applications and services and added value data service creation. New functionality and technology combinations will require re-thinking existing concepts and models, extending usage scenarios.

Further development of the proposed PRIaaS and its components will be done in the ongoing project SLICES-DS. This work also intends to contribute to the EOSC Architecture Working Group.

Acknowledgments This work is supported by EU-funded projects SLICES-DS (Grant Agreement No. 951850), FAIRsFAIR (Grant Agreement No. 831558), and GN4-3 (Grant Agreement No. 856726). The authors value wide discussions in the RDA and EOSC forums on the different aspects of the existing research infrastructures, data management, and ongoing research and developments.

References

1. The digitalisation of science, technology and innovation: Key developments and policies. Paris: OECD Publishing. (2020). [online]. https://doi.org/10.1787/b9e4a2c0-en.
2. *Measuring the digital transformation: A roadmap for the future*. Paris: OECD Publishing [online]. https://doi.org/10.1787/9789264311992-en.
3. EC COM. (2020). *A European strategy for data*. 66 final, Brussels, 19.2.2020 [online]. https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf
4. European Open Science Cloud (EOSC) [online] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
5. Research Data Alliance [online] https://rd-alliance.org/
6. *Turning FAIR into reality*. Final Report and Action Plan from the European Commission Expert Group on FAIR Data, 2018 [online] https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf
7. GÉANT Project GN4–3, Accelerating research, driving innovation and enriching education [online]. https://www.geant.org/Projects/GEANT_Project_GN4-3
8. FAIRsFAIR Project: Fostering FAIR data practices in Europe [online]. https://www.fairsfair.eu/
9. SLICES-DS Project [online]. http://slices-ri.eu/
10. Demchenko, Y., De Laat, C & Membrey, P. (2014). *Defining architecture components of the Big Data Ecosystem*. In International Conference on Collaboration Technologies and Systems (CTS 2014), May 19–23, 2014, Minneapolis, USA.
11. Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). *Addressing big data issues in scientific data infrastructure*. In Proc. 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), May 20–24, 2013, San Diego, California, USA. ISBN: 978-1-4673-6402-7.
12. Demchenko, Y, Van Der Ham, J, Ngo, C, Matselyukh, T, & Filiposka, S. (2013). Open cloud exchange (OCX): Architecture and functional components. IN Proc. 2013 IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom2013), 2–5 Dec 2013, Bristol, UK.

13. European Research Area (ERA) [online]. https://ec.europa.eu/info/research-and-innovation/strategy/era_en
14. Landscape of Research Infrastructures and evolution of the ESFRI Roadmap. ESFRI Roadmap 2021 InfoDay 25 Sept 2019 [online]. https://www.esfri.eu/esfri-events/roadmap-2021-infoday?qt-event=5#qt-event
15. Making Science Happen: A new ambition for Research Infrastructures in the European Research Area. ESFRI Whitepaper, March 2020 [online]. https://www.esfri.eu/esfri-white-paper
16. GEANT. https://www.geant.org/
17. PRACE – Partnership for Advanced Comp in Europe. https://prace-ri.eu/
18. European Cloud Initiative - Building a competitive data and knowledge economy in Europe [online]. https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe
19. Building the European Open Science Cloud, By Drago, Federico; Ferguson, Nicholas, 19 March 2020 https://zenodo.org/record/3716192#.X1aYzXkzZPY
20. Ferrari, T., Scardaci, D., & Andreozzi, S.. The Open Science Commons for the European Research Area, Part of the ISSI Scientific Report Series book series (ISSI, Vol. 15) [online]. https://link.springer.com/chapter/10.1007/978-3-319-65633-5_3
21. Solutions for a Sustainable EOSC. A tinman report from the EOSC Sustainability Working Group, Draft 2 December 2019. https://www.eosc-nordic.eu/content/uploads/2020/03/Tinman_draft_19_compressed.pdf
22. Industry 4.0: the fourth industrial revolution – guide to Industry 4.0 [online]. https://www.i-scoop.eu/industry-4-0/
23. AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science, July–September 2019 [online]. https://www.anl.gov/ai-for-science-report
24. AI for Science, by Barbara Helland, AI for Science Town Hall, Oct 2019 [online]. https://science.osti.gov/-/media/ber/berac/pdf/201910/Helland_BERAC_Oct2019.pdf
25. AI Development life cycle: explained [online]. https://www.devteam.space/blog/ai-development-life-cycle-explained/
26. Data quality in the era of Artificial Intelligence, blog by George Krasadakis [online]. https://medium.com/innovation-machine/data-quality-in-the-era-of-a-i-d8e398a91bef
27. 5G and The Cloud, 5G Americas White Paper, December 2019 [online]. https://www.5gamericas.org/5g-and-the-cloud/
28. The Evolution of Security in 5G- 5G Americas White Paper, 5G Americas, July 2019 [online]. https://www.5gamericas.org/wp-content/uploads/2019/08/5G-Security-White-Paper-07-26-19-FINAL.pdf
29. INSPIRE-5Gplus Deliverable D2.1: 5G Security: Current Status and Future Trends, 2020 INSPIRE-5Gplus Consortium Parties [online]. https://www.inspire-5gplus.eu/wp-content/uploads/2020/05/i5-d2.1_5g-security-current-status-and-future-trends_v1.0.pdf?x51934
30. Parker, G. G. (2016). *Platform revolution: How networked markets are transforming the economy and how to make them work for you*. W. W. Norton & Company.
31. Key Enablers for a Hybrid Infrastructure Platform, TMForum, 15–18 May 2017, Nice, France [online]. https://dtw.tmforum.org/wp-content/uploads/2017/05/12.-Stephen-Fratini-Milind-Bhagwat-Takayuki-Nakamura.pdf
32. In the Ecosystem Economy, What's Your Strategy? By Michael G. Jacobides, Harvard Business Report, Issue Sept–Oct 2019 [online]. https://hbr.org/2019/09/in-the-ecosystem-economy-whats-your-strategy
33. IG1167 TM Forum Exploratory Report ODA Functional Architecture, 31 Jan 2020 [online]. https://www.tmforum.org/resources/exploratory-report/ig1167-oda-functional-architecture-v5-0/

34. IG1157 Digital Platform Reference Architecture Concepts and Principles v5.0.1, 21 July 2020 [online]. https://www.tmforum.org/resources/reference/ig1157-digital-platform-reference-architecture-concepts-and-principles-v5-0-0/
35. 11 main benefits of hyper-converged infrastructure, August 2020 [online]. https://searchconvergedinfrastructure.techtarget.com/tip/11-main-benefits-of-hyper-converged-infrastructure
36. IDSA Reference Architecture Model (RAM3.0), Version 3.0, April 2019 [online]. https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-Reference-Architecture-Model-3.0.pdf
37. Trusted Connector, Industrial Data Space, Draft DIN Spec 27070 [online]. https://industrial-data-space.github.io/trusted-connector-documentation/
38. DevOps & Change Management In The Enterprise World [online]. https://clearbridgemobile.com/devops-change-management-in-the-enterprise-world/
39. MLOps: Methods and Tools of DevOps for Machine Learning, 23 July 2020 [online]. https://www.altexsoft.com/blog/mlops-methods-tools/
40. Design a machine learning operations (MLOps) framework to upscale an Azure Machine Learning lifecycle, Microsoft Azure [online]. https://docs.microsoft.com/bs-latn-ba/azure/architecture/example-scenario/mlops/mlops-technical-paper
41. Ten Reasons to Dive into the Smart (Semantic) Data Lake Cambridge Semantics, 2017 [online]. https://blog.cambridgesemantics.com/ten-reasons-to-dive-into-the-smart-semantic-data-lake
42. Building your Data Lake on Azure Data Lake Storage gen2, Blog by Nicholas Hurt, March 2020 [online]. https://medium.com/@Nicholas_Hurt/building-your-data-lake-on-adls-gen2-3f196fc6b430
43. Kumar, H. A., Julita, V., & Ralph, D. (2020). A blockchain platform for user data sharing ensuring user control and incentives. *Frontiers in Blockchain, 3*, 1022. [online]. https://www.frontiersin.org/article/10.3389/fbloc.2020.497985
44. Confidential Computing Consortium [online]. https://confidentialcomputing.io/
45. A Technical Analysis of Confidential Computing v1.1, Whitepaper, Confidential Computing Consortium, January 2021 [online]. https://confidentialcomputing.io/wp-content/uploads/sites/85/2021/03/CCC-Tech-Analysis-Confidential-Computing-V1.pdf
46. ESFRI Roadmap 2018 [online]. http://roadmap2018.esfri.eu/media/1066/esfri-roadmap-2018.pdf
47. EOSC Catalog [online]. https://catalogue.eosc-portal.eu/
48. Virtual private Cloud [online]. https://en.wikipedia.org/wiki/Virtual_private_cloud
49. Demchenko, Y., Los, W., & de Laat, C. (2018). Data as economic goods: Definitions, properties, challenges, enabling Technologies for Future Data Markets. *ITU Journal, 1*, 2. ICT Discoveries, Special Issue "Data for Goods".
50. Persistent Identifiers, Groups of European Data Experts, 27 Nov 2017, RDA [online]. https://www.rd-alliance.org/system/files/PID-report_v6.1_2017-12-13_final.pdf
51. Data management, Extension of the Open Research Data Pilot in Horizon 2020, Horizon2020 Manual [online]. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
52. RDA Interest Group on Data Economics (IG-DE) [online]. https://www.rd-alliance.org/groups/data-economics-ig

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part III
Use Cases and Data Ecosystems

Chapter 16
Silicon Economy: Logistics as the Natural Data Ecosystem



Michael ten Hompel and Michael Schmidt

Abstract The "Silicon Economy" is synonymous with a coming digital infrastructure (digital ecosystem) based on the automated negotiation, disposition, and control of flows of goods, enabling new, digital business models (not only) for logistics. This infrastructure requires and enables the trading of data without losing sovereignty over the data. It is the digital infrastructure and environment for the highly distributed AI algorithms along value networks. In contrast to oligopolistic developments in the B2C sector (amazon.com, AirBnB, Alibaba, Uber, etc.), the Silicon Economy is a federated and decentralized platform ecosystem, the basic components of which are made available to the general public as open source for free use.

The Silicon Economy ecosystem is becoming an enabler of supply chain ecosystems in which goods, autonomously controlled by Artificial Intelligence (AI), undergo orchestrated processes according to the situation.

This article focuses on the origins and potentials but also on the technological foundations and challenges of the transformation toward a Silicon Economy.

16.1 The Digitization of Everything and Artificial Intelligence in Everything Will Change Everything for Everyone

By the end of the 1920s, AI algorithms will determine, regulate, and control nearly everything in the logistics sector—and not only there. Platforms will hoard data and generate knowledge. Swarms of autonomous robots will explore their surroundings, negotiate with each other, and organize themselves.

A new "Silicon Economy" is emerging. It will outclass the business models of Silicon Valley and turn the world upside down. And there is no alternative to the introduction of AI—human intuition and hierarchical order have failed in the attempt

M. ten Hompel (✉) · M. Schmidt
Fraunhofer-Institute for Material Flow and Logistics IML, Dortmund, Germany
e-mail: michael.ten.hompel@iml.fraunhofer.de; michael.b.schmidt@iml.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_16

to master the complexity of existing networks and processes. AI algorithms and their machine learning will define the game. Logistics and supply chain management are the crucial domains where the initial stage of this new B2B competition will be decided.

The Coincidence in Time Is Crucial

The introduction and universal application of AI characterizes the era we live in. Autonomously interacting entities increasingly determine the course of development. Driven by the hardware development of digital semiconductors such as memory, low-power sensors, and processors, the automation of entire processes and supply chains on the basis of autonomous entities in software and hardware is now becoming focus of attention. The decisive factor here is the coincidence in time of a wide range of technical developments:

- Industry 4.0 and the Internet of Things including AI in devices (Edge AI)
- Real-time networking (5G, Wi-Fi 6)
- AI-based platforms (AI Platform as a Service)
- Blockchain (distributed ledger) and automated negotiation (smart contracting)
- Swarms of autonomous robots (LoadRunner®)
- Virtualization and simulation (simulation-based AI)
- Immersive technology such as augmented reality (AR) and virtual reality (VR), which connect humans with AI
- Cognitive computing
- Quantum computing

The common element in each case is the universal application of AI—albeit in a wide variety of forms: from the simplest, rule-based systems in the trackers of our containers, the support vector machines in "intelligent" sensors, and simulation-based reinforcement learning of swarms of autonomous vehicles to deep learning algorithms in supply chain management. Obviously, it seems to be logistics where all these technologies are now breaking through simultaneously. Due to its comparatively simple processes performed millions of times and yet its enormous systemic complexity as a whole, logistics is the sector that is virtually a prime example of AI application. For example, the automatic identification and measurement of individual packages via camera and AI is already a market worth billions. However, the real market potential will be leveraged when the process chains on the AI platforms of future supply chain management close and AI algorithms fully permeate logistics networks both vertically (from the sensor to the cloud) and horizontally (along logistics processes).

There is no single development that is currently leading to a disruptive change or by which the entire era is named. It is the temporal coincidence that concentrates a multitude of exponential developments on just one point. However, the outcome is indeed singular. And then, in turn, a "connected and autonomous supply chain ecosystem" [1] emerges: the Silicon Economy.

Social Change

It's not just about engineering and technology but also about an essential change in our society. Humans will no longer be the "decisive authority," but will hand over the reins of action to machines and their algorithms. The change is universal and will not take the form of a machine man, as Fritz Lang once depicted in his film Metropolis. On the contrary, imitating humans in robot form would essentially be a waste of resources and the corruption of a technology that can do many things, but is by no means human. It will be essential to relocate humans and their position in relation to AI.

In the first three industrial revolutions, mechanical work was transferred to machines and robots. For such an industrial application, it was pointless to think about whether computers could develop creativity or intuition. Today, their abilities are far beyond human capabilities in certain areas. For example, even painting pictures and composing pieces of music can be learned comparatively easily by a computer via AI [2]. Even experts are no longer able to distinguish whether some works were based on human or machine creativity.

In the Silicon Economy, intellectual work is increasingly being transferred to machines. In logistics, for example, this will manifest itself in the planning, control, and scheduling of processes and in new business models. In this context, Anders Indset et al. [3] speak of an emerging "knowledge society" in which we as humans only react primarily to predefined knowledge from our search engines and databases that has been algorithmically processed by AI and which must therefore be overcome. This would make humans increasingly obsolete—at least in terms of repetitive skills or the representation of knowledge.

After the automation of assembly processes in the third industrial revolution and the associated loss of jobs, cashiers at the supermarket checkout **could** lose their jobs or the banker who **migh**t only reproduce what the automated check via AI revealed. Today, operations are supported by AI. A surgeon who operates on a cataract might be replaced at some point in future, or the teaching profession might be enriched by artificial avatars—at least as far as pure knowledge transfer is concerned. The consequence of this development is the demand for a change from a purely reflective knowledge society to an "understanding society" in which a return to humanistic values and the abilities for philosophical, artistic, and scientific discourse are considered essential for human beings.

It is a technical question how we leverage the potentials of neural networks in our computers; it is another question how AI changes the neural networks in our brains. Elon Musk, founder of Tesla and SpaceX, faced this question and came to the conclusion that we have to combine the human brain with AI in order to avoid ending up as its own pet. In his characteristic consistency, he founded the company Neuralink in 2016 and now intends to connect the human brain with a computer to enable paralyzed people to use computers. However, this should only be the first step in ensuring the intellectual participation of humans in future and in connection with the machine world.

However, AI will develop in relation to humans, and one thing seems indispensable: A profound debate is needed about what it means to be human today and tomorrow.

Sharing Economy

The universal challenge of the ubiquitous introduction of AI raises the question of how to ensure the participation and sharing of many people and companies.

On the one hand, the aim is to prevent AI from becoming independent, as feared by Elon Musk et al. (see above).

On the other hand, however, the dimension of this development exceeds what can be achieved by a single organization—no matter how large it may be. At the same time, "sharing" is the new generation's leitmotif of developers who have grown up with the principle of swapping and sharing on the Internet and have internalized a different logic of giving and taking: "Using instead of owning" is their motto. The principle has spread to large areas of the economy and has become the basis for new value creation models.

This leitmotif is followed by the open-source software movement, i.e., the freely accessible provision of source code, which offers people and businesses the opportunity to use, adapt, and distribute this source code. The publication of construction plans as open hardware or the provision and use of data as open data are also expressions of the sharing mindset, as are open innovation processes with internal and external forces (open innovation). Common to all these trends is the underlying confidence that business potentials generated by intact and open ecosystems can be better leveraged together—for example, through greater innovative strength, through better stability and IT security, or through the avoidance of licensing costs, etc.

Open-source software is now an integral part of the digital economy in Germany and a constituent part of almost all innovation processes—across countries and with the participation of numerous organizations. This does not only apply to the Internet economy but also to industrial production where 50% of the code base is now built on open-source software [4]. It is impossible to imagine today's world without it. The digital transformation and therefore also the Silicon Economy will not succeed without using open source.

16.2 Potential of the Silicon Economy for Logistics and Supply Chain Management

The importance of logistics has increased strongly in recent decades in parallel with the growth in world trade. Logistics forms the basis of global trade. It connects places and companies in global networks—from the physical flow of materials and goods to the exchange of data in the flow of information and the flow of finance in logistics management. In this respect, logistics is one of the most important factors influencing free world trade.

Before presenting the potential of the Silicon Economy in this domain, the terms Logistics and Supply Chain Management should be defined.

Logistics and Supply Chain/Logistics Management

Logistics describes the reasonable movement of things, in places, through time, and in relations. It is a fundamental principle that permeates everything physical and its movement. At the same time, it is an expression of man's striving to set things in motion. Based on Delfmann et al., we will define logistics as an applied science, as an industry, as well as an operational function. Logistics analyzes and designs economic systems as flows of objects (above all, but not exclusively: goods and people) in networks, supplying recommendations for action on the design, implementation, and operation of these networks.

Across Europe, the logistics market amounts to around 1.050 billion euros. Important economic functions are the control of goods and information flows and the transport and storage of goods.

More than any other industry, it is highly standardized and thus ideal for the widespread use of digital platforms, blockchains, and AI processes. AI-equipped technology such as intelligent containers and pallets that negotiate autonomously and route and pay themselves to the recipient, or swarms of autonomous vehicles in factories, exemplify that and how value chains will function in the future.

As with the abovementioned definition of logistics, no single definition or use of the term Supply Chain Management (SCM) has been established. Overlapping with the statements made at the beginning, SCM (or: Logistics Management) should be understood as follows: SCM encompasses "both the targeted development and design of company-related and cross-company value creation systems according to logistical principles (strategic management) and the targeted control and monitoring of the flow of goods and information in the value chains under consideration (operational management)."

On the one hand, SCM understood in this way addresses the basic, goal-oriented design, which describes the initial planning as well as the structural organization of a logistic process, system, or network—in order to create it as an object and as a unit capable of action for the reasonable movement of goods and people. On the other hand, it includes the ongoing, permanent planning and design of logistical processes, systems, or networks in terms of continuous, goal-oriented further development. The execution and realization of logistical activities and their monitoring and control are also largely assigned to SCM. As a central and increasingly important component of management, SCM should be understood as an integrative, cross-functional perspective on and along the entire life cycle of logistics processes, systems, and networks. The primary elements of SCM therefore include design and organization, planning, execution/implementation, and monitoring.

The Open and Federated Approach of Silicon Economy

The potential for optimizing processes or designing new digital services and new business areas appears almost endless. Digital platforms and their AI are crucial for this. Companies like Amazon have demonstrated how a new business model can completely change and even dominate a market within a few years through the

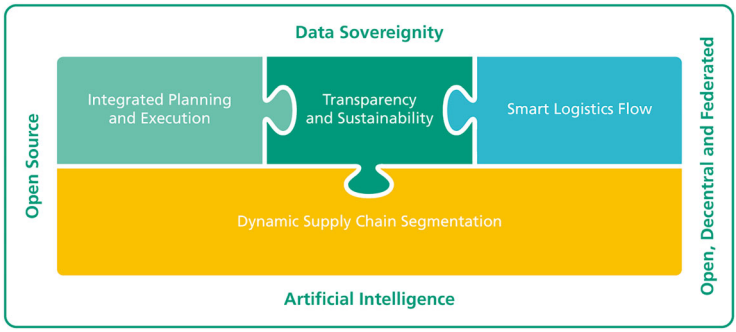


Fig. 16.1 Characteristics of Supply Chain Ecosystems in the Silicon Economy (adapted from [1])

intelligent combination of logistics and IT. The consequences of this development are already evident for companies in the logistics industry (see Fig. 16.1). The market shares of a coming platform economy have not yet been allocated in the B2B sector, but the race is on. The winners will be digital platforms with AI algorithms that permeate the entire logistics sector and thus the economy. Globally, the logistics industry is coming under the scrutiny of technology developers and investors [5, 6]. Given the high degree of standardization in logistics, it can be assumed that within a few years, logistical AI will negotiate, control, and schedule the flow of goods in this world. At the same time, the consistent expansion of the Silk Road reflects China's extraordinary commitment to the field of physical logistics in an increasingly globalized world economy.

The Silicon Economy is being developed in Germany: with over three million employees and more than a quarter of a trillion euros in annual turnover, logistics is the third largest industry in Germany, ahead of mechanical engineering and telecommunications. Deutsche Post DHL is considered the largest logistics company in the world. DB Schenker and Kühne + Nagel are two other companies from German-speaking countries among the global top ten. The same applies to the technology sector with SSI Schäfer (2nd place) or Beumer (largest manufacturer of sorting machines in the world at 8th place) [7].

The development is slower in logistics than in the times of the e-commerce hype and is therefore not perceived as decisive by the public. This is partly due to the much higher complexity of B2B logistics applications. However, this effect is increasingly being compensated for by high investments in technology and start-ups. The classic methods of Silicon Valley, which is focused on the B2C sector, are increasingly giving way to long-term commitments in terms of a Silicon Economy.

The challenges lie at both the operational and strategic levels. And the venture of implementing a Silicon Economy must develop both technical and management solutions to overcome existing limitations.

- *Heterogeneous and fragmented system landscapes:* Historically grown and highly fragmented system landscapes result in data silos and a lack of information transparency.

- *Specialized and multimodal value chains:* Different logistics areas and segments have very specific requirements for their digital infrastructure.
- *New business models & (digital) competitors:* Companies like Amazon.com or financially strong Chinese companies occupy logistics based on their B2C platforms and business models.
- *Limited financial and human resources:* Particularly in the areas of digitization and AI, there is a lack of human resources, and, due to low margins in logistics, in-house developments in this area are limited to the bare essentials.

No single company by itself has sufficient motivation, market power, or resources to succeed on its own in the logistics of a Silicon Economy. Open, federated, and strong consortia from business and science, in which technologies, de facto standards, and new business models are quickly brought together and developed, would be able to create the basis for economic use of AI solutions with new services, technologies, and applications in logistics and supply chain management and enable decisive participation for (German) SMEs. It is necessary to create open and federated platforms that all can benefit from.

The Emergence of Supply Chain Ecosystems

The leitmotif of the change toward a Silicon Economy is a new type of cooperation in global, digital ecosystems. Today's rigid and well-defined value chains are being replaced by flexible, highly dynamic and globally connected value networks. The availability and transparency of relevant data are a key prerequisite for this [8] and a decisive driver of innovation and growth. In this context, data sovereignty—understood as the ability of a natural or legal person to decide in an exclusive and sovereign way on the use of data as an economic asset—plays a key role. On the one hand, data sovereignty acts as an enabler for the use of AI applications and thus automation and autonomization in supply networks. On the other hand, it represents a basic prerequisite for the cooperation or connection of previously separate value chains and networks. The silo-like, discontinuous vertical linking of companies along the value creation process, which is currently mostly dominated by producers of an end product (OEM), can be expanded to include horizontal and spontaneous or situational cooperation between chains or networks that were previously separate or in competition.

Supply chains will be connected at all levels—autonomously and in real time. Logistics services will be traded, scheduled, and supervised via platforms. Devices will negotiate and pay autonomously. The control loops of logistics planning and scheduling will be closed. Supply chains will plan, organize, and optimize themselves autonomously. Consequently and finally, an autonomous logistics ecosystem will emerge.

Synergy potentials that clearly exceed the potential of an isolated and optimized chain are the result. These include the reduction of emissions through optimization and consolidation of transports with a simultaneous acceleration of throughput times through transport networks; the reduction of logistics costs and the vulnerability of transport chains to errors (increase in resilience); the setting of impulses for an ecologically and economically sustainable, cycle-based economy; and much more.

16.3 Silicon Economy inside

16.3.1 Big Picture/Vision

The "big picture" of the Silicon Economy shows the complete data chain: from data generation in the Internet of Things (IoT Broker) to the trading and booking of data (Blockchain Broker) to the organization of (logistical) processes (Logistics Broker) with the all-connecting secure data space (International Data Spaces (IDS)) and the platforms above it for the realization of new digital business models (see Fig. 16.2).

This digital infrastructure enables end-to-end transparency in value networks and creates trust along complete supply chains—from raw material suppliers to end customers—perhaps the most important prerequisite for the participation of all companies. Many of the technologies required for the "big picture" are already available. Starting with logistics, this comprehensive vision could be successively translated into products and business models. The key to realizing this vision is to combine the following key areas and lines of action into a holistic solution in the Silicon Economy sense.

Integration and Connectivity of Infrastructures The basic constitution of the technical infrastructure must be based on European values. Data protection, IT security, and data sovereignty must take a central role. This can be achieved in a Silicon Economy by using the components of International Data Spaces [9] to create secure data spaces that ensure data exchange between a network of companies while maintaining data sovereignty (trust anchor, trusted platform, data usage control, and no transfer of ownership rights) as a central criterion of sovereign management of data.

Realization of Open and Federated Digital Infrastructures and Platforms Participation in the digital ecosystem of the Silicon Economy should take place by means of open and barrier-free access to all basic technologies (open source; see opensource.org). The goal must be to minimize the entry thresholds into the Silicon Economy for companies and developers. These, in turn, are free to build new data-driven business models or adequate services, etc. The open basic technologies also include methods of AI.

Capabilities for Real-time Connection of Things The basis of all digital business models or services is made possible by current developments, particularly in the field of information and communications technology. By developing components for networking devices of an industrial Internet of Things with open and federated platforms, a technological basis for new services and process models will be created.

Smart Services Companies that understand how to use data as a basis for creating unique customer offerings are among the most successful companies in the world: on the list of most valuable companies, Alphabet (Google), Amazon.com, and Facebook are at the top positions with their data-driven business models, and 80% of the approximately 260 Unicorns existing in 2018 had data-driven business models

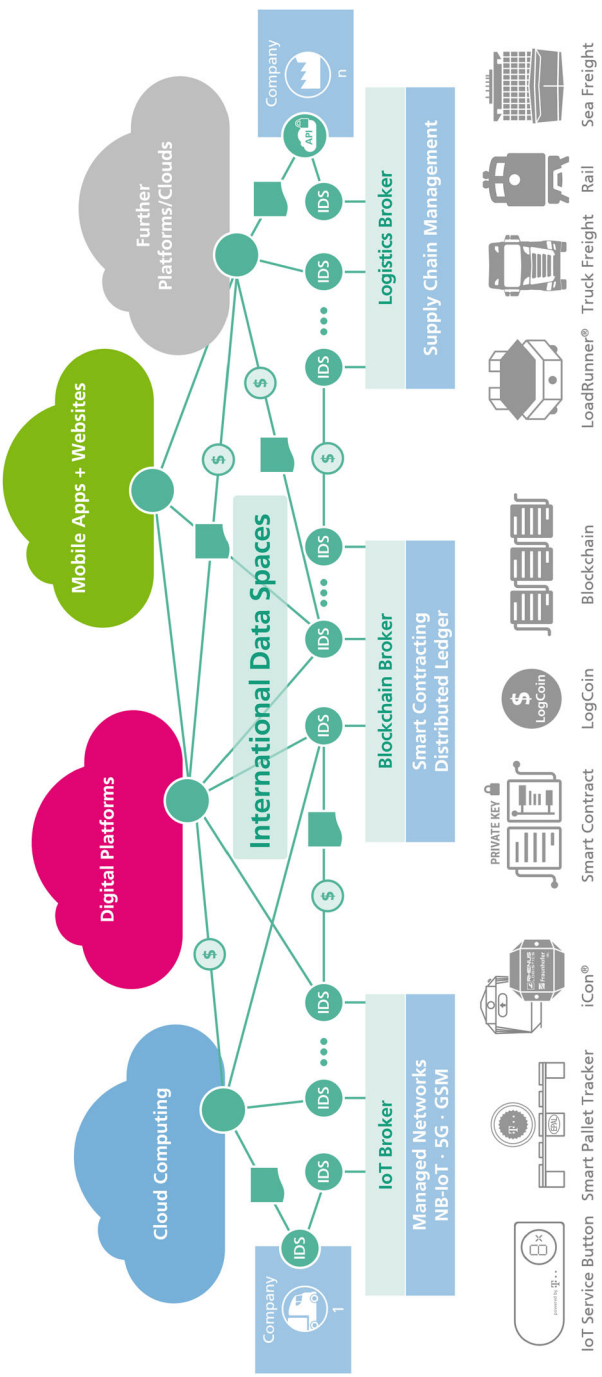


Fig. 16.2 An open and digital ecosystem as a digital infrastructure for autonomously acting and highly dynamic supply chains (© 2020, Fraunhofer IML)

[10]. This includes, for example, new solutions for digitally negotiated contracts (smart contracts), receipts and payment models by using distributed ledgers—for example, for booking and billing logistical services (transport, handling, storage), or also platforms and digital environments for autonomous planning and scheduling processes.

16.3.2 Silicon Economy Architecture

The architecture of the Silicon Economy can be characterized firstly by the central architectural patterns used. Secondly, it is characterized by its essential architectural components. Both architectural patterns and components will be briefly presented below.

16.3.2.1 Architectural Patterns

An architectural pattern is a general, reusable solution to a commonly occurring problem in software architecture. Central architectural patterns of the Silicon Economy include microservices, self-contained systems, application containers, application container orchestration, and event-driven communication.

Microservices A microservice architecture [11] is a version of a service-oriented architecture (SOA). The target system combines a set of small-scale services ("micros") that allow for easy, independent distribution, as well as independent changes and extensions. Each microservice has a high degree of autonomy and isolation and can be developed autonomously and deployed in its own Docker container (see also descriptions of the Application Container architectural pattern). Each microservice can be implemented using a different technology; they communicate with each other using lightweight protocols (fast, data-efficient protocols such as REST).

The goals of this pattern are reuse, high cohesion, low coupling, separation of concerns, single responsibility, and information hiding. Its advantages are modularity and maintainability, as well as faster adaptation to changing requirements (scaling). Furthermore, these goals are supported by the use of the additional architecture pattern called self-contained systems.

Self-contained Systems Self-contained systems (SCS) [12] divide a system into independent web applications, in this case built with microservices. They communicate preferably via asynchronous application programming interfaces (API). Here, the preferred architecture is based on the Independent Systems Architecture (ISA) best practice guidelines [13].

Application Container In the Silicon Economy, a microservice is always delivered and executed in exactly one application container. Containers do not only include the application or the microservice itself but also all the required dependencies. These

include, for example, runtime environments and system libraries. In contrast to virtual environments or virtual machines (VM), containers use core functionalities of the underlying operating system and are therefore more lightweight in comparison.

Application Container Orchestration The use of the architecture patterns described above leads to a large number of containers. This results in a high effort for the management of the containers. This is exactly where application container orchestration comes into play [14, 15]. Orchestration solutions, such as Kubernetes, perform the following tasks, for example:

1. Managing resources, such as storage
2. Management of nodes on which individual containers are run
3. Allocation of resources, such as memory and network
4. Scaling containers based on redundancy requirements
5. Monitoring containers for functionality and resource usage

Event-driven communication Software components are loosely coupled via known interfaces. In a purely event-oriented system, this knowledge is no longer necessary, since events can simply be triggered and assigned to receivers via certain criteria (e.g., topics). This enables asynchronous or event-based communication, in which the sending and receiving of data takes place asynchronously and, for example, waiting for a response from the recipient does not block the process. Events can be triggered both from outside the system, e.g., by user input or by sensor values, and internally by the system itself.

There are several implementations of this architectural pattern. In [16, 17], different asynchronous messaging protocols with wide distribution are compared and referred to the respective standards and technical description documents of the protocols. In the Silicon Economy architecture, MQTT [18, 19] and AMQP [20, 21] are used.

16.3.2.2 Architectural Components

International Data Spaces The International Data Spaces (IDS) address the design of a reference architecture model and associated reference implementations for industrial data spaces. The basis is built by the so-called IDS connectors. The functionality and security of these connectors are based on the three topics of trust anchor, trustworthy platform, and data usage control. One of the fundamental principles of IDS is to maintain sovereignty over one's own data. This principle excludes the transfer of ownership rights to any central entities or providers. IDS provides a generally applicable technical infrastructure for the exchange of any kind of data and has no direct technical reference to the logistics application field.

IDS connectors establish connectivity between the individual platforms while maintaining data sovereignty. They are used to communicate securely with the outside world.

IoT Broker, Blockchain Broker, Logistics Broker Central to the concept of the Silicon Economy are so-called brokers.

IoT Brokers are important data sources of a Silicon Economy. They connect cyber-physical systems (CPS), such as smart containers and pallets, the same way they securely connect smart machines via 5G technology, NarrowBand IoT, or conventional networks and offer data over the Internet. An IoT broker encapsulates IoT devices and their low-level protocols (data is typically sent in binary representation) and also real-time capable protocols and transforms the messages into open standards (e.g., HTTP(S), AMQP, or MQTT(s)) and open data format JSON, except for visual data (e.g., images, point clouds/3D sensor data).

Blockchain Brokers offer integrated and standardized blockchain solutions for horizontal and vertical networking in value networks. The IT architecture required for this is provided by the Blockchain Broker setup and is integral to the Silicon Economy ecosystem. Contracts (i.e., smart contracts) can be signed via Blockchain Brokers. One of the central building blocks of the Blockchain Broker is a component for smart-contract-based billing of services. Payments via crypto tokens and micropayments are also among the services offered by the brokers. Performed transactions are announced, immutably chained, and validated. Current developments in e-money and cash-on-ledger are taken into account. A framework will be created that considers the requirements of international trading operations. An integrated payment system acts as an enabler for new and disruptive business models in Industry 4.0, supporting instant payments at the value-added and enterprise level on the one hand and micro-transactions at the system level between individual CPSs on the other.

Logistics Brokers provide connectivity between services in the Silicon Economy that run on different platforms. Logistics services and their execution are organized via Logistics Brokers. They connect providers of logistics services with customers and users. This applies equally to both internal logistics/facility logistics and external logistics, e.g., the internal transport or picking of a customer order as well as the (road or rail) transport and handling of goods. In addition, the Logistics Broker is responsible for orchestration, i.e., combining several Silicon Economy services to form a meaningful business process. This means that even complex IT service business processes can be automated.

Silicon Economy Services Silicon Economy services (SE services) are developed and operated on the basis of the abovementioned infrastructure and architectural patterns. From a technical perspective, SE services consist of several microservices. Cross-company use of services and brokers takes place via IDS. Generally, independent and exchangeable web applications (SE services with web user interface as self-contained systems) are developed for specific use cases. These consist of individual software as well as standard products. For each SE service, it can be decided individually which system platform is to be used, although there are basic specifications that must be met (IDS, Web User Interface technology, programming language portfolio, database system portfolio, tool portfolio). Developer guides and style guides provide the necessary standardization. Viewed from the outside, an SE service forms a decentralized unit that communicates with other SE services (as asynchronously as possible) only via IDS. The business logic is usually implemented as microservice. Due to the clear, isolated functional scope, an SE service can be developed, operated, and maintained by one team.

16.3.3 The Role of Open Source

The abovementioned developments present a challenge for traditional logistics service providers. An already competitive market, in which profits are made through standardization efforts, is put under pressure by the requirement of increasing integration into complex, digital supply chains. Consequently, logistics service providers are facing a conflict between offering their traditional services and providing and developing new, digital products and services to meet the (digital) requirements of their customers. At the same time, already established platforms such as Amazon or Uber are increasingly entering the B2B market, followed by startups that act as fourth-party logistics providers, for example, by decoupling technologically driven, smart services and solutions from the actual logistics service [22, 23].

Platforms, through their inherent characteristics of strong network effects and the ability to incorporate complementary goods, offer traditional companies the opportunity to expand their product portfolio and value proposition. By including other companies in a platform-based ecosystem, platform service providers can achieve integration into their customers' supply chains in addition to their core logistics tasks. Until now, B2B platforms are mostly still very specialized and hardly benefit from strong, indirect network effects [24]. Moreover, the integration of complementary providers rarely succeeds and is associated with high costs. In particular, the challenge of trust relationships and the question of orchestration are crucial for platform building [25]. Since logistics, which is in any case a link between the individual supply chain partners, must establish trust and orchestrate processes and tasks, it can be attributed a central role here in building a B2B platform economy.

Open source developments in particular play a decisive role here. On the one hand, open source is a driver of a federated platform economy, as the open provision of processes and implementations enables integration into further platform and offerings from other partners and thus contributes to the growth of the ecosystem. On the other hand, open source in combination with collaborative software development is by definition a good way to work collaboratively, openly, and transparently, thus increasing the trust relationship between the partners involved.

That is why the crucial aspects of a coming federated platform economy are linked to strategies such as open source, open innovation, and collaboration. No company (in logistics) has sufficient motivation, market power, or resources to implement the "big picture" of a Silicon Economy on its own (see above). Only together open and federated platforms can be developed and thus technologies, de facto standards, and new business models can be established quickly. Consequently, it is about a "Linux for logistics" and thus about the joint foundation of a European Open Logistics Community as a driver of open developments of a Silicon Economy. Through the joint development and use of open source software and hardware, efficiency and participation are to be achieved in equal measure. Common standards, tools, and services are created, which in turn enable successful commercial use in companies, act as growth drivers for the industry, and become the starting point for new products and services that can be generated from them.

Design of an Open Source Concept for the Silicon Economy

The core component of the Silicon Economy ecosystem is a repository in which components for infrastructure (platforms and brokers) and applications (Silicon Economy services) are provided. These Silicon Economy components are mostly not a finished program or a finished software platform. They provide a reusable, common structure and (AI) algorithms for applications and devices and are generally developed with the goal of multiple use in a wide variety of logistics areas. Brokers provide a framework through which companies can connect Silicon Economy applications (e.g., web services or service platforms such as freight exchanges), services, or devices (e.g., IoT and blockchain devices). Components are developed by the open source community. They are made available as open source. Companies build their own applications on top of them and extend them in such a way that they meet their specific requirements. In sum, a logistics operating system is created, a "Linux for logistics."

Initially, this "operating system" will primarily cover existing logistics services and components that are either shared by a large number of companies or form the basis for individual implementations. This will create, use, further develop, and continuously improve a common source code basis. Previously very different IT implementations can converge both technically and functionally by using the common basis and thus realize greater interaction with less integration effort. Typical logistics use cases, such as Track&Trace or the integration of freight forwarders into corporate IT, can be easily used and reused via existing components from a repository (see Fig. 16.3).

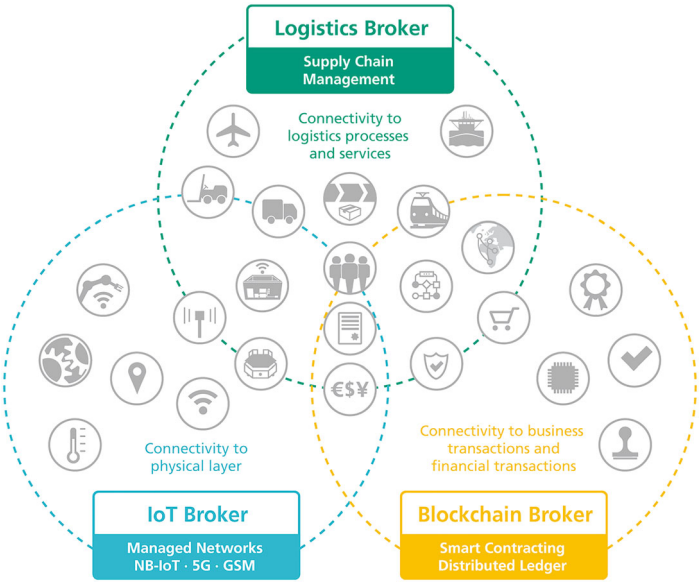


Fig. 16.3 Silicon Economy Repository (© 2020, Fraunhofer IML)

16.4 Conclusion

The world is no longer divided into East and West, but into digital and non-digital. The motto is: Whatever can be digitized will be digitized. Supply chains will be networked independently and in real time at all levels (link + virtualize). Logistics services will be traded, planned, and controlled via platforms (trade – plan – control). Devices will negotiate and pay independently (smart contracting + blockchain). The control loops of logistical planning and scheduling will close (closed loop), and supply chains will independently schedule, organize, and optimize themselves (plan – organize – optimize). Due to secure communication and data spaces, this will happen without losing sovereignty over data. All in all, an autonomous logistics ecosystem will emerge—in short: the Silicon Economy. This is too complex an undertaking for one company alone, so that open source developments and open innovation (must) take on a central role.

Acknowledgments The research reported herein was performed in the "Silicon Economy Logistics Ecosystem" project. This project is funded by the Federal Ministry of Transport and Digital Infrastructure.

References

1. PwC. (2020). *Connected and autonomous supply chain ecosystems 2025*.
2. Barreau, P. *AIVA – Artificial Intelligence Virtual Artist*
3. Indset, A. Was kommt nach der Digitalisierung? Handelsblatt 03.12.2018, Handelsblatt GmbH.
4. Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. (BITKOM) (Ed.) (2020). Open Source Monitor. Studienbericht 2019.
5. Konrad, A. (2019). Freight startup Flexport Hits $3.2 billion valuation after $1 billion investment led by SoftBank. Hg. v. *Forbes*. https://www.forbes.com/sites/alexkonrad/2019/02/21/flexport-raises-1-billion-softbank/?sh=48c3bd125650.
6. CB Insights (Ed.) (2018). Manbang Group. Online verfügbar unter. https://www.cbinsights.com/company/manbang.
7. World Bank. The Logistics Performance Index 2018.
8. Plattform Industrie 4.0 (Ed.) (2019). Digitale Ökosysteme global gestalten.
9. Fraunhofer Gesellschaft (Ed.) (2020). *International data spaces*. https://www.fraunhofer.de/de/forschung/fraunhofer-initiativen/international-data-spaces.html
10. CB Information Services (Hg.) (2017). *The increasingly crowded Unicorn Club in one infographic*. https://www.cbinsights.com/research/increasingly-crowded-unicorn-club/
11. VMware. (2020). *What is vSphere hypervisor?* https://www.vmware.com/de/products/vsphere-hypervisor.html
12. NNOQ. (2020). *SCS: Self-contained systems. Assembling software from independent systems*. https://scs-architecture.org/
13. INNOQ. (2020). *ISA: Independent systems architecture. Principles for microservices*. https://isa-principles.org/
14. Kubernetes. (2020). *What is Kubernetes?* https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/
15. Mirantis | Pure Play Open Cloud. (2020). *What is container orchestration?* https://www.mirantis.com/blog/container-orchestration/

16. Naik, N. (2017). *Choice of effective messaging protocols for IoT systems: MQTT, CoAP, AMQP and HTTP*. In 2017 IEEE International Systems Engineering Symposium (ISSE), Vienna, pp. 1–7. https://doi.org/10.1109/SysEng.2017.8088251
17. Dizdarević, J., Carpio, F., Jukan, A., & Masip-Bruin, X. (2019). A Survey of Communication Protocols for Internet of Things and Related Challenges of Fog and Cloud Computing Integration. *ACM Computing Surveys, 51*(6), 1–29. https://doi.org/10.1145/3292674
18. OASIS. (2014). *MQTT Version 3.1.1*. http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html
19. OASIS. (2019). *MQTT Version 5.0*. https://docs.oasis-open.org/mqtt/mqtt/v5.0/mqtt-v5.0.pdf
20. ISO. (2014). *ISO/IEC 19464:2014. Information technology — Advanced Message Queuing Protocol (AMQP) v1.0 specification*. https://www.iso.org/standard/64955.html
21. OASIS. (2020). *AMQP- AMQP Working Group*, 0-9-1, 0. https://www.amqp.org/specification/0-9-1/amqp-org-download
22. Seiter, M., Autenrieht, P., & Schüler, F. (2019). Logistikdienstleister im Zeitalter digitaler Plattformen. In M. Schröder & K. Wegner (Eds.), *Logistik im Wandel der Zeit. Von der Produktionssteuerung zu vernetzten Supply Chains*. Springer Gabler. S. 585–600.
23. Sucky, E., & Asdecker, B. (2019). Digitale Transformation der Logistik. Wie verändern neue Geschäftsmodelle die Branche? In W. Becker, B. Eierle, A. Fliaster, B. Ivens, A. Leischnig, A. Pflaum, & E. Sucky (Eds.), *Geschäftsmodelle in der digitalen Welt. Strategien, Prozesse und Praxiserfahrungen*. Gabler Verlag. S. 191–212.
24. Bundesministerium für Wirtschaft und Energie (BMWi). (2019). *Die volkswirtschaftliche Bedeutung von digitalen B2B-Plattformen im Verarbeitenden Gewerbe*. Berlin.
25. Tian, J., Vanderstraeten, J., Matthyssens, P., & Shen, L. (2021). Developing and leveraging platforms in a traditional industry: An orchestration and co-creation perspective. *Industrial Marketing Management, 92*(1), 14–33.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 17
Agricultural Data Space



Ralf Kalmar, Bernd Rauch, Jörg Dörr, and Peter Liggesmeyer

Abstract The digital transformation strongly affects the agricultural domain. Still, there is a lot of potential for optimization in many work and business processes. In the current agricultural digital ecosystem, numerous isolated, often non-interoperable solutions exist. In this chapter, we motivate the need and added value of an "Agricultural Data Space" (ADS for short). We outline an ADS concept, which resulted mainly from the Fraunhofer lighthouse project "Cognitive Agriculture" (COGNAC) and describe the necessary prerequisites and technical solution approaches. Complemented by the possibilities of a transparent and open marketplace for data, digital products, and software services, such a data space would address many of the existing obstacles to widespread acceptance and take-up of digital technologies. Overall, an ADS as part of an extended digital ecosystem will significantly advance digitalization in agriculture. In the end, we provide application scenarios for which an agricultural data space can add value.

R. Kalmar (✉) · B. Rauch
Fraunhofer Institute for Experimental Software Engineering IESE, Kaiserslautern, Germany
e-mail: ralf.kalmar@iese.fraunhofer.de; bernd.rauch@iese.fraunhofer.de

J. Dörr
Fraunhofer Institute for Experimental Software Engineering IESE, Kaiserslautern, Germany

Chair of Digital Farming, TU Kaiserslautern, Kaiserslautern, Germany
e-mail: joerg.doerr@iese.fraunhofer.de

P. Liggesmeyer
Fraunhofer Institute for Experimental Software Engineering IESE, Kaiserslautern, Germany

Chair of Software Engineering: Dependability, TU Kaiserslautern, Kaiserslautern, Germany
e-mail: peter.liggesmeyer@iese.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_17

17.1 Digital Transformation in Agriculture

17.1.1 The Agricultural Domain

Agriculture is the oldest domain of mankind, lasting back more than 10,000 years. With the industrial revolution in the last century, both crop harvesting and livestock breeding have been highly optimized. In 1960 a German farmer fed 17 people—in 2017 this number had risen to 140 people—an enormous increase. Feeding the world's population is a great challenge, and zero hunger is one of the UN sustainable development goals. For 8.9% of the world's population, food is a scare resource and so is farmland. Most of the landmass is already cultivated, and erosion and desertification are putting farmable land at risk. Optimizing yield output is therefore one of the prime goals in agriculture. In the past years, a second goal has gained attention by society and also policy makers: sustainability. This addresses in particular environmental sustainability but also includes economic and social sustainability.

There is still a lot of potential for optimization in many processes in the agricultural value chain, especially when looking at the "big picture." Working with plants and animals faces agriculture with the complexity of the biosphere of our planet and many physical and biochemical processes. This complexity and unpredictable weather impact make it much more difficult to control and optimize than, for example, a production environment in a factory. For digitalization, this results in the challenge building many complex models, with lots of parameters and a high amount of required data.

Also, on the dimension of the horizontal value chain, agriculture combines many different stakeholders in the production chain, including farmers, contractors, manufacturers of agricultural machinery, resource providers, public authorities, traders, food processors, stores, and finally consumers.

Among these stakeholders, there is a large variety in different aspects like size and operating style, from global operating companies down to small farms operated by one family.

17.1.2 Agricultural Digital Ecosystem

Improvement potential in agriculture mainly exists in the work processes and the higher-level planning and decision processes. This requires comprehensive contextual information from the past, the present, and the predicted future. Collection, processing, and interpretation of this requires automation (i.e., through software) in order to be recorded and provided in the necessary quantity and quality.

To enable optimal operational management supported by software services, all data required for decision-making and process automation must therefore be available in digital form.

As of today, many processes in agriculture cannot be interconnected and automatically provided with the necessary information because of lacking interoperability regarding data, interfaces, and protocols. There is a fragmented landscape of islands of smaller domain ecosystems. One larger group (of yet not fully interoperable systems) is formed by machine manufacturers that comprise processes executed by agricultural machinery and the corresponding data, such as field data, machine data, or crop care documentation. Another group of ecosystems are grouped around business processes for farm management, including planning processes, resource management, sales, certification, or taxes. Bringing these and others together, so that, for example, a service provider for crop care consultancy could offer the same service on many platforms using the same implementation and being provided with the necessary data, would greatly improve the infrastructure and give a boost for digital transformation.

> An agricultural digital ecosystem should provide an infrastructure for efficiently supporting all agricultural business and work processes with services and information, provide means for flexible adaptation of needs from different stakeholders, and enable new business models.

Each digital ecosystem is formed by two or more stakeholder groups. In agriculture, these stakeholder groups comprise farmers, contractors, consultants, public authorities, research organizations, machine manufacturers, operating resource producers (seeds, fertilizer, forage, etc.), traders, logistics, services, food processors, commerce, and customers—just to name the big ones. It becomes clear that bringing these together, even in smaller steps, is a great endeavor.

17.1.3 Domain-Specific Challenges and Requirements

Apart from the great variety of processes, existing solutions, and resulting lack of data interoperability, there are other domain-specific challenges:

- Connectivity and network infrastructure: many rural areas have no high-speed Internet access or even none at all. Despite political promises, improvement is very slow.
- Offline-data collection (as a result of the above) and the need of synchronization and integration of data from different sources.
- Agriculture gets more and more attention from the general public, especially with the stronger emphasis on sustainability. It operates in public space—cultural landscape is also a public space, which impacts the environment, and therefore the interests of people.

- SME-structured farmers need to be educated in digital technologies. This is important to create acceptance and qualify farmers to actively participate in creating added value with the data.

Especially for the acceptance of systems by farmers, data sovereignty is a key factor [1]. Farming industry has defined a "Code of Conduct on Agricultural Data Sharing",[1] which addresses basic rules for using shared data. The initiative "Ag Data Transparent"[2] goes one step further and tries to certify products regarding 11 key questions of data usage. However, in most cases today, as soon as farmers provide their own data for evaluations in digital services, they currently feel to lose their sovereignty over this data. The platform providers are called upon to create solutions with which farmers can control and monitor data sovereignty easily and in a self-determined manner. Interoperability for universal data usage should not end with the farmer, however; rather, it should be enabled along the entire value chain from processing operations to the consumer. This is the only way that all stakeholders can benefit from data analyses and decision support based upon them and the only way in which comprehensive transparency can be achieved.

17.2 Agricultural Data Space (ADS)

In the agricultural domain, there are many different, partially isolated platforms and systems with redundant data, services, and software solutions. Even as integration of platforms is moving on with bilateral connectivity and the emergence of data routers, there is still no thorough connectivity for components across the digital domain. Furthermore, specific challenges of the agricultural domain, like offline capabilities or a lack of IT infrastructure at farms, have to be considered for any integration.

An important aspect to keep in mind is that even though there is a lack of interoperability and data sovereignty, the design of an ADS needs to consider the already emerged digital ecosystems. New designs cannot start from scratch but have to be adapted and integrated in the already existing environment by accepting certain boundaries and specifications. This is also where we see the main contribution of a domain-specific adaptation of IDS concepts. The implementation of IDS functionality is possible as well as promising, but it needs to be adapted to the status quo and possibly extended. Furthermore, the agricultural domain is strongly influenced by the business interests of single, big market players, which could hamper a fully interoperable data space when they strive to concentrate data within their respective realms.

[1] EU CODE OF CONDUCT ON AGRICULTURAL DATA SHARING BY CONTRACTUAL AGREEMENT, 05/2020. https://copa-cogeca.eu/Publications

[2] Ag Data Transparency Evaluator Inc., https://www.agdatatransparent.com/

The Fraunhofer lighthouse project Cognitive Agriculture[3] (COGNAC) researches and develops concepts for such an ADS while placing challenges like data sovereignty and interoperability in the middle of a thriving, digital domain ecosystem for agriculture. In the following sections, we describe the domain architecture as we perceive it, explain possible levels how IDS concepts can be used for an ADS, and briefly explain the benefits of an interoperable ADS.

17.2.1 Domain Architecture

Following the concepts of the International Data Space (IDS), the ADS comprises all components of digital ecosystems that generate, store, manage, or consume data and are interconnected. Just as in typical digital ecosystems, the ADS needs one or more digital platforms as key part of its infrastructure. One of the key goals is the best and most holistic possible integration of the various components of the digital agricultural domain ecosystem. To this end, we do not envision the ADS as being enabled by a single or specific digital platform as the sole player in the ecosystem. The reason for this is not only to reflect the current situation in an already emerged domain ecosystem but also the vast diversity of agricultural business and work processes and, consequently, the broad variety of specific existing (sub-)ecosystems in the domain. In order to integrate actors, services, and data, we propose a framework or reference architecture for ADS-ready components that fulfill basic requirements like interoperability and data sovereignty.

Figure 17.1 depicts an exemplary and conceptual ADS with multiple digital (sub-)ecosystems like digital platforms of machine manufactures, a routing platform, and service-specific platforms like Farm Management Information Systems. The single digital platforms would not need to be interconnected to all other existing components; data transfer or service interoperability can also be achieved by utilizing routing platforms. Connectivity would comprise syntactical and semantical interoperability for APIs and data.

For the ADS, we assume that there is no need for one big ecosystem that completely connects each and every system and player. There are many diverse facets; some players don't need to interact. Rather, the ADS framework would enable the fundamental connectivity so that all participants could engage in collaborations, but we moreover expect the ADS to develop industry-based factions like arable farming, livestock breeding, produce refinement, and so on that will keep a certain autonomy. While those factions are loosely coupled and share small portions of the value network, a tighter coupling can be expected within those factions which will develop more deeply integrated sub-ecosystems.

In this context, due to the existing (sub-)ecosystems, we strongly advocate flexibility when it comes to variants of communication channels. This means, e.g.,

[3] www.cognitive-agriculture.de

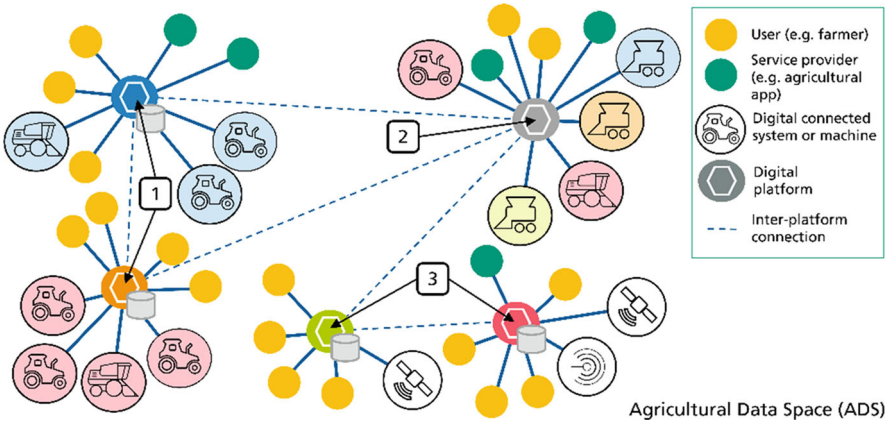


Fig. 17.1 The agricultural data space as a domain ecosystem with interconnected digital platforms and specific digital (sub-)ecosystems (machine manufacturers platforms (1), routing platform (2), and service specific platforms (3)). (illustration ©2021, Fraunhofer IESE)

that in an envisioned ADS, it should not be mandatory to utilize IDS connectors as the sole means for data exchange and protection, but there could be different mechanisms bound to the needed level of data sovereignty. If some entity is in the need of high level of data protection, IDS connectors should be used. If not, data could just be transferred via existing access controlled interfaces.

Given the diversity of the agricultural domain and the status quo with already developed ecosystems, the ADS needs to provide a high level of flexibility in both concepts and technologies to be accepted in the domain. On the other hand, the IDS provides functionalities and concepts that would fulfill yet unmet demands of the domain and thus contribute to a successful digital transformation in agriculture.

17.2.2 Possible Levels of IDS Integration

Given a multitude of perspectives in the agricultural domain, we see various entry points for an integration of IDS concepts. Currently, most activities working on connectivity and interoperability are based in a context where farmers' data is worked on and where farmers increasingly demand data sovereignty. On the other side, industry players have also an interest in data sovereignty as they seek to protect sensitive machine data like exact fuel consumption in certain situations.

If one would like to segment the implementation of an ADS in phases, we would recommend to start with the farmers' context, which can be further divided in arable and livestock farming. Farmers increasingly use software solutions and digitized machinery when handling their work processes. Farm management information systems (FMIS) collect, organize, and process data that is produced as well as consumed by machinery in the field. Figure 17.2 gives an exemplary illustration of

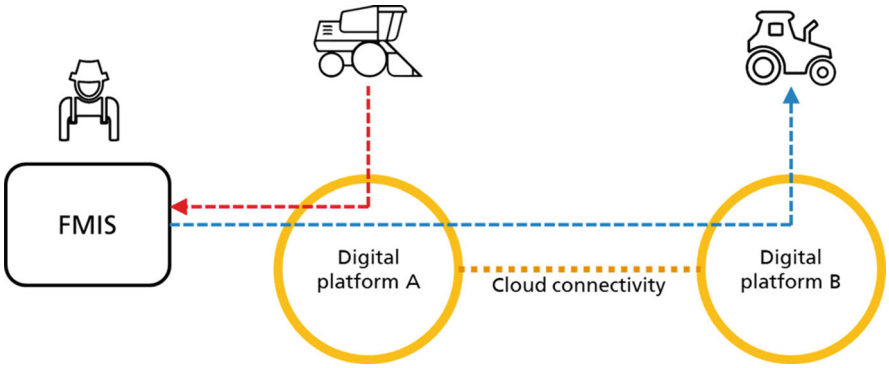


Fig. 17.2 Data flow between machinery and systems crossing platforms. (illustration ©2021, Fraunhofer IESE)

a typical data flow in arable farming, where data flows from different machinery to a software system (FMIS) and vice versa while being exchanged across different platforms. Crossing platforms is here a necessity. Machinery manufacturers often make data and API access to their machines exclusively available via own digital platforms.

In such a scenario, there is a demand for interoperability and data sovereignty between the farmers' software, digital platforms, and systems like agricultural machines. There are many activities in research and industry working on connectivity and interoperability, but data sovereignty so far is most times just implemented as access control, which is not sufficient to assure full sovereignty (Chap. 8).

As stated initially, IDS concepts could be applied on multiple levels. One general aspect to consider is that there are typically three actors involved in a scenario that focuses on farmers:

- The farmers as the data owners.
- The Platform providers (often machine manufacturers).
- Third-party system or service providers (FMIS, digital and farming services, etc.)

Farmers as well as third-party providers are participants in an ecosystem that is enabled by a digital platform, but they still are independent entities. Consequently, they strive to have data sovereignty regarding data from their respective assets against platform providers. In such a context, one could think of different levels to integrate IDS concepts. We start with a discussion of independent IDS integration possibilities, discuss possible problems in adoption, and conclude with a proposal for a hybrid approach.

- Inter-platform connectivity

  This approach would be the analogy to having IDS protection between companies, but in our scenario, it is about the data of the farmers and not the platform providers. IDS connectors can be used to connect the different platforms. A data broker could be aware of all existing services offered on the various

platforms. The viewpoint of the end-user is less integrated compared to other options.

- System and IoT connectivity

  In this approach, IDS connectors can be used to connect IoT devices and the various systems. Data could be protected end-to-end between system or IoT and software like FMIS. While this would be a very thorough protection, it is unlikely to succeed on short term as it would contradict the current machine manufacturers' goals and would be very demanding to implement.
- User context connectivity

  Here, IDS protection would be implemented at a user's perimeter, meaning that farmers protect their respective data with own infrastructure while system and service providers do the same on their side, again, a highly demanding approach where farmers would need to build up or outsource extensive infrastructure, which is limiting the applicability of this solution.

As a result, a hybrid approach could be a good way to go. In this, IDS concepts could be applied on platform level, while typical systems and users within the platform context build up trust and/or legal frameworks towards the platform owners. For more sensitive data, one still can power up single systems and contexts with IDS protection at their respective perimeters as needed. In order to realize such a hybrid approach, further requirements and concepts have to be considered and can help to show the value of an interoperable data space in agriculture.

Federated Services

Another requirement to look at are federated services for the domain. As initially stated, we do envision the ADS as a digital ecosystem consisting of decentralized digital platforms, systems, and users. Those need functions to navigate the ADS like marketplaces for services and data in the ecosystem. In a day-to-day-example, a service operator in ecosystem A would like to provide a service to a farmer that has his or her data stored in ecosystem B. Via a data marketplace, the service provider can find the needed data for the service and fulfill its provision.

Digital Twins

In order to exploit the full benefit from making data across platforms and systems accessible, the domain could build up an infrastructure for digital twins. For arable farming those would be digital field twins, while for livestock farming, single animals could be twinned. In such a context, central functions like clearinghouse logging can be done in the twin directly to capsule data with access information. In addition, data usage policies would also be integrated in those twin objects. Such a twin concept supports the idea of a decentralized environment, where twins can exist at any arbitrary platform in the ADS as long as they fulfill the requirements for interoperability and can be used across multiple platforms. This also helps in organizing data of physical assets, as it is no longer distributed across different systems.

Semantic Interoperability

For semantic interoperability, we do not propose another broad, common standard but moreover a common, basic meta-model for principals of data exchange along

with flexible mechanisms that supports semantic interoperability between datasets. Incorporating vocabularies and ontologies, like in the IDS, one can integrate conversion functionality directly in the digital twins.

17.2.3 General Benefits of an Interoperable Agricultural Data Space

The discussed interoperable data space, the ADS, supports the digital transformation of agriculture by enabling thorough data sovereignty and interoperability. A common framework for digital twins could enable multiple actors and systems to work collaboratively on specific sets of data, which supports the robustness of processes and enables the ecosystem to develop a reliable data economy. By ensuring data sovereignty in various levels as needed, an ADS can enable the willingness to share data, which further supports the digital transformation of agriculture. More data will become available for new business models and new services that add value to the domain. Still, the data providers can keep control over their data. In the next section, we will describe concrete usage scenarios how an ADS infrastructure can offer benefit to the domain and its stakeholders.

17.3 Application Scenarios

In the following, we describe three different application scenarios where an ADS can provide value.

17.3.1 Sustainable Management of Nutrient Cycle

One application example for the ADS and corresponding services is the evaluation of ecological and economic sustainability via the nutrient cycle. In agriculture, a balanced and appropriate nutrient cycle forms the core of the efficient, productive, and sustainable production of plant as well as animal products. In addition to documentation, e.g., for monitoring by public authorities, the focus is increasingly on the optimization of the nutrient cycle. In our example, we consider a dairy farm with crop cultivation and grassland farming. Optimization can only be achieved by linking different specialist areas. On the one hand, there is agricultural machinery data, which records both the quantities of fertilizer applied and, indirectly via yield mapping, the nutrients applied. Various suppliers of sensor systems record soil, plant, and weather data. In the dairy farm sector, conclusions can be drawn about the nutrients entering and leaving the cycle by looking at data about the feeding together with the milk yield. Fig. 17.3 depicts the general steps in which nutrients are

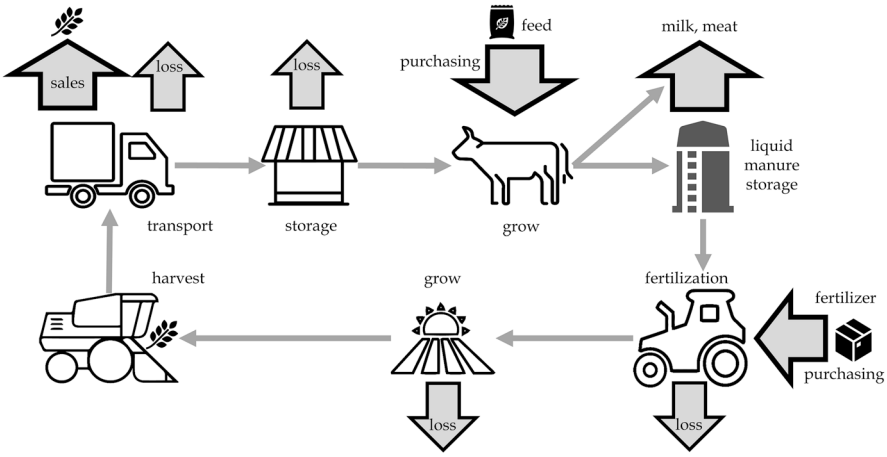


Fig. 17.3 Nutrients are brought in and out at many stages in the nutrient cycle and therefore require multiple measurements and data processing

generated. However, in almost all places (air, soil, groundwater, etc.), nutrients are leaked, and these losses can be reduced to a certain level (although practical implementability must be given and feasible).

In this example, many different stakeholder groups are represented, which either elicit, collect, or evaluate data, or record and store data for legal documentation. In this case, the ADS approach can offer supporting services and at the same time protect the farmer's data, as not all data necessarily needs to be and should be accessed by everyone. In this example, data sovereignty plays an important role. Data is generated from different sources and stakeholders. In order to optimize the nutrient cycle, the relevant data must be complete and must be available in sufficient quality. Important components here are interoperability, uniform ontologies, and cognitive processing of the data. Missing data, for example, must be interpolated or modeled accordingly. By representing the nutrient cycle in the form of a digital twin, the farmer can get information about their current nutrient balance and thus identify possible problem areas. Based on the digital twin, services offering the farmer appropriate decision-making aids can be purchased on the service marketplace for a fee. Since the nutrient cycle is a highly complex representation of various parameters, there are several sub-areas in which services can provide support. Examples include optimal feeding aimed at reducing the amount of nitrogen in the liquid manure, subplot-specific fertilization, or improved utilization of the nutrients in the liquid manure.

17.3.2 New Business Models and Fulfilling Legal Obligations with Data in the ADS

Agriculture is a diverse sector, and farms are part of a complex network with various interest groups. A large amount of different data is generated on a farm, which in the

future will be increasingly collected, stored, evaluated, and documented by many different systems. For this reason, the approaches of the ADS with its possibilities for integrated data access and data usage control are of essential importance. The collected data can now be used in various ways bringing benefits for the farmer or the whole agricultural ecosystem. On the one hand, the documented data of the farmer can be made accessible for service providers who have an interest in this data (e.g., to compare yield, evaluate the effectiveness and efficiency of equipment, get data on soil quality, or provide transparency in the food chain). The farmer can participate in the profit made with these data-intensive applications. On the other hand, the documented data can be used by the farmer in order to fulfill legal obligations. Here, the IDS mechanisms of usage control can be used to its full potential: in general, the farmers are very sensitive towards which data from their farming activities will be given to which public authority. On the other hand, the farmers experience a strong burden by many obligations to document activities. Therefore, the farmer can specify usage control policies and give specific public authorities usage rights to obtain the data. It is important to mention that in this scenario the farmer needs to be in full control of this data usage.

17.3.3 Governmental Platforms

Besides the farmers themselves and the companies in the agricultural domain, the public authorities are important stakeholders in the domain. Recently, a study was published on the feasibility of a governmental data platform in the agricultural ecosystem [1]. In this study, various concepts are outlined how the public authorities could build up a (set of) platforms in order to handle data from public authorities and also specific data from the farmers for various purposes. Such purposes are providing information from governmental institutions, providing information on regulations, but also obtaining data from farmers that would like to get subsidies or for fulfilling regulations (see also previous section). This governmental platform can conceptually be seen as an own (sub)ecosystem in an agricultural data space. One result of this study was that the data should not be made available for the consumption by the farmers via portals, but also in a machine-readable, interoperable fashion. In order to exploit the full benefit of interoperability and data exchange in the domain, such governmental platforms should therefore be connected to the nongovernmental IT systems and digital platforms. If this is achieved, farmers as well as companies in the agricultural sector can benefit from the governmental data that is made available.

17.4 Summary and Outlook

In this chapter, we have motivated and presented the concept of an Agricultural Data Space, which can greatly advance digitalization in agriculture. To do so, the Agricultural Data Space takes up the concepts of the International Data Spaces and adapts and extends them with solutions for the agricultural sector.

This data space integrates data and services from different plat- forms without restricting them. Enabling technology is required for this, into which further data and services can be integrated successively, provided that other platforms implement a corresponding connector and describe data access via a service directory.

Many of the elements outlined above address current challenges, but even greater investments are required on the part of providers and users to realize the vision of a common data space for the agricultural sector.

The current activities around the GAIA-X initiative (Chap. 4) which are supported by the IDS can be a strong facilitator to address these challenges. In GAIA-X, agriculture is an own domain, and first use cases are realized, e.g., by the nationally funded projects Agri-Gaia [2] and NaLamKI [3]. Those projects aim to make use of AI components to leverage the potential of agricultural data. For this, the ADS and its concepts can be a foundation to get access to the different data sources. As the challenges in agriculture are typically not solved on national levels, but on a broader scale, and GAIA-X is a European initiative supported by many countries, it has the potential to support the initialization and realization of concepts for a European ADS.

References

1. Bartels, N., Doerr, J., & Fehrmann, J., et al., (2020). *Machbarkeitsstudie zu staatlichen digitalen Datenplattformen für die Landwirtschaft*. https://www.bmel.de/DE/themen/digitalisierung/datenplattform-machbarkeitsstudie.html; zuletzt besucht April 11, 2021.
2. Webpage of the Project Agri-Gaia. https://www.bmwi.de/Redaktion/EN/Artikel/Digital-World/GAIA-X-Use-Cases/agri-gaia.html. Last visited April 14th, 2021.
3. Webpage of the Project NaLamKI. https://www.digitale-technologien.de/DT/Redaktion/DE/Standardartikel/KuenstlicheIntelligenzProjekte/KuenstlicheIntelligenzProjekte_ZweiterFoerderaufruf/ki-projekt_NaLamKI.html. Last visited April 14th, 2021.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 18
Medical Data Spaces in Healthcare Data Ecosystems



Thomas Berlage, Carsten Claussen, Sandra Geisler, Carlos A. Velasco, and Stefan Decker

Abstract Exchange of sensitive medical data between healthcare providers and researchers requires a particularly high level of trust and security. Involving patients and citizen into this process increases transparency and may improve the outcome of preventive, diagnostic, and therapeutic measures. We propose to build the structure of a healthcare ecosystem on the basis of "apps" that not only hold (and exchange) health data but support user interaction and healthcare process management. Emphasis on process support may also be used to improve data quality, which is an important prerequisite for evidence-based medicine and the training and usage of future AI tools.

18.1 Introduction

The state of digital transformation in the healthcare sector shows a very heterogeneous picture. While radiology departments use advanced AI tools to more accurately interpret their medical images, many resident physicians still rely on fax communication. Patients use their smartwatch to record ECGs, while their doctor records blood pressure in a paper file.

The biggest challenge currently is the transfer of relevant information between different organizations in the healthcare system. This is particularly relevant in a complex healthcare system such as in Germany. Different public and private healthcare providers (hospitals, resident practitioners, and specialized centers)

T. Berlage · S. Geisler (✉) · C. A. Velasco · S. Decker
Fraunhofer Institute for Applied Information Technology FIT, Schloss Birlinghoven, Sankt Augustin, Germany
e-mail: thomas.berlage@fit.fraunhofer.de; sandra.geisler@fit.fraunhofer.de; carlos.velasco.nunez@fit.fraunhofer.de; stefan.decker@fit.fraunhofer.de

C. Claussen
Fraunhofer-Institute for Translational Medicine and Pharmakology ITMP, VolksparkLabs, Hamburg, Germany
e-mail: carsten.claussen@itmp.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_18

have specific roles in this system. While providers use information technology for their own purposes, an overarching digital infrastructure is just starting to appear. One of the reasons is the lack of agreements or regulations concerning the provision or use of data by actors in the healthcare sector.

We are looking at this sector from a data ecosystem perspective. In a data ecosystem, exchange of data results in an overall benefit. In this case, patients should receive a more effective and more efficient treatment enabled by a collaboration of healthcare providers. For example, continuous treatment after a hospital visit should be facilitated. A technical infrastructure for data exchange is necessary, but not sufficient. Digital workflows also require harmonized processes and data structures. Healthcare providers need incentives (or regulations) from the ecosystem to establish collaborative workflows. A technical infrastructure alone will not push the digital transformation of the healthcare system.

A further challenge for the healthcare sector is to collect high-quality datasets of significant size for the introduction of precision medicine. Again, the technical challenge is being able to share datasets among organizations, but equally important is the harmonization of the procedures the data are based on. Appropriate incentives are necessary to reflect the value of such datasets and to properly share the benefit among the research community.

And finally, patients and consumers have to be integrated into this ecosystem, both as providers of their personal data and consumers of data-based prevention, diagnosis, and treatment.

Similar to logistics, healthcare is largely a real-time process, aiming at an immediate outcome to the patient's benefit. In contrast, medical research has often been a separate process where the results slowly flow back into healthcare via publications and, later, formal guidelines. In precision medicine, the processes of healthcare and associated research are much more integrated. Patient data, such as genome sequences, are immediately used to identify the best therapy in the light of available data from recent cases. Subsequent patient engagement and patient monitoring will directly deliver detailed outcome information for the benefit of future patients.

In the case of precision medicine, digital processes not only benefit efficiency and customer experience, but will be indispensable for more effective therapies and successful prevention. Therefore, we expect this sector to be the key driver to the digital transformation of the healthcare system in the future.

Therefore, building a medical data space is contingent on establishing innovative data-based healthcare concepts. Technical developments and healthcare innovation have to go hand in hand. In this paper, we are trying to capture a sufficient part of the heterogeneous healthcare ecosystem via three representative scenarios:

1. Health and disease management
2. Integrated care
3. Precision medicine

In most of these scenarios, we have been involved in multiple innovation projects. We have tried to compile the experiences in a more abstract way in these scenarios.

18.2 Elements of Medical Data Spaces

The main healthcare providers are hospitals and resident practitioners, including various forms of collaborative medical institutions. Each healthcare provider needs to keep electronic medical records for all patients seen and treated. In countries with a centralized healthcare system, all providers are part of the same organization and can share medical records, while in countries such as Germany, the healthcare providers are independent and partially competing with each other.

But even within organizations, various independent information systems exist. For example, while a hospital uses a hospital information system (HIS) to organize the care process, various specialized information systems, such as radiology information systems (RIS) or operating room management systems, often from different vendors, cover more specialized needs. Each information system manages a subset of patient-related data.

This already constitutes the example of a data space. A *data space* is a collection of independently administered data sources or repositories [1]. Their independence promotes modularity in the overall technical and socioeconomic system, which in turn fosters local specialization and innovation.

In medicine, there are already approaches for such data spaces within organizations. Medical data protection regulations might already restrict visibility and use of certain data within the organization. The IHE (Integrating the healthcare enterprise) standards, particularly HL7/FHIR® (Fast Healthcare Interoperability Resources; [2]), define data exchange protocols and formats needed to synchronize the different information systems along specific workflows (see Fig. 18.1).

The architecture defined by the International Data Spaces Association [3] is an example how a data space can be constructed from individual data providers and consumers. The key is a suitable meta-model that allows participants to define and execute data exchange contracts in a trusted way.

We talk about a *data ecosystem* when different organizations are involved [4–6]. In this case, a contractual relationship typically needs to be established [7]. Medical data transfer requires a legal basis, for example, by laws concerning medical registers or by informed consent of the patients. The receiving institution might also have to rely on quality standards implemented by the sending institution to trust the data.

18.3 Scenario 1: Health and Disease Management

Many diseases require monitoring and treatment over an extended period of time. Specifically for chronic diseases, treatment regimes can be lifelong, such as in diabetes. Others are resolved into a stable state over the course of months or years. In such a situation, both patients and healthcare providers profit from a disease-

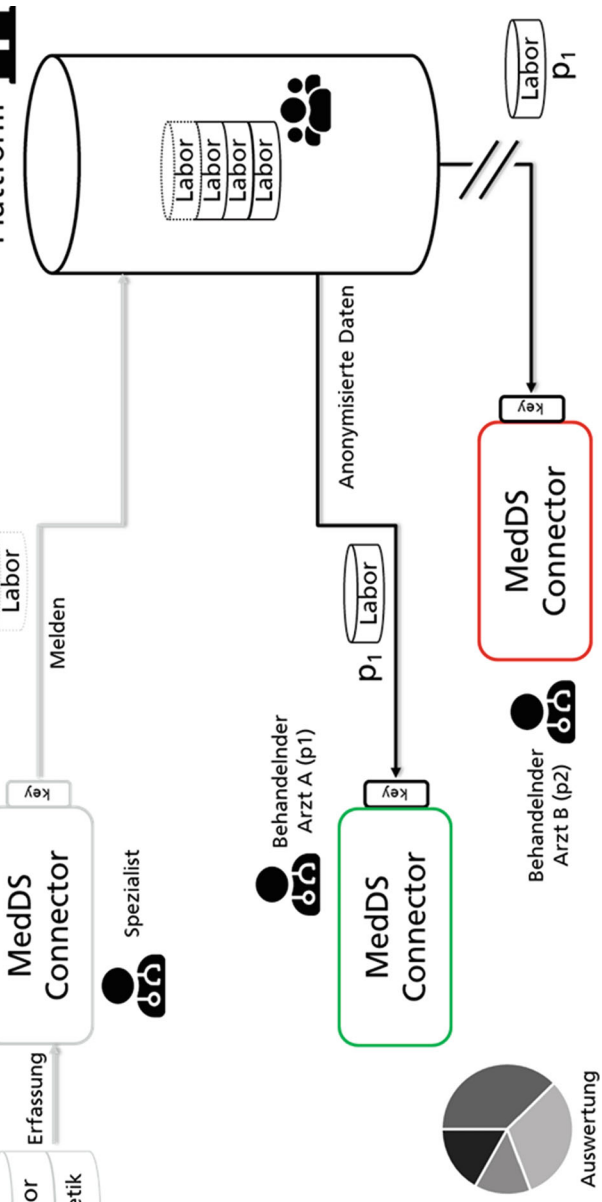


Fig. 18.1 Using the IDSA architecture to aggregate patient data from different healthcare providers

oriented data exchange platform that facilitates and documents decisions and interventions and improves outcomes and quality of life.

SALUS is a project financed by the German Innovationsfonds that implements a new approach to health management for glaucoma patients.[1] In the SALUS approach, progression of a glaucoma disease is monitored with the help of an electronic health record app that includes multiple doctor visits and also allows data to be collected at home reducing time-consuming hospital visits. Usually, a profile of the intraocular pressure (the main parameter influencing the progression of the disease) of 4 h interval measurements is created over several days to assess day-and-night variation of the pressure. This currently necessitates hospitalization. Using a new consumer-operable measurement device in conjunction with a specialized app, SALUS avoids this hospitalization for both financial and comfort benefits.

The process is supervised by a resident ophthalmologist who can interact with the app in this role. A web-based information system collects the data and the status of the process. A web-based system is used because:

- the new protocol will be used by hundreds of ophthalmologists that do not use the same OS,
- data will be sent from the mobile devices directly to the system using interfaces that traditional medical information systems are not equipped with,
- the patient should have access to the information via a mobile device.

SALUS intends to improve therapy by optimizing medication and improve decisions for interventional measures. Medication is obviously relevant for other medical decisions outside ophthalmology and will also be documented (partly) in other information systems. Information about any previous interventions would also be maintained elsewhere. While the SALUS app currently interfaces with manufacturer apps of the measurement devices (Fig. 18.2 shows an example view of measurements), it would be highly desirable to exchange medication and intervention data as well.

18.4 Scenario 2: Integrated Care

The World Health Organization (WHO) defines integrated care as:

> A concept bringing together inputs, delivery, management and organization of services related to diagnosis, treatment, care, rehabilitation and health promotion. Integration is a means to improve services in relation to access, quality, user satisfaction and efficiency [8].

As pointed out by the "European Commission Report on the Impact of Demographic Change" [9], the European healthcare systems are facing several challenges because of the growth of the ageing population, combined with an increasing number of

[1] https://www.ukm.de/index.php?id=salus-glaukom

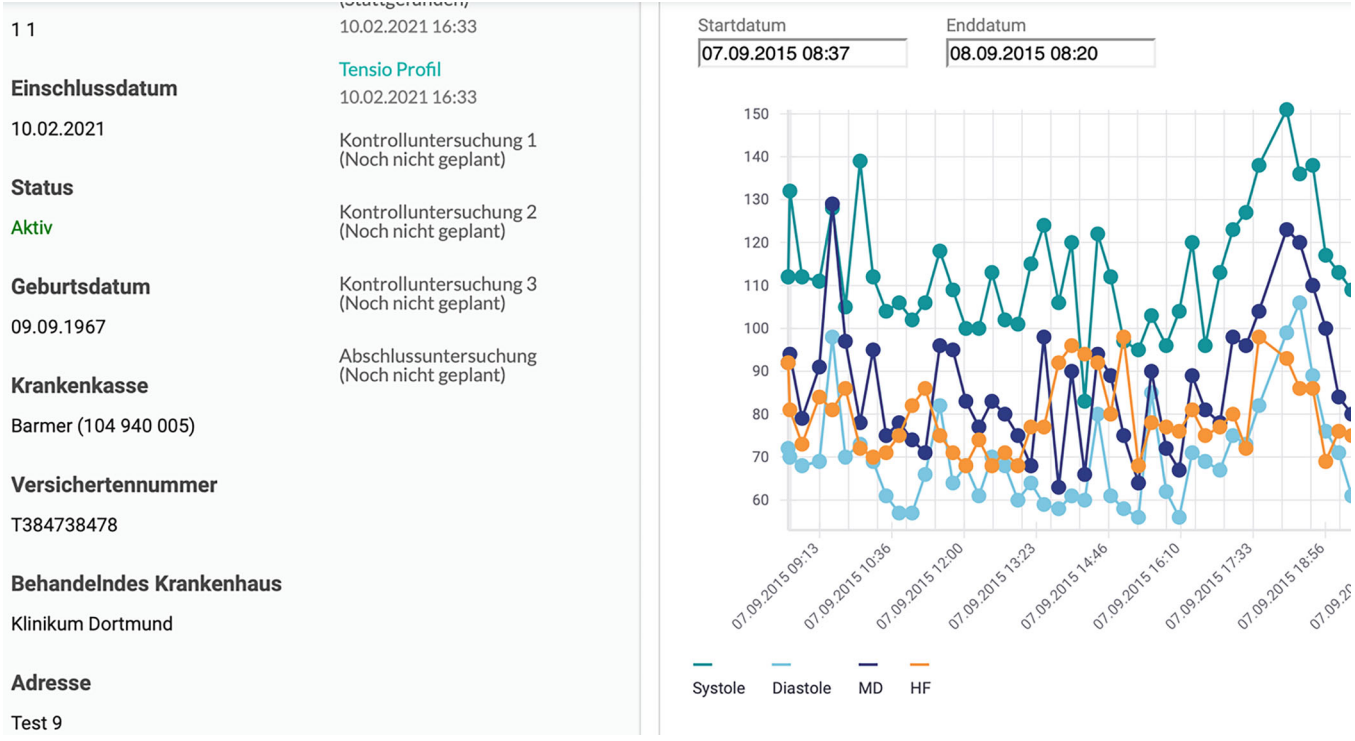


Fig. 18.2 Partial display of the SALUS "app" that allows different healthcare providers to enter and view data related to glaucoma treatment. The display here shows results of a 24 h blood pressure measurement as a reference for the possible variation in intraocular pressure

chronic diseases. These challenges demand a search for integrated care solutions that promote a longer stay of patients at home, improving their quality of life and optimizing the use of health resources. This section presents the results of the PICASO project,[2] a project partially funded by the European Union's Horizon 2020 research and innovation program under grant agreement No.: 689209. The PICASO solution supports a coordinated care of patients who have multimorbid chronic conditions (rheumatoid arthritis, together with a cardiovascular disease, and Parkinson, together with a cardiovascular disease) that involve different medical specialists and caregivers across multiple organizations. It enables the sharing of a patient's complete clinical pathway with tools to monitor the health status, predict risks, collaborate, and adjust care.

The platform and its tools have been tested at two pilot sites with patients in Germany and Italy.

The PICASO platform consists of three major cloud components:

- Multiple legacy care information systems (Hospital Information Systems, HIS), each operating as a private cloud or cloud-like internal business structure with strict access control and limited access rights including secure storage of patient data. Within PICASO, this is called the Care System Private Cloud. This includes components such as the traditional EHR within the hospital, together with different security and privacy components and an Operational Data Store (ODS) which translates the information between PICASO and the clinical EHR.
- The PICASO Integration Platform Public Cloud operating as a public cloud solution providing the central integration service platform such as management of secure data exchange between the multidisciplinary actors, secure data collections from patients' homes, and secure execution of care plan services. In this cloud, the patient data are pseudonymized.
- Multiple patient and environment monitoring systems running in the patients' homes. Each of them exposes cloud web services and is thus regarded as the Patient Private Cloud.

Figure 18.3 schematically shows the interaction between these components. All software for care management and decision support is hosted in the public cloud, while the patient data resides inside the care organizations.

A Distributed Validation Authority (DIVA) ensures that data is only shared if all policies and transaction-specific privacy and security requirements are met. DIVA combines an Access Manager, Identity Manager, and Policy Manager [10, 11], which ensure the following:

- A secure connection to the PICASO Public Cloud as well as communication between Public and Local Clouds.
- Identification and authentication of users.
- Verification of data transactions.

[2] https://www.picaso-project.eu/—last visited 2021-04-19.

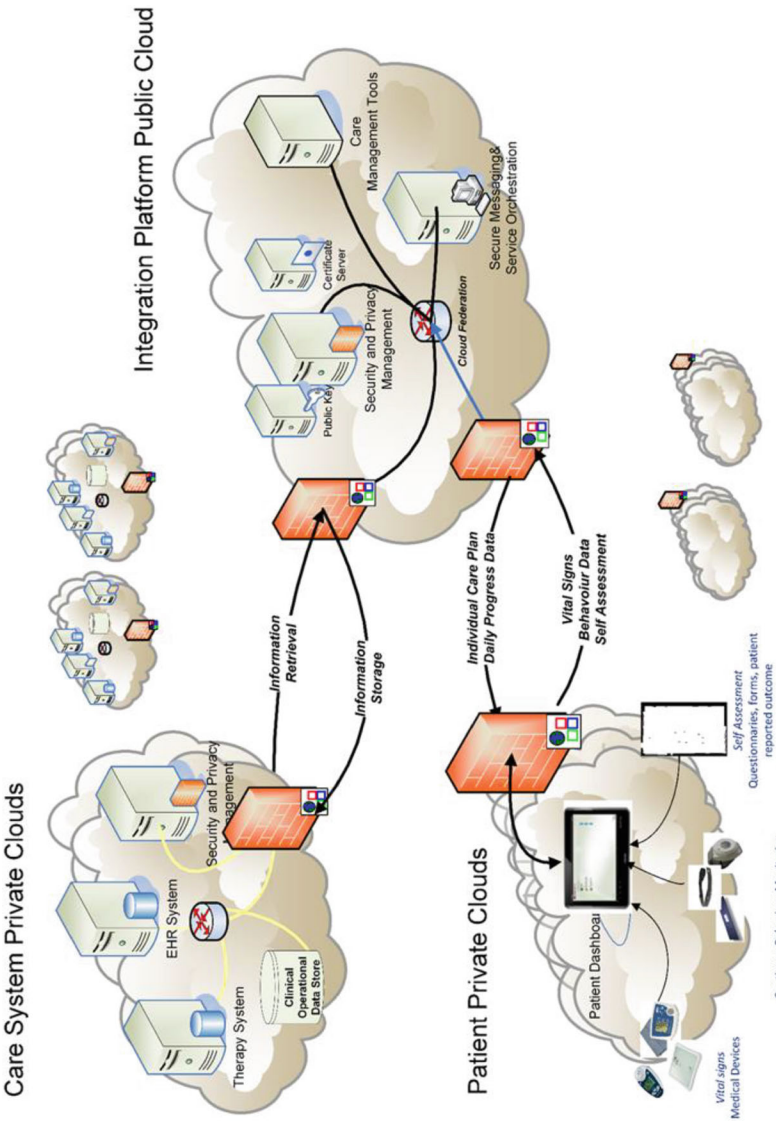


Fig. 18.3 PICASO platform overview [10] © 2021, Richter et al. [CC BY 4.0]. No changes made

The complete system was tested at the two pilot sites, with a positive feedback from patients and clinicians, especially from the perspective of improvement in the management of the clinical processes and patient empowerment. The system also demonstrated the integration of different clinical data spaces into a homogeneous environment.

PICASO developed the following applications.

18.4.1 Care Management as a Service

PICASO offered a complete solution for the management of care plans across different organizations, so that clinicians can follow up the different treatments of the patients in different institutions. That reduces the risk of polypharmacy and facilitates the coordination of tests, together with the stop/start of different medications (see Fig. 18.4).

18.4.2 Patient Self-Monitoring as a Service

The Patient Self-Monitoring Solution is supported by a patient dashboard and an app. Unlike existing solutions, this service integrates all three types of monitoring schemes:

- Scheduled vital signs measurements with medical devices.
- Continuous activity and behavior monitoring using wearable sensors.
- Patient self-assessment and medication recording (PROMs—Patient-reported outcome measures).

18.4.3 Risk Manager

The Risk Manager provides patient-specific risk estimates, merging standard charts with novel risk scores into one single tool. The novel risk scores use machine learning to merge data from an entire patient's history into a single integrated risk score.

18.4.4 Data Resource Browser and Patient Data Viewer

Clinicians can browse all available patient data in the Data Resource Browser, with an intuitive user interface. In the Patient Data Viewer, the clinician can look at patient data received by other care providers if authorized to do so and see what measurements have been performed, together with interventions and care plans (see Fig. 18.5).

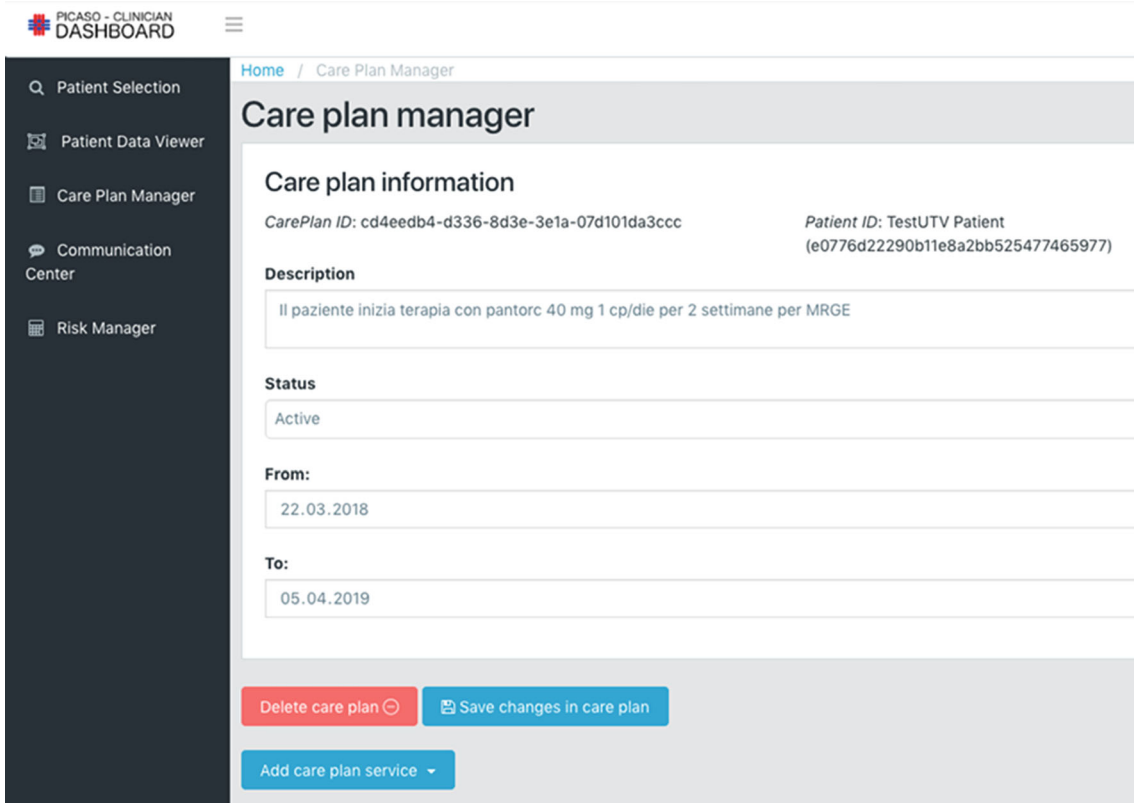


Fig. 18.4 PICASO Care Plan Manager entry page

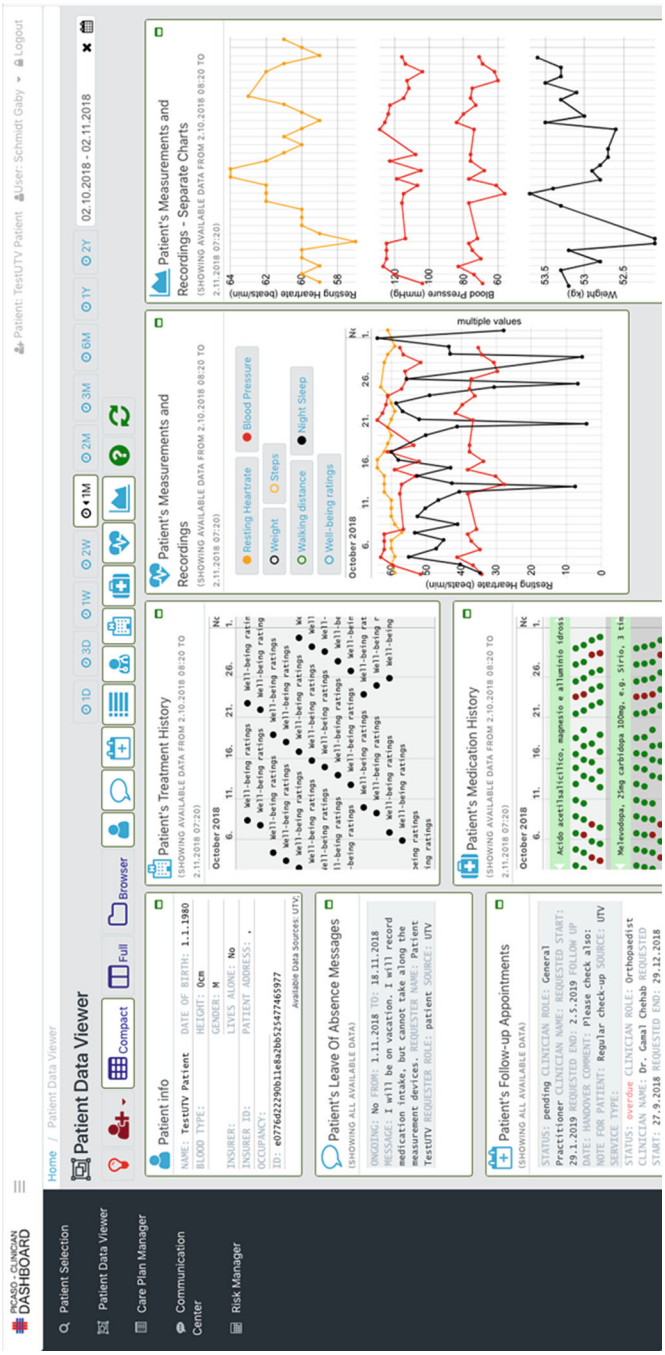


Fig. 18.5 PICASO Patient Data Viewer (dashboard view)

This example shows that it is important to support a hierarchy of apps. For example, the risk manager uses data collected by other apps. While the PICASO components were developed together, in a healthcare ecosystem, it should be possible for the risk manager to have a flexible interface for data collection so that it can interoperate with a variety of primary apps.

18.5 Precision Medicine

Genome sequencing technology has advanced in the recent years, so that whole genome sequencing can be routinely employed as a diagnostic mechanism (e.g., in rare diseases) and as a selection criterion for therapy optimization (e.g., in cancer). This technology has been introduced in many countries. In most of these countries, it has necessitated a new organization of healthcare provision due to the demands of the technology.

Genome sequencing is a high-throughput laboratory technique that produces huge datasets. The main challenge of data analysis is to decide which of the thousands of genomic variations of an individual are of diagnostic or therapeutic relevance. This knowledge increases rapidly and is tightly bound to data analysis within larger datasets. Data analysis employs many special tools at the research level.

To make this technology applicable for regular patient care, the sample processing and data analysis have been centralized in larger, research-oriented units. The actual patient care is also provided by specialized regional centers, which are distributed throughout the country to give all patients equal access. Because genomic data are very sensitive, they are typically kept in a specially secured environment, while the clinical data of the patient reside with the specialized regional centers. Furthermore, genomic data together with standardized clinical records can also be shared internationally when the patient permits.

In this situation, data are distributed between at least three different organizations. The clinical data are highly disease-specific and can also vary with healthcare providers, while the genomic data are more standardized for exchange (see Fig. 18.6).

The collaboration among the different (sub-)organizations is typically defined by regulatory procedures. Although these procedures vary between countries, the overall structure is pretty similar. This field, therefore, provides a sufficiently complex example of a medical data space. In some countries with centralized healthcare systems, IT has been centrally organized in this area. Other countries with a decentralized healthcare system still struggle with establishing the necessary infrastructure. Germany has recently started an initiative to introduce this model into

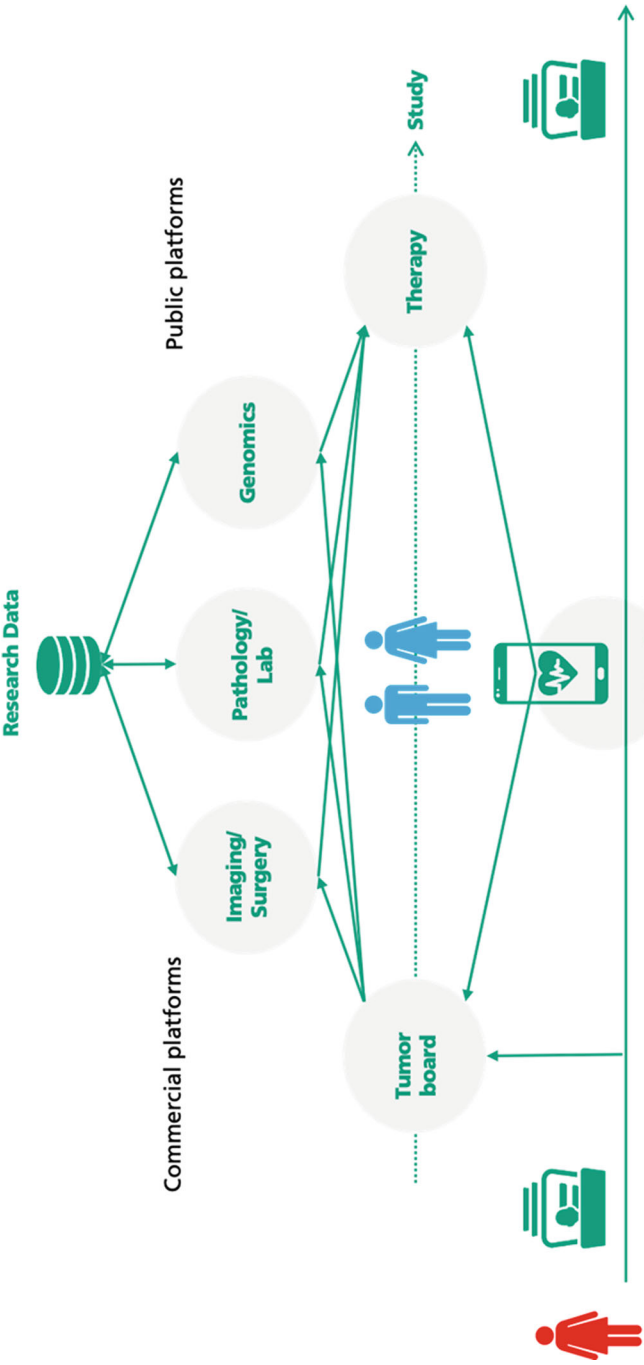


Fig. 18.6 Longitudinal record of a cancer patient with different aspects of the patient collected by different organizational units

the German healthcare system (genomDE[3]). It will be based on a federated and distributed data structure.

Because of the highly sensitive nature of genomic data, not all patients will agree to share their data. Therefore, techniques of remote algorithm execution are needed [12, 13] to enable researchers to comparatively and statistically analyze genomic variation in large cohorts.

18.6 Healthcare Data Ecosystems

In healthcare, the model of data providers connected to data consumers is too simple because most data are personal. The Personal Data Ecosystem Consortium (PDEC)[4] therefore introduces the "Principal" role, which in healthcare corresponds to the patient.

Quite often, patients are regarded as passive subjects. In a patient-centric healthcare system, patients have an active role, executing their sovereignty over their personal data maintained by other parties and also adding further input into the process via personal monitoring. A further role distinguished by the PDEC is the "Custodian." This role is important because few patients are actually data managers, so they might employ another party to help managing data and to coordinate with the other actors.

Therefore, a crucial element of a healthcare data space is the patient role. Patients may take actions on the usage of their personal data, but they are also the "end-customers" of healthcare services and active players on behalf of their own health (patient empowerment). The aspect of personal data coupled with an active participation of each person has not yet been addressed in detail in data space research.

Furthermore, the view of a data provider either as a repository (as in data management) or as a data stream (as in the Internet of Things) mostly ignores that healthcare data are typically generated as part of a healthcare service, such as an examination or an intervention. The data are a result of an interaction of patient, caregiver, medical devices, and IT services. Examples are:

- Visiting a general practitioner. The anamnesis data are entered into the local patient record.
- Medical imaging. Data are stored, e.g., in a radiology imaging system.
- Blood analysis in a lab. Data are sent to the general practitioner (on paper).
- Blood pressure or blood sugar measurement on a personal device. Data are collected via a mobile app.

[3] https://www.bundesgesundheitsministerium.de/themen/gesundheitswesen/personalisierte-medizin/genomde-de.html

[4] http://pde.cc

To avoid the confusion between healthcare services (which are legal and financial transactions as well as personal interactions) and IT services (which are operated to manage the former), we will use the term "app" to refer to an interactive IT service that interacts with a user (doctor/patient) or a medical device to manage and coordinate healthcare services. In that sense, data providers and data consumers in a healthcare data space on the one hand represent human or organizational actors, but on the other hand have corresponding "apps" that manage the digital transcript of the interaction.

In the digital ecosystem view, actors may not only be providers of traditional healthcare services but also information providers, disease management systems, study companions, or procurement services. Actors and the supporting apps are typically regulated with respect to safety and security, as well as reimbursement and documentation.

18.7 Structure of a Healthcare Data Space

The technical components of a healthcare data space are interactive information systems (owned by independent service providers) called "apps" that encapsulate functions and data. Functions may include data capture from users and technical devices, data analysis, data visualization, and remote interaction. Functions may also include execution of transactions such as appointments and referrals and (maybe bidirectional) data transfer to other apps. Apps support multiple users in different roles. These apps are distributed among different organizations and connected through an appropriate technical platform, ideally a modern, flexible, and trustworthy multi-cloud such as the Gaia-X architecture.[5]

Apps can support processes like

(A) Intra-organizational care, e.g., in a hospital using a hospital information system (HIS).
(B) Inter-sectoral collaborations, e.g., outpatient management and home care.
(C) Personal monitoring and prevention, e.g., fitness tracking or diabetes monitoring.

Another reason to use the "app" terminology is that we want to incorporate mobile technology from the beginning. While category C is already built on a consumer perspective through wearables and mobile devices, B and A still mostly face the transformation task to a modern digital platform.

There are numerous examples of such digital apps. However, most of them are specialized for particular diseases, interventions, or risks. Many of them aggregate data relevant for the decision process in a certain context, e.g., a disease-specific guideline. Apps have to be specialized to be useful, effective, and user-friendly.

[5] https://gaia-x.eu/; https://cloud.ionos.de/gaia-x

For example, in diabetes management it is not sufficient to just collect data (glucose level, weight, etc.) and display them. Most diabetes management services include coaching and social interaction mechanisms with different roles (e.g., doctor/coach/other participants). Apps that include medical devices or are classified as medical products themselves will also include more advanced data analysis functionalities.

While apps can tailor therapies for the individual patient based on the data collected, apps can also provide data for medical research and healthcare system optimization. This is particularly relevant both for chronic and rare diseases. Effectively, the healthcare system will learn from as many examples as possible.

While early smartphone apps stored data on the phone itself, most apps now use some form of cloud storage to provide a backup and to allow cross-device usability. While medical data protection in typical cloud implementations is seen as problematic, just storing the data on the device is no alternative in terms of backup and sharing. Therefore, healthcare apps should use an appropriately certified cloud service or should not store the data themselves at all, but directly acquire them from another certified app or medical record service.

18.8 User-Centered Concepts in a Healthcare Data Space

An important feature of a healthcare data ecosystem is to permit the patient to exercise sovereignty over their personal data. The difficulty with this goal is that we typically cannot expect (elderly) patients to be expert data managers, in particular in situations when they are in direct need of healthcare. Internet privacy is plagued by similar issues, e.g., when consenting to the use of cookies.

Data spaces are typically described from the perspective of data owners. Each institution is owner of the data it maintains and could provide to others. In digital health, the situation is a little bit more complicated. The concept of ownership is not directly applicable, as both patient and healthcare institutions have different legal rights and obligations on the same data. Therefore, patients have dual roles as legal entities and as additional users with a consumer perspective distinguished from the provider perspective of the healthcare providers.

18.8.1 Trusted Users

All users (patients as well as healthcare professionals) need a trusted identity in a healthcare ecosystem. This ensures that data can only be exchanged with trusted participants. In the projects described above, such a service is specifically set up for each project. A national or regional healthcare system, however, needs to establish such a system. It should cover an extensive range of users including home care personnel, medical supply stores, or even relatives as trustees or guardians. Fortunately, technical interfaces for identity management are mostly standardized, as they

are also used in other business areas. All apps in a healthcare data ecosystem should be prepared to employ such a system-wide identity service.

Trust is established by a service provider certified via IDSA-compatible processes that lets individual users register and validates their identity. For example, an insurance company could be the identity provider for its customers, while associations of medical professionals could do the same for their members. Authentication preferably uses mobile devices (plus other factors) for ease of use.

18.8.2 Single Entry Point

For ease of use and convenience, every user should have a single entry point to the ecosystem. For patients, this is called the *health account*, and it provides

- A directory of apps that the person subscribes to/has accounts with. Accessing the app will execute a login action to this app or show the dashboard (e.g., if I am a trusted relative).
- A directory of consents given to each app and to each data exchange channel between apps. Each consent should have a legal document, and explanatory document, a date of signature, and a way of withdrawing the consent. Preferably, electronic consent should be promoted.
- A transaction log of any data exchange between apps, being able to inspect which data were exchanged when.

Informed consent is an important element in data protection. In a data space, consent is primarily given at the level of apps (instead at the level of institutions and/or data elements). The main advantage is that apps themselves can establish and control fairly detailed rules on who can access which data in what role. For example, if I fill the general practitioner (GP) role in an app with my preferred GP, the access rights follow from that role. In an app about mental disorders, a psychiatrist would be authorized to look into my detailed protocols, while the GP role would only see a brief summary (if the GP role is filled at all).

Of course, this means that apps have to be trusted to implement access rights correctly. Given that medical apps will be regularly scrutinized for various properties in the future, in particular for safety and security, this should pose no additional problems. While this is a relatively coarse granularity of user influence, it is much more user-friendly than highly granular access rights, as the latter have a high error potential even for experienced users.

In a data space, in addition to influencing who will fill each role, patients can give permission for one app to exchange standardized data with another. A typical example would be a medication plan. While a medication plan could be maintained centrally, an app would have to manage it anyway unless the central service is mandatory for all patients. Managing the medication plan inside the app also gives the possibility to extend the data stored useful for a particular use case, while only the exchange would stay at the standard definition.

Therefore, we have defined exchange permissions as a central task in the patient's health account.

18.8.3 Unified Health Report (aka Virtual Health Record)

One of the usage scenarios of a common EHR is that actors not originally involved in the treatment of the patient should be able to get a brief overview of patient status and history as relevant to health-related use cases. For example, in ophthalmology, people in care of an elderly person should be aware of a glaucoma diagnosis, any medication, and the need to regularly visit an ophthalmologist. Such a summary could be provided in a common EHR, but it can also be provided as a health dashboard of the glaucoma app.

The EHR solution has the advantage that the data by the different apps could be analyzed by further tools (maybe using artificial intelligence) to detect interactions. It requires the summaries to be fairly standardized, though. Just visualizing the information in the health dashboard makes the apps more independent of each other and is fairly easy to implement by each app. The medical data space provides both options.

There is a standard interface to call up a health dashboard from all the roles that are authorized by the patient (see Figs. 18.3 and 18.7). This may particularly apply to care personnel and relatives, which are easily enabled to participate in the care process, even remotely. The advantage is that in the dashboard, the visualization can be optimized towards general information demands for specific diseases, distinguished by the levels of relatives, professional caregivers, and general practitioners. Again, no specific permissions have to be given by the patient except assigning a person to the particular role.

18.9 Data Quality in the Medical Data Space

A common problem with EHRs is reliability of data generated by a different organization or even by the patient. With a health record regarded as just a set of data, even secure audit trails provide only limited context about how the data have been generated. With health records maintained by apps, it is one of the main purposes of the app functionality to encourage or guarantee a quality management process. The quality of the data depends on appropriate measures being implemented by the app. Data quality in this scenario is highly dependent on the quality of the app.

For example, in SALUS there are provisions to make sure that data are correctly handled by the patients. Furthermore, a review process tries to detect suspicious data. The data quality of the SALUS data is not just provided by the database, but by the treatment process organized around the SALUS app. The SALUS system allows ophthalmologists to review measurements in a context of the source devices, subsequent examinations, and general consistency as regards the disease. The whole

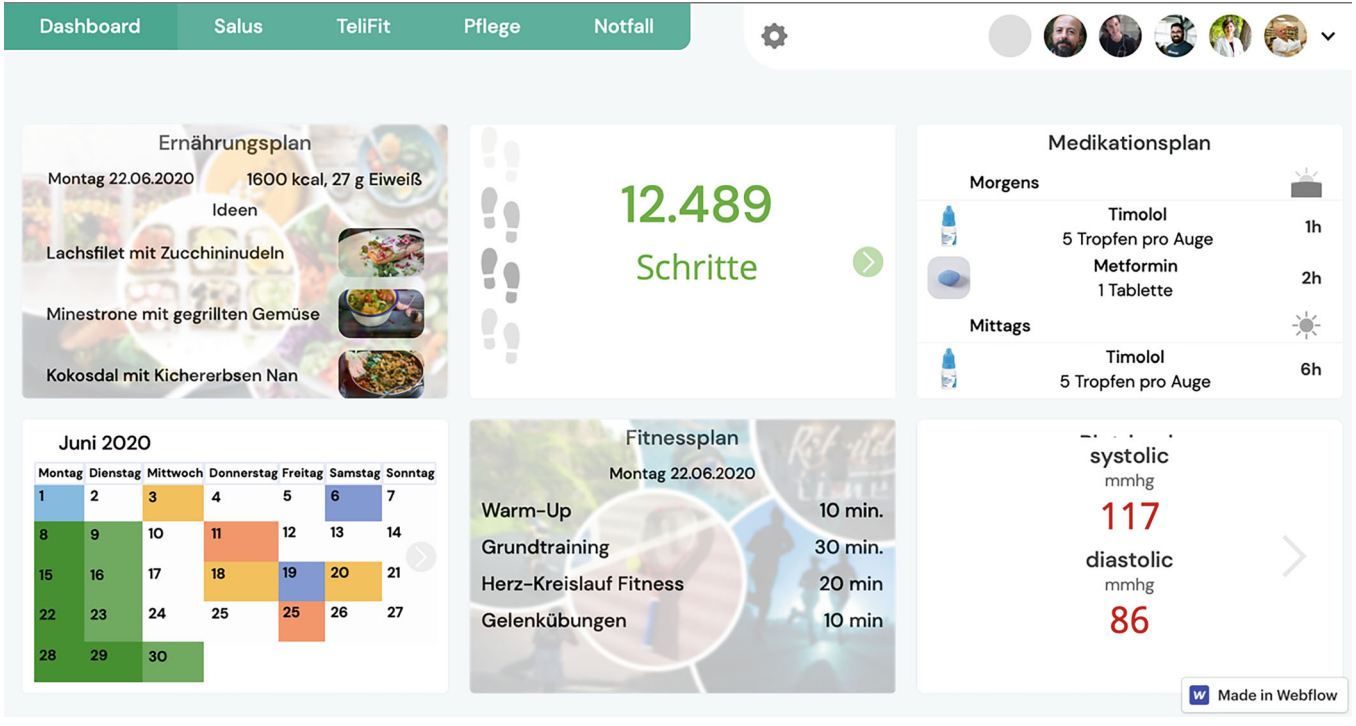


Fig. 18.7 Draft design of a patient dashboard that is composed from individual displays from several different apps

process determines the final quality of the data. Therefore, an app that stimulates a quality-driven process establishes trustworthiness of the data for other uses. In SALUS, data may be uploaded by the patient, but they are reviewed by an ophthalmologist. Consequently, the data and in particular the data summary can be reliably used by other specialties or by nurses and caregivers.

For medical research, a hierarchical aggregation process of such data is necessary. Quality-controlled data exported from an app can be converted into a structure suitable for big data analytics and combined with other data sources. This aggregation can be performed using ad hoc techniques, but assuming that apps like SALUS will be much more common in the future, it will be easily possible to establish permanent pipelines of data using IDSA broker services as aggregators. User consent will be necessary for such studies, but with a transparent mechanism, it is much easier for users to give (and redraw) such consent. For the research process, an important aspect are the anonymization and pseudonymization of data, which will allow the aggregation of data coming from different patients as input to different machine learning algorithms and data analytics processes facilitating new research and diagnostics tools for clinicians, as in the PICASO project.

18.10 Conclusion

The architecture presented here is a promising approach to handle future demands for a health data ecosystem. However, just like the whole ecosystem, the architecture will have to evolve as well. Currently, we are trying to validate the approach in many different areas from prevention and healthcare to medical research and epidemiology. A particular emphasis will be on the new role of the patient. Further studies have to be conducted that find out what is a most appropriate level of control and how we can simplify the interaction and still inspire trust.

References

1. Berlage, T., Claussen, C., Hofmann-Apitius, M., Fischer, R., & Jarke, M. (2016). *Medical data space white paper*. Fraunhofer-Gesellschaft.
2. HL7. (2021). *HL7/FHIR® Resources*. https://www.hl7.org/fhir/resourcelist.html
3. IDSA (2019, April). International Data Spaces Association. *Reference architecture model*, Version 3.0.
4. Gelhaar, J., & Otto, B. (2020). *Challenges in the emergence of data ecosystems*. In Proceedings of the 23rd Pacific Asia Conference on Information Systems (PACIS'20).
5. Oliveira, M. I. S., & Lóscio, B. F. (2018, May). *What is a data ecosystem?* In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age (pp. 1–9).
6. Geisler, S., Vidal, M. E., Cappiello, C., Lóscio, B. F., Gal, A., Jarke, M., et al. (2021). Knowledge-driven data ecosystems toward data transparency. *ACM Journal of Data and Information Quality (JDIQ), 14*(1), 1–12.

7. Gelhaar, J., Groß, T., & Otto, B. (2021). *A taxonomy for data ecosystems*. In Proceedings of the 54th Hawaii International Conference on System Sciences (p. 6113).
8. Gröne, O., & Garcia-Barbero, M. (2001). Integrated care: A position paper of the WHO European Office for Integrated Health Care Services. *International Journal of Integrated Care, 1*, e21.
9. European Commission. (2020). *European Commission report on the impact of demographic change*. Accessed April 19, 2021, from https://ec.europa.eu/info/sites/info/files/demography_report_2020_n.pdf
10. Richter, J. G., Chehab, G., Schwartz, C., Ricken, E., Tomczak, M., Acar, H., Gappa, H., Velasco, C. A., Rosengren, P., Povilionis, A., Schneider, M., & Thestrup, J. (2021). The PICASO cloud platform for improved holistic care in rheumatoid arthritis treatment—Experiences of patients and clinicians. *Arthritis Research & Therapy, 23*, 151. https://doi.org/10.1186/s13075-021-02526-7
11. Povilionis, A., et al., (2018). *Identity management, access control and privacy in integrated care platforms: The PICASO project*. 2018 International Carnahan Conference on Security Technology (ICCST) (pp. 1–5), Montreal, QC, Canada. https://doi.org/10.1109/CCST.2018.8585716.
12. Beyan, O., Choudhury, A., van Soest, J., Kohlbacher, O., Zimmermann, L., Stenzhorn, H., . . . Dekker, A. (2020). Distributed analytics on sensitive medical data: The personal health train. *Data Intelligence, 2*(1–2), 96–107.
13. UK Health Data Research Alliance. (2020). *UK Health Data Research Alliance, Trusted Research Environments (TRE): A strategy to build public trust and meet changing health data science needs*. Green Paper v2.0 dated 21 July 2020. https://ukhealthdata.org

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 19
Industrial Data Spaces



Thomas Usländer and Andreas Teuscher

Abstract This chapter describes the application of the IDS principles, architectural artifacts, and technologies to the application domain of industrial production and smart manufacturing, in particular as drafted by the Platform Industrie 4.0 in their Reference Architecture Model Industrie 4.0 (RAMI4.0) and follow-on specifications about the Asset Administration Shell (AAS). It elaborates on the working approach of the IDS-Industrial Community (IDS-I) for the analysis of requirements on data sovereignty. This activity is motivated by the vision 2030 of the Platform Industrie 4.0 that states autonomy, including data sovereignty, as one strategic field of action.

The chapter presents how IDS-I aims at systematically deriving and analyzing data sovereignty aspects from the two reference use cases, Collaborative Condition Monitoring (CCM) and Smart Factory Web (SFW), in order to identify architectural and technological synergies and gaps between the International Data Spaces (IDS) and the specifications of the Platform Industrie 4.0.

19.1 Motivation for Industrial Data Spaces

Digitalization leads to the creation and usage of data. In the context of this social, economic, and technical development, data has become an independent product, also in the domain of industrial production and smart manufacturing. As an economic good, they form the basis for new value-added processes and business models. In daily business practices, data are used very often, however, exchanged rather rarely. Companies are still too worried about losing control over their data—and thus their valuable corporate knowledge. This is where International Data Spaces (IDS) come

T. Usländer (✉)
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Karlsruhe, Germany
e-mail: thomas.uslaender@iosb.fraunhofer.de

A. Teuscher
SICK AG, Waldkirch, Germany
e-mail: andreas.teuscher@sick.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_19

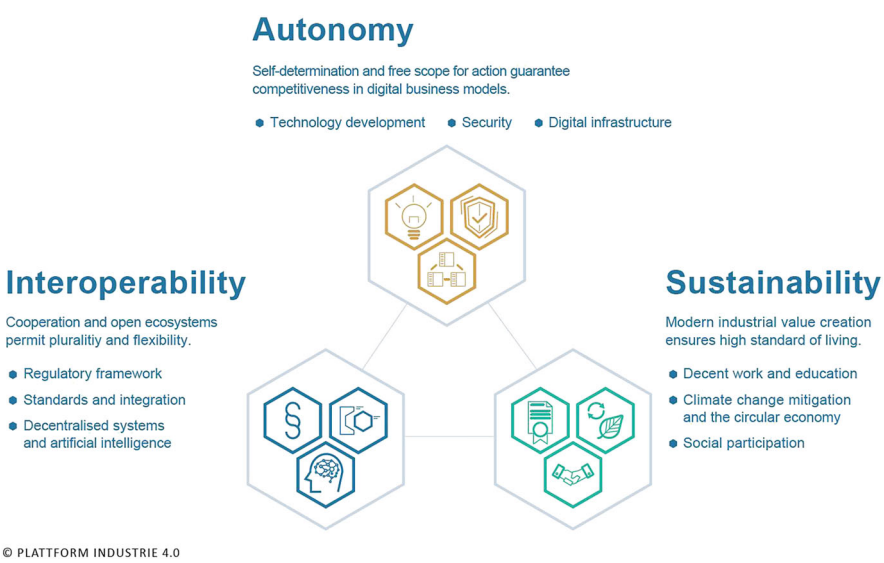


Fig. 19.1 Strategic fields in the Vision 2030 of the Platform Industrie 4.0 [1]

into play: with an architecture for virtual data spaces that guarantees the secure and sovereign exchange of their data.

The overriding objective of the IDS is to help companies and institutions to take advantage of the benefits of digitalization without increasing their risks. The means for this is a trustworthy architecture for data management with standards for sovereignty and secure data exchange.

The International Data Spaces Association (IDSA) represents the interests of more than 120 international companies and institutions. The IDSA bundles the requirements for all IDS application domains, organizes the exchange of knowledge between research and industry, and develops guidelines for the certification, standardization, and utilization of the results resulting from the various IDSA-related research projects at European and national level.

In its vision for 2030, published in August 2019, the Platform Industrie 4.0 formulated a holistic approach to the shaping of digital ecosystems and re-orientated the further development of Industrie 4.0 according to this vision [1]. At the heart of the design of digital ecosystems are the three strategic fields of action autonomy, interoperability, and sustainability (see Fig. 19.1).

For data sovereignty the strategic field of autonomy is highly relevant:

- Autonomy is the freedom to take independent decisions and to interact in conditions of fair competition—from a chosen business model to an individual's decision to make a purchase.
- Autonomy requires an open digital infrastructure for everyone, data protection, IT and information security, and technology-neutral research, development, and innovation.

The freedom to take independent decisions and the request for fair conditions are key characteristics of the demand for digital sovereignty—on the technological and infrastructural level as well as on the data level. The European initiative GAIA-X [2] focuses on the former level and aims at creating a secure, federated infrastructure that meets the highest standards of digital sovereignty while promoting innovation. The data level, however, is the key concern of the IDS.

Within IDSA, the application domain of networked industrial production and smart manufacturing, hence Industrie 4.0, is represented by a dedicated community entitled IDS-Industrial, abbreviated by IDS-I. This chapter describes how IDS-I handles the problem of analyzing the requirements on data sovereignty for Industrie 4.0.

As a consequence, the mission of the IDS-Industrial Community is summarized as follows:

1. To gather requirements on data sovereignty incl. Data sharing, data usage monitoring and control, as well as data provenance tracking by means of reference use case specifications.
2. To map these requirements systematically to the standards, capabilities, and recommended technologies of the International Data Spaces Association and the Platform Industrie 4.0.
3. To derive profiles of IDS/Industrie 4.0 specifications that support the requirements in industrial business ecosystems based upon standards and by means of common governance models.
4. To validate and demonstrate the applicability of these specifications by means of reference testbeds, e.g., Smart Factory Web and GAIA-X use cases.
5. To contribute to the outreach of the IDS architecture and specifications to the community of industrial production and smart manufacturing.

In this chapter, it is described how the individual statements of this mission are carried out.

19.2 Industrial Perspective

Establishing globally accepted framework conditions for possession, ownership, transfer, and usage of data is a basic prerequisite for the realization of an efficient digital ecosystem.

In order to meet the requirements for the networking of industrial equipment, machinery, and sensors with cloud/data services in the Internet, going hand in hand with digitalization, “smart machinery” or “cyber-physical systems (CPS)” must become part of the Internet of Things (IoT). The production environment, in particular, is becoming increasingly digital and intelligent. Especially automation, mobility, flexibility, and individuality are pushed. Digitalization is thus becoming the key to making things in the cyberspace addressable, visible, and, in this way, usable, e.g., for applications such as condition-based monitoring, predictive

analytics, and maintenance. It increases efficiency as well as attractiveness of a location for business development and competitiveness.

From the industrial perspective, there are the following central requirements:

1. **Identification of users without doubt when they access equipment, machinery, and sensor data.**

   In the real world, this problem is addressed with documentary proof such as an identity card or an employee badge. In the digital world, a one-to-one digital identity, which is accepted beyond a security domain, is necessary. Examples of this exist in telecommunications technology, which allows a mobile phone user to be identified in the telecommunications network via a subscriber identity module. Subsequently, the mobile communications provider provides data connections to the user.

2. **Upon successful identification, the transfer of data must be safeguarded against manipulation before the data flow starts.**

   In view of the increasing data volumes, the transfer must be at the same time secure and reliable as well as efficient (high throughput) and scalable. Ideally, as these objectives may be contradictory, the users themselves should be able to define the required level of the technical security. Thus, it would be possible to select the security mechanisms to the match the data protection class.

3. **Usage characteristics shall be assigned to the data itself, further-on communicated to the recipient, and, in an ideal case, enforced and monitored.**

   As a consequence, data could be linked to a dedicated contract associated with a data usage control policy, e.g., to delete the data elements after a certain period of use or number of accesses in order to prevent any further use. However, this would require the data user to receive, understand, and confirm the terms of use. It lies in the nature of digital data that reproduction without any technical problems and quality loss results in a transfer of possession that is undesirable to the owner. This can only be prevented and controlled by clear rules and their technical implementation. These measures must be transparent and trustworthy for all economic actors. To this end, it is necessary to create independent institutions and technologies ensuring compliance with the socially necessary and accepted rules.

The IDS-Industrial Community has come to entertain the view that an individual company that centrally controls and manages data and its possession and/or the transfer of data ownership cannot fulfill the conditions of a free market economy, because such an approach would result in a global monopoly on data.

The current practice of signing bilateral contracts and bilateral technical agreements to lay the foundation for secure data sharing generates high costs for establishing and maintaining a contract and, consequently, high transaction costs, and it would hinder further digitalization.

If we aim at leveraging a diversified, competition-driven global economy, we need to standardize in terms of laws and regulations and technically implement a data ecosystem for the transfer of ownership and the associated transfer of possession.

The IDS-Industrial Community is convinced that the International Data Spaces promotes digital sovereignty of data owners in the industrial data economy and thus provide the basis for an open, generally accepted technical procedure for all data transactions. The development of such a trust level with the label "Made in EU" could not only significantly promote global data sharing and thus the growth of future data-based business models but in general support easy value creation as an enabler for the economy.

19.3 Requirements Analysis in the IIoT

The question of how to handle requirements on data sovereignty in joint Industrie 4.0/IDS and GAIA-X service-oriented environments falls into the general problem of Agile Service Engineering in the Industrial Internet of Things (IIoT) [3].

As illustrated in Fig. 19.2, an agile approach is recommended to reduce the conceptual and terminological gap between the views of the thematic expert (typically an industrial, mechanical, and/or an electrical engineer) and the IT expert (typically, a computer scientist). Driven by the business strategy, the thematic expert expresses his/her functional and non-functional requirements about the system's behavior and characteristics, whereas the IT expert "answers" in terms of (mostly technical) system capabilities and service registries. Usually, both descriptions cannot be matched without additional, tedious discussions and additional explanations.

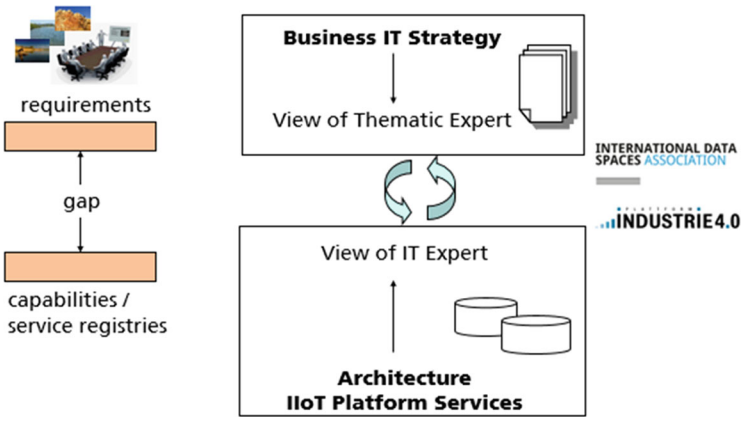


Fig. 19.2 Mapping of requirements in IoT platform environments

ID	<<will be defined later>>
Name	Name of the use case
Priority	[Low, medium, high]
Reference use case	[Smart Factory Web, Collaborative Condition Monitoring or other...]
Description	Textual description of the use case: • motivation • involved stakeholders • objective • constraints • ...
Comment	Optional further comments
Preconditions	what is required before the use case may be started or deployed
Workflow	The following steps are required to perform the use case: 1. ... 2. ... 3. ... ... Note: may have loops and jumps (if ... then go to step X)
Postconditions	Describe the situation after the use case was carried out
Sources	Literature or references
Authors	Name of the authors
Date	Date of last change

Fig. 19.3 Use case template as used in the IDS-Industrial Community

The idea of the SERVUS methodology [3] is to use semi-structured descriptions of use cases for this activity, following semi-structured use case templates [4] (see Fig. 19.3). With SERVUS, this idea of a semi-structured description of analysis and design artifacts applies, too, when mapping the use cases step-by-step to other design artifacts such as requirements and when matching them with abstract, technology-independent capability descriptions of IIoT platforms.

The terminological gap is approached by attaching semantic annotation labels to the basic terms used. The IIoT platforms of interest are those specified by the Platform Industrie 4.0, IDS, and GAIA-X. In order to master this stepwise mapping

process, all the analysis and design artifacts are stored in an IIoT Platform Engineering Information System (IIoT-PEIS), which also provides documentation, information retrieval, and visualization support [3].

19.4 IDS-I Reference Use Cases

The IDS-I Community decided to describe use cases stemming from two so-called reference use case domains (Fig. 19.4), also used by other initiatives:

- Collaborative Condition Monitoring (CCM), provided by the Platform Industrie 4.0 applied as GAIA-X use case.
- Smart Factory Web (SFW), an accepted IIC Testbed of the Industrial Internet Consortium (IIC).

19.4.1 Collaborative Condition Monitoring (CCM)

The CCM reference use case deals with the collection and use of operating data to optimize the reliability and service life of machines and their components during operation [5]. In the real world, installed machines come from different machine suppliers that are equipped with different products from different component suppliers. In these multi-stakeholder environments, the exploitation of the data of components, machines, and the factory plant to provide higher-level services such as predictive maintenance is still a challenge.

This is due not only to the lack of interoperability, which may be solved by encapsulating the assets by the concept of the Asset Administration Shell (AAS), but also due to the uncertainty of how to control the access and usage of the datasets

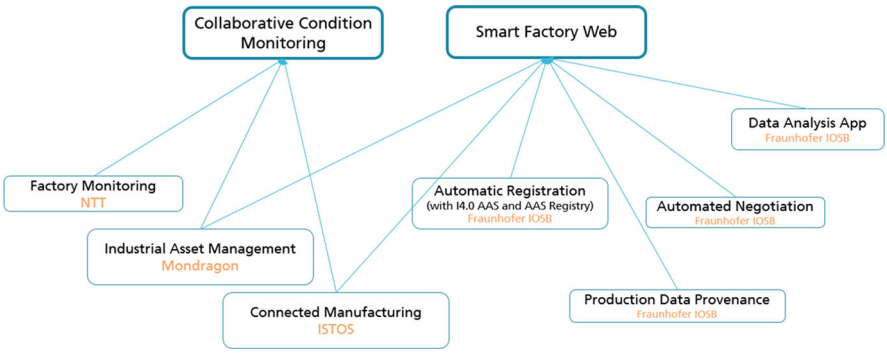


Fig. 19.4 Usage control as an extension to access control. ©2020, International Data Spaces Association (IDSA)

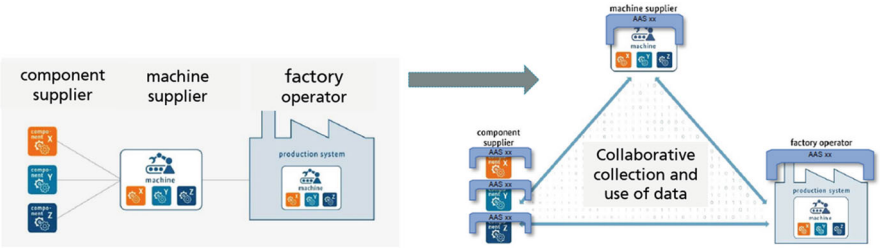


Fig. 19.5 CCM and SFW use cases under investigation in IDS-I

associated with and provided by the assets. IDS-I aims at investigating the detailed requirements and concerns (Fig. 19.5).

19.4.2 Smart Factory Web (SFW)

The SFW reference use case provides a blueprint architecture for open sustainable and resilient production ecosystems [6]. One important SFW application is an industrial marketplace for industrial production following the platform-driven economy of other branches such as tourism or mobility. As illustrated in Fig. 19.6, the demands of higher resilience, sustainable production, more flexibility, higher product variance, manufacturing on demand, and smaller lot sizes do not only address the factory level, e.g., the shop floor environment, but also the supply chain level, the so-called connected world of the Reference Architecture Model Industrie 4.0 (RAMI4.0).

A marketplace is highly needed that does not impose business dependency constraints upon the suppliers, but is designed on the principles of openness,



Fig. 19.6 Problem illustration of the Collaborative Condition Monitoring use case [5]. ©2020, Platform Industrie 4.0

fairness, and transparency. It allows a user to quickly search for new and alternate suppliers in a supply chain network. More flexibility is demanded in case a given supply chain is at risk or about to fail due to broken transport lines, natural catastrophes, pandemics, or material shortage.

In order to enable searching for alternate suppliers and matchmaking by the marketplace, adequate data about the capabilities and assets of factories in the supply chain is required. IDS-I aims at investigating the detailed requirements and concerns about sharing this type of data in a marketplace and within the supply chain network.

19.5 Requirements Analysis for Data Sovereignty

When considering and analyzing the requirements for data sovereignty in case of scenarios spanned by these two reference use cases, one has to distinguish between the classical aspects of access control (to data and operations) and data usage control.

Usage control is an extension to traditional access control [7]. After access to data and operations has been permitted, the question remains what happens to the data after the access and delivery of the data (as part of operation results). Hence, usage control is about the specification and enforcement of restrictions regulating what must (not) happen to data. Usage control is concerned with requirements that pertain to data processing (obligations), rather than data access (provisions) as illustrated in Fig. 19.7. In general, usage control is relevant in the context of intellectual property protection, compliance with regulations, and digital rights management.

As IDS-I aims to ensure data sovereignty in industrial value chains, requirements on data usage control are analyzed according to a common scheme, following the list of obligations proposed by [7]:

- **Secrecy:** Classified data must not be forwarded to nodes which do not have the respective clearance.
- **Integrity:** Critical data must not be modified by untrusted nodes as otherwise their integrity cannot be guaranteed anymore.
- **Time to live**: A prerequisite for persisting data is that it must be deleted from storage after a given period of time.



Fig. 19.7 Problem illustration of the Smart Factory Web use case

- **Anonymization by aggregation**: Personal data must only be used as aggregates by untrusted parties. A sufficient number of distinct records must be aggregated in order to prevent de-anonymization of individual records.
- **Anonymization by replacement**: Data which allows a personal identification must be replaced by an adequate substitute in order to guarantee that individuals cannot be de-anonymized based on the data.
- **Separation of duty:** Two data sets from competitive entities (e.g., two automotive OEMs) must never be aggregated or processed by the same service.
- **Usage scope:** Data may only serve as input for data pipes within the connector, but must never leave the connector to an external endpoint.

The IDS-Industrial Community has started a process to analyze the two reference use cases in more detail by means of use case descriptions that fall into these categories. For each of the use cases, requirements on access control (both role-based and attribute-based), usage control (according to the obligation categories described above), and data provenance tracking (where does the data come from) are gathered and analyzed.

19.6 Major Concepts of the International Data Spaces (IDS)

In order to understand the use case analysis below, the major architectural concepts of the International Data Spaces (IDS) are described [8] in the following and illustrated in Fig. 19.8.

- **Data spaces** unlock the value of data.
- An **IDS Connector** is a dedicated software component that allows participants to attach usage policies to their data in a data space, enforce the usage policies, and seamlessly track the provenance of received data. Hence, an IDS Connector acts as a kind of gateway for data and services. Furthermore, it provides a trusted environment for the execution of apps.
- **Data owner and data provider**: The data provider is a device that transfers the owner's data to the data space via the IDS Connector. It allows others to use the data while retaining control over the who, how, when, why, and at what price.
- **Data user and data consumer**: The data consumer is a device that processes data on behalf of the user. The data is offered by data providers by their usage policies and with confidence in the data quality and reliability.
- An **Identity Provider** creates, maintains, manages, and validates identity information of and for participants in the data space such as data providers and data consumers.
- **App Stores** provide software applications that can be deployed in IDS Connectors.
- **Apps** may be downloaded from the App Store into the trusted environment of the IDS Connector. Apps perform tasks such as transformation, aggregations, or analytics of data.

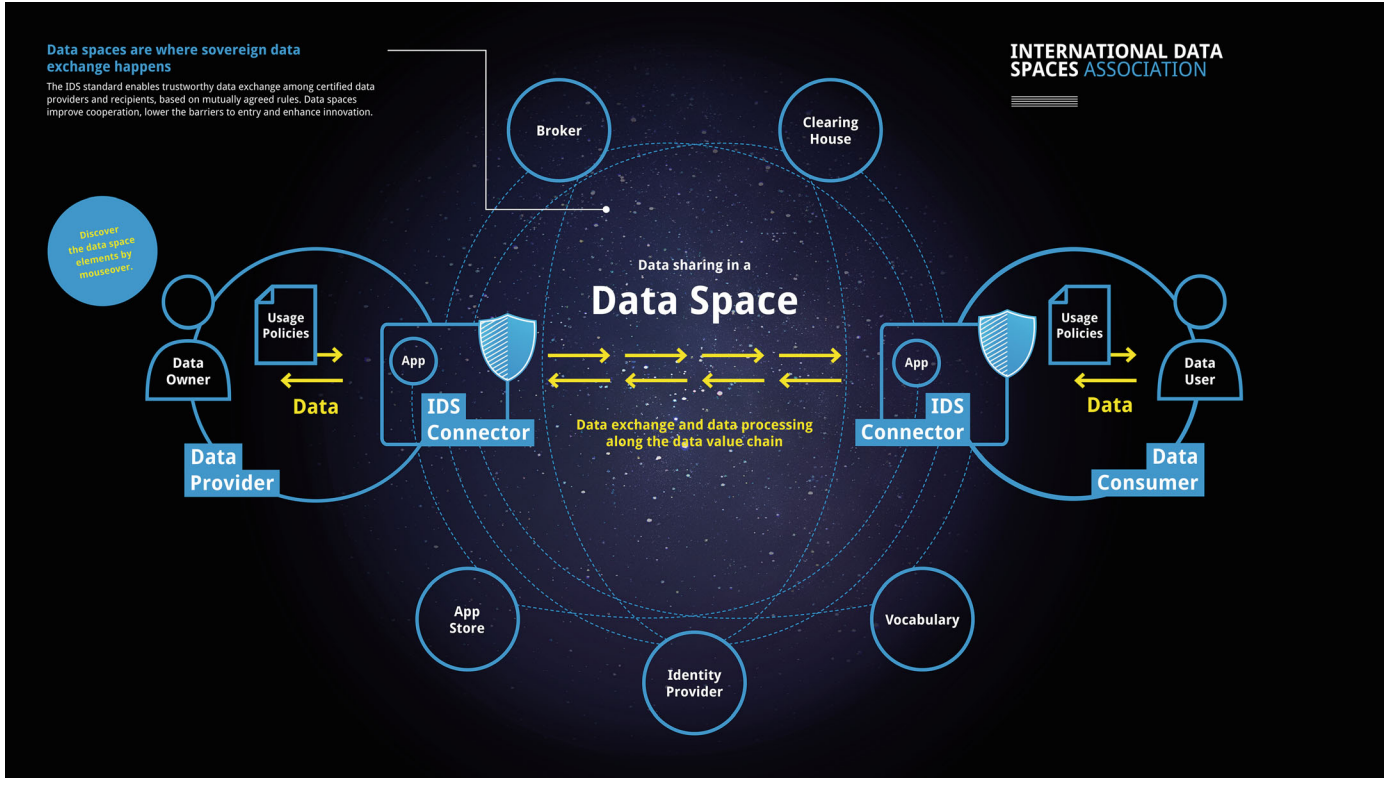


Fig. 19.8 Basic Architectural Concepts of the International Data Spaces (IDS). ©2021 International Data Spaces Association (IDSA)

- A **Broker** provides information about data sources in terms of content, structure quality, currency, and other features.
- A **Clearing House** provides clearing and settlement services for all data exchanges and financial transactions within a data space.
- **Vocabularies** provide standardized descriptors for data based on accepted best practices.

19.7 Exemplary Use Case Analysis

As an example for a detailed use case analysis, we take the "Automated Negotiation" use case in the context of the Smart Factory Web.

Name	Automated Negotiation
Priority	Medium
Referenceuse case	Smart Factory Web (SFW)
Description	**Motivation** The Smart Factory Web enables the design of an open matchmaking platform for industrial production to find new business partners and negotiate with them (Manufacturing as a Service—Maas). In most industries today, this negotiation is carried out by manual agents with telephone and e-mail. During the negotiation, sensitive information such as price, availability, capacity, and process durations are revealed. There is a need for a data infrastructure to let automatic negotiation agents handle the negotiation without revealing any information to potential partners. Negotiation should take place in isolated data containers (e.g., connectors) that interact with each, but cannot leak any information elsewise. After negotiation, the successful terms of a contract are presented, but dynamic variables about the production shall remain hidden. The companies then have the opportunity to sign the contract proposed by the negotiation. **Involved Stakeholders** Smart Factory Web portal, Company A, Company B **Objective** Hide sensitive data during negotiations **Constraints** The negotiation apps need to be compatible with each other, meaning that successful negotiation is usually achieved in the same industry branch, where variables and prices can be compared. Additionally, these apps need to be licensed by some authority so that information cannot be extracted
Pre-conditions	When realized in an IDS environment: both companies have IDS Connectors deployed with negotiation apps downloaded from certified IDS app stores.
Workflow	The following steps are required to perform the use case: 1. The production capabilities and/or production assets of company B are registered in the SFW portal. 2. Company A searches for an adequate production capability and issues a search request to the SFW portal. 3. Company A finds Company B on the SFW portal and places an order for a component or service.

(continued)

	4. The SFW portal performs the search and returns an endpoint address of company B to company A. 5. Company A contacts Company B via the IDS Connector and initiates the negotiation. 6. Negotiation data is exchanged between the IDS connectors involved according to the negotiation needs. 7. The negotiation is handled by the negotiation app without sensitive information ever leaving the IDS connectors including their apps. 8. The information gets deleted and the result of the negotiation presented (contract with terms). 9. Company A and B sign the contract, or they may proceed to the next negotiation step.
Post-conditions	Successful negotiation and contract between previously unknown partners.
Sources	https://www.smartfactoryweb.de
Authors	Employees of Fraunhofer IOSB

When reflected at the data sovereignty aspects presented before, this use case covers the following aspects:

- **Secrecy:** Negotiation data must not leave the IDS connectors and the negotiation apps.
- **Integrity:** Negotiation data must not be modified.
- **Time to live**: Negotiation data shall be deleted after the negotiation process.
- **Anonymization by aggregation/replacement**: not necessarily required.
- **Separation of duty:** Negotiation data shall only be used for the negotiation between companies A and B.
- **Usage scope:** Negotiation data may only serve as input for the negotiation process.

Looking at basic architectural patterns when realizing this use case in the context of an Industrie 4.0/IDS System environment, one option is as follows:

1. The capabilities and/or assets of company B are represented to the SFW portal according to the meta-model and constraints of the Industrie 4.0 asset administration shell and related AAS sub-models.
2. The SFW portal accesses the capabilities and/or assets of company B by means of one of the AAS application programming interfaces (e.g., RESTful API or OPC UA).
3. The SFW portal just acts as mediator between the search request of company A and is not directly involved in the negotiation process.
4. The negotiation process is carried out in a trusted IDS industrial data space by means of interacting IDS connectors.

Please note that other architectural patterns are possible. For instance, the SFW portal may also be connected to the industrial data space via IDS connectors and may act as active negotiation broker including a negotiation policy. In such a setting, the SFW portal may also learn from previous negotiations based on anonymized data in order to improve the negotiation based upon additional information. In this case, the data sovereignty aspects need to be interpreted and applied in a different manner.

This use case analysis shows that the level and aspects of data sovereignty are dependent on the use case to be supported. The basic IDS architecture is designed in a generic manner such that a multitude of use cases may be implemented. By means of a detailed use cases analysis, the IDS-Industrial Community validates this approach and proposes reusable architectural options in the context of a joint Industrie 4.0/IDS system environment.

19.8 Outlook

The mission of the IDS-Industrial (IDS-I) Community is to enable the design, setup, and operation of International Data Spaces tailored to the needs of the application domain of industrial production and/or smart manufacturing. IDS-I is an open community associated with the IDS Association, currently comprising 47 companies and research institutes around the world. The IDS-I Community is convinced that the handling of data sovereignty according to international regulations and standards will become a critical success factor for the European manufacturing industry [8], or may be even world-wide. After the requirements analysis on data sovereignty was published in an IDS-I position paper in June 2022 [9], IDS-I will consider the architectural and technological consequences in joint Industrie 4.0/IDS and GAIA-X system and infrastructure environments.

References

1. Platform Industrie 4.0 (Ed.). *Position Paper 2030 Vision for Industrie 4.0*. https://www.plattform-i40.de/PI40/Redaktion/EN/Downloads/Publikation/Positionspapier%20Leitbild%20(EN).html
2. GAIA-X. https://www.bmwi.de/Redaktion/EN/Dossier/gaia-x.html
3. Usländer, T., & Batz, T. (2018). Agile service engineering in the industrial internet of things. *Future Internet, 10*, 100. https://doi.org/10.3390/fi10100100
4. Cockburn, A. (2001). *Writing effective use cases*. Addison-Wesley.
5. Platform Industrie 4.0: CCM Webinar, 2020. https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Webseminar-Collaborative-data-driven-business-models.pdf?__blob=publicationFile&v=8
6. Fraunhofer IOSB (Ed.) (2020). *IIC testbed smart factory web*. https://www.smartfactoryweb.eu
7. IDS Association. (2019, November). *Usage control in the international data spaces*. Position paper of the IDSA, Version 2.0. Accessible at https://www.internationaldataspaces.org/wp-content/uploads/2019/11/Usage-Control-in-IDS-V2.0_final.pdf
8. Hillermeier, O., Punter, M, Schweichhart, K., & Usländer, T. (Eds.). *Data sovereignty - Critical success factor for the manufacturing industry*. Position Paper of the IDS-Industrial Community Accessible at https://internationaldataspaces.org/download/21213/
9. Usländer, T. (Ed.). *Data sovereignty – Requirements analysis of manufacturing use cases*. Position Paper of the IDS-Industrial Community. Accessible at https://internationaldataspaces.org/download/32789/

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 20
Energy Data Space



Volker Berkhout, Carsten Frey, Philipp Hertweck, David Nestle, and Manuel Wickert

Abstract The energy sector is in a dynamic transition from centralized systems with large fossil power plants to a decentralized system with a high number of renewable energy assets and a rapidly increasing number of additional flexible loads from storage solutions, e-mobility, or power-to-heat applications.

To operate the system reliably, demand and supply have to be matched at all times very closely. Thus, the sector is very data and communication intensive and requires advanced ICT solutions to automate processes and deal with the enormous complexity.

The Energy Data Space can enable the digitalization of the energy transition by providing an architecture to make data available in order to increase the efficiency in asset and system operation.

Data provision and market communication within the system operations of electricity grids is a key use case due to its central role in the sector. Next, the integration of data from the smart meter rollout could as well be built on Data Space technology. Further use cases include predictive maintenance and the energy supply of buildings.

V. Berkhout (✉) · M. Wickert
Fraunhofer Institute for Energy Economics and Energy System Technology IEE, Kassel, Germany
e-mail: volker.berkhout@iee.fraunhofer.de; manuel.wickert@iee.fraunhofer.de

C. Frey
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Advanced System Technologies Branch (AST), Ilmenau, Germany
e-mail: carsten.frey@iosb-ast.fraunhofer.de

P. Hertweck
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Karlsruhe, Germany
e-mail: philipp.hertweck@iosb.fraunhofer.de

D. Nestle
Smartrplace GmbH, Kassel, Germany
e-mail: david.nestle@smartrplace.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_20

Initial research projects have demonstrated the feasibility of basic use cases. On the European level, the Platoon project will provide seven pilot applications by 2024.

20.1 New ICT Solutions for Decentralized, Data-Intensive, and Distributed Processes

The decentralization of generation, consumer, and prosumers as new participants in flexibility markets, sector coupling, and the use of new decentralized storage technologies are leading to a massive increase in complexity in the planning and operation of energy systems. To reduce these complexities, the automation of operational processes in the energy system is vitally important.

The digitalization of the energy system thus becomes a necessity for the transformation of the energy system towards 100% renewables.

International Data Spaces are currently emerging as a European alternative to the data lake approach of hyperscalers. The basic idea of preserving companies' sovereignty over their data and to promote the exchange of data for new data-driven services seems particularly promising in the energy industry, where energy system and energy economy processes are distributed across a whole range of devices and companies.

Based on the reference architectures of the International Data Space Association and the GAIA-X initiative, a European Energy Data Space is being created. The Energy Data Space supports the linking of data silos within the companies in order to enable innovative process automation for the energy industry while at the same time preserving the data sovereignty of the participating companies.

20.2 Use Cases in the Energy Data Space

The early use cases implemented in the Energy Data Space are laying the necessary foundations for a digital ecosystem in which more and more companies in the energy industry are establishing themselves, developing and providing digital tools for IDS-supported processes in the energy industry, and creating data structures for the standardized use of applications.

Using the complete value creation processes, demonstrators will be used to show that an efficient linking of information and software solutions is possible in a scalable manner in a European energy data space. In particular, the cross-company and sovereign data exchange should enable new processes that are able to increase the economic efficiency of the energy supply system without having to compromise on the security of supply. The following are some use cases that may seed future ecosystems.

20.2.1 Communications in Electrical Grids

The operation of transmission and distribution grids for electricity supply is a key part of the energy economy. The operation of grids requires information from a large number of generating units as well as an even larger number of electrical loads and consumption. Supply and demand have to match very closely with a very high time resolution to ensure the operation of the electricity system within the range of the grid codes and to provide reliable power supply to electricity consumers.

In addition to these characteristics, regulation adds complexity to transmission and distribution system operation as these systems are regulated monopolies governed by specific regulations to ensure fair access and competition in the electricity market under the European unbundling regulations.

Besides energy grids are critical infrastructures and as such subject to special regulations, e.g., on IT security and data confidentiality.

International Data Spaces can provide a framework and building blocks to meet the high requirements on ICT technology in grid operations. IDS use cases can build on existing standards for data that are already in place and increase efficiency in the existing system. Data spaces will make data easily available for further use like the training of AI models for forecasting or predictive maintenance.

Regulations may be translated into usage control policies and may enable data usage that is prohibited if the transfer of grid data was required. If the sensitive data is only used but not shared, new applications may become viable and benefit operational decisions.

20.2.2 Predictive Maintenance

As in other industries, predictive maintenance is a major trend to increase the equipment efficiency and availability by detecting anomalies in the operational behavior of assets and react with maintenance actions before a failure and unplanned downtime occurs.

Within the energy sector, wind farms are a good illustration of the advantages of data exchange and the formation of data ecosystems.

Wind turbines are similar worldwide in their design and function. This technical homogeneity makes it a promising scalable market for standardized applications if data can be accessed and interpreted easily. In the present state, manufacturers largely rely on proprietary data management approaches and offer digital solutions as added services to their customers.

The full potential of data, however, is only unlocked by combining different data sources and services based on them. To resolve the conflicts around data use and for establishing collaboration in data ecosystems in the wind industry, secure, i.e., trustworthy, infrastructures and independence are needed.

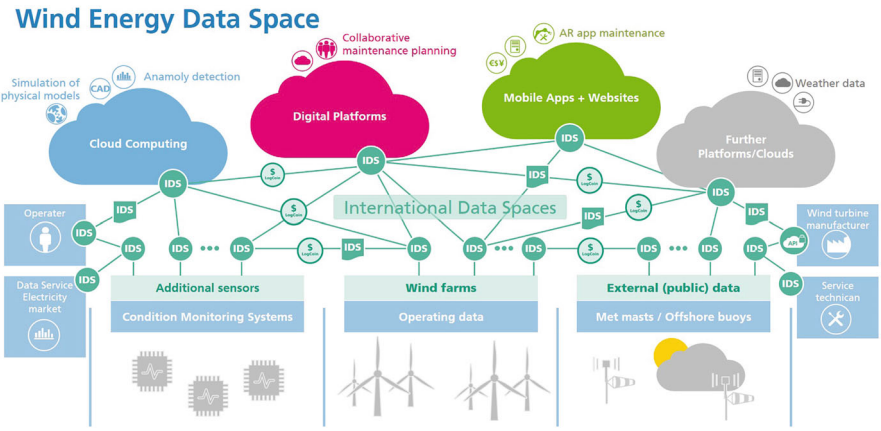


Fig. 20.1 Participants and connections for predictive maintenance in an energy data space (©2020, Fraunhofer IML)

The vision of an Energy Data Space is one of a purpose-built, domain-specific, and prosperous data ecosystem (Fig. 20.1). Following the overall picture of a Silicon Economy, it combines the data basis of operational and plant data with digital platforms and cloud technologies. Everything is networked and interoperable. Different participants, services, and data assets are found and dynamically used via brokers. The case for predictive maintenance for wind farms has been described in a white paper [1] by Fraunhofer IEE and IML.

This information from predictive maintenance applications is particularly valuable for the plant operator and the maintenance service provider. Furthermore, the evaluation of this data is also interesting for the manufacturer of wind turbines as they may promote product development and design the turbines more efficiently and more resistant to faults. The same applies to component suppliers. In addition, there are other interest groups that can benefit from access to data such as consulting services or insurance companies.

Another potential is that anomaly detection data and results are assessed by experts for specific components or types of equipment. Enriched by their expert knowledge, the identification of the fault or the prediction of possible damage is complemented. In the context of a data economy, new business relationships between stakeholders can thus emerge and new business models developed along the usual business processes around operation and maintenance. The various functionalities in the business process are performed by different participants in the data ecosystem.

20.2.3 Energy Management Gateway: From the Perspective of an SME

Smartrplace is a spinoff company of Fraunhofer IEE and offers fully integrated Smart Building energy solutions for offices, schools, shopping malls, and community buildings. The initial core of the product is a building automation solution that enables demand-driven individual room control of heating, ventilation, and air conditioning. As a next step, Smartrplace offers the full digitalization of building energy supply. This also includes "smart districts," which comprise entire residential complexes or neighborhoods. The number of players and stakeholders that need to be considered in such solutions leads to complex requirements. These players include tenants, janitors, building cleaners, facility managers, etc. There are significantly higher requirements in the areas of maintenance and reporting compared to individual buildings.

This also means a tight integration of operational planning of local electricity and heat energy supply as well as integration of electric vehicle charging requirements. Within the SINTEG beacon program, Fraunhofer IEE has worked with many partners in the C/sells project to look at both sides of the activation of flexibilities: their provision by plant operators, suppliers, and service providers and their use by grid operators. The researchers have paid particular attention to the backbone of activation—information and communication technology infrastructures and smart metering. With the EnergyPilot a new energy management software for activating flexibilities was developed that can be used in conjunction with the Virtual Power Plant offered by the institute or alternative control systems.

With Fraunhofer IEE's EnergyPilot, the Smartrplace solution can be expanded into an energy management system that mobilizes the flexibility of consumers and generators in the building for the grids (Fig. 20.2).

For SMEs providing smart energy solutions for building operation, the IDS architecture and standard provides an excellent basis for scaling, as very different systems can be mapped via the cloud. The servers of the various operators and manufacturers can easily be made EDS-capable, which has high potential of facilitating the integration of the different interfaces at the field level in each case.

Relevant logic in the form of connectors for data processing can either be installed at the data source in a protected environment and perform data pre-processing, or at the data user side. In the first case, only the aggregated data or data computed specifically for the use case needs to be transferred, which saves transfer volume or is often the only viable solution for large amounts of data, while ensuring that no unauthorized use of the full base data occurs.

As significant parts of the API are publicly available [2] and decisive components are offered as Open Source under the Apache License [3], IDS implementations can be implemented more and more easily. Furthermore Fraunhofer provides a range of information, services, and additional implementations [4].

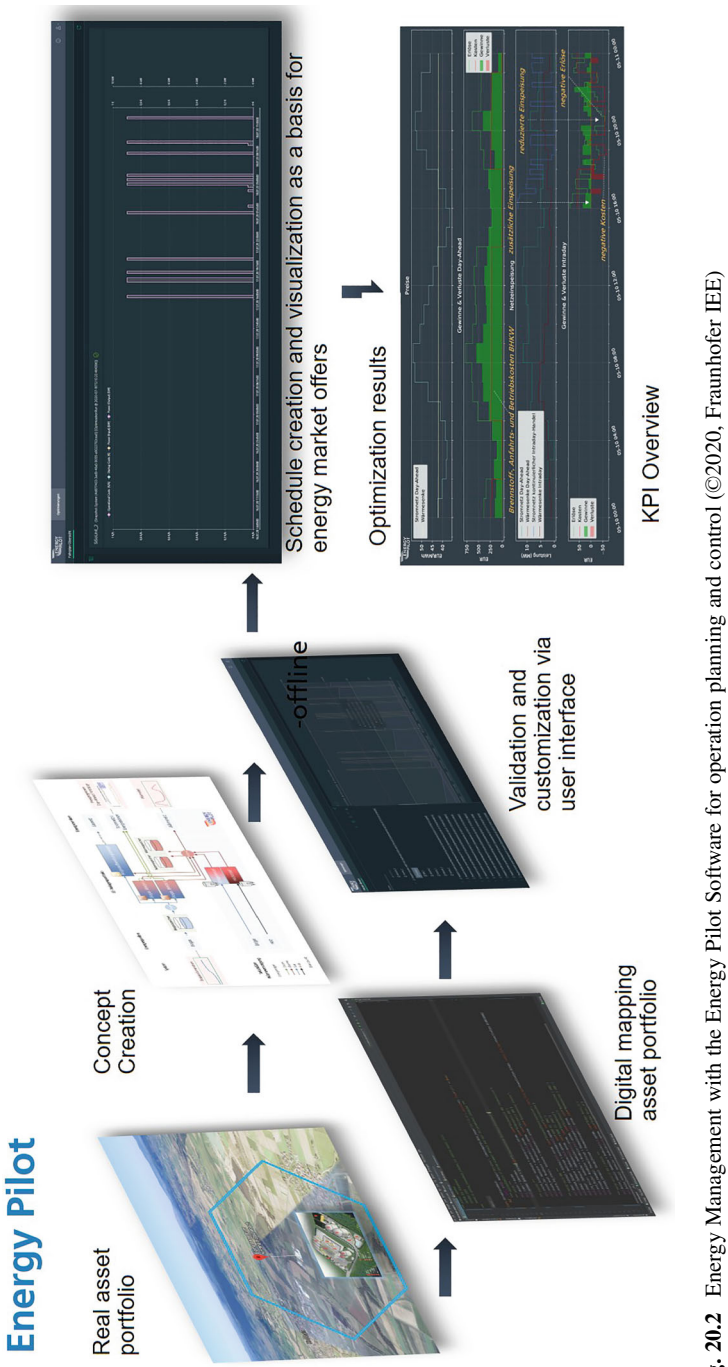


Fig. 20.2 Energy Management with the Energy Pilot Software for operation planning and control (©2020, Fraunhofer IEE)

20.3 Early Demonstration Projects

A number of projects to initially demonstrate the use of the technology of the IDS in the energy sector have already been started. Within the EU Horizon 2020 program, the R&D project platoon [5] is working in a 3-year project since the beginning of the year 2020 on the digitalization of the energy sector through data governance for multi-party data exchange via IDS-based connectors. The French energy utility company ENGIE is coordinating the project with 19 partners among which are TECNALIA, Fraunhofer IAIS, and several European universities as research partners.

The project's objective is to demonstrate the use of the IDS in seven pilot use cases:

1. Predictive Maintenance of Wind Farms
2. Electricity Balance and Predictive Maintenance
3. Electricity Grid Stability, Connectivity, and Life Extension
4. Office Building: Operation Performance with Physical Models and IA Algorithms
5. Advanced Energy Management System and Spatial (multi-scale) Predictive Models in the Smart City
6. Energy Efficiency and Predictive Maintenance in the Smart Tertiary Building Hubgrade
7. Energy Management of Microgrids

The European strategy for data calls for a "Common European energy data space, to promote a stronger availability and cross-sector sharing of data, in a customer-centric, secure and trustworthy manner" (European Commission [6], S. 22), which will be further developed with additional European research funding which will be available from the Horizon Europe program.

20.3.1 Fraunhofer Demonstration Project "EnDaSpace"

As a kick-off for the demonstration of data spaces in the energy sector, Fraunhofer funded a project to learn and show how the IDS works with energy data. Data from a wind turbine operated by Fraunhofer was provided to the EMS-EDM PROPHET® as digital service to calculate schedules for the operation of an electrolyser at the Fraunhofer power-to-gas research facility (Fig. 20.3).

All data was communicated through IDS connectors. A subset of operational data of the wind turbine had been defined, pre-aggregated, and provided as a data resource. The EMS has been extended to be able to communicate through the IDS Connector with other IDS instances. To achieve this functionality, the internal time series management was wrapped as an interface to a message broker.

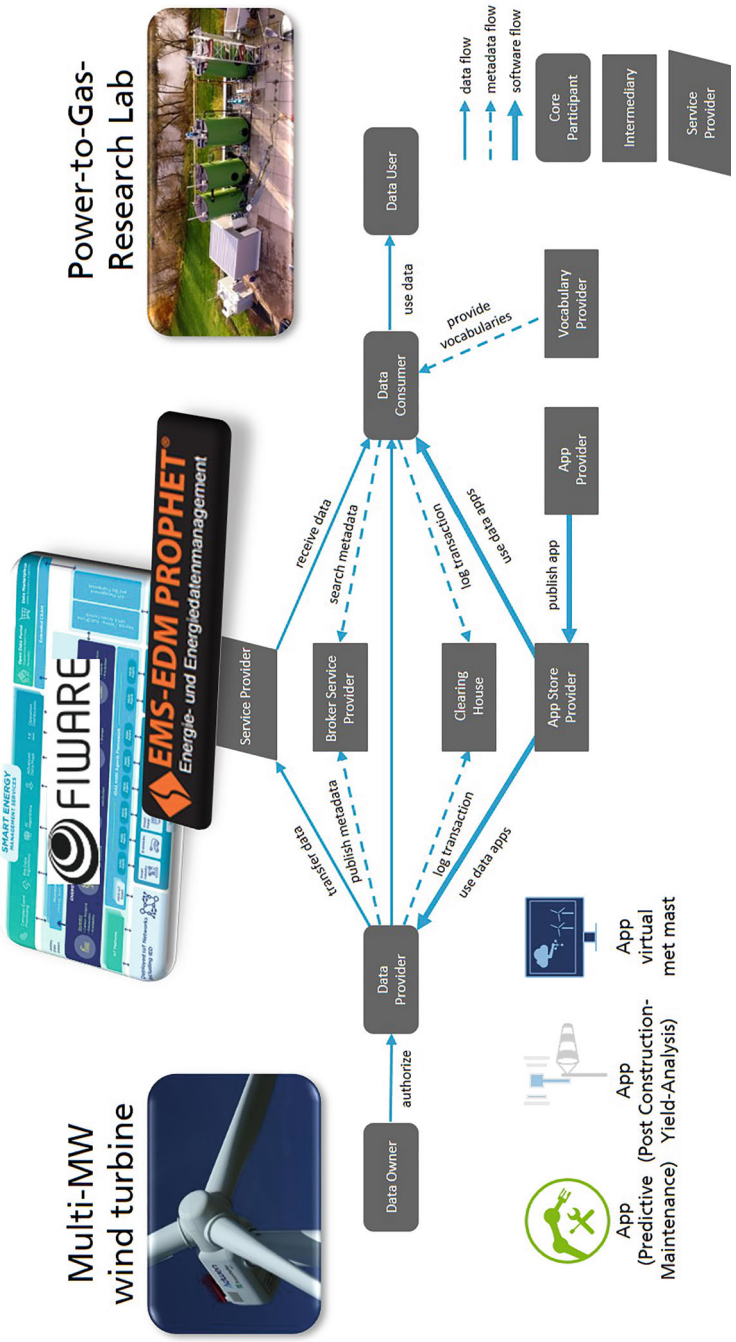


Fig. 20.3 Elements in the IDS role model in the EnDaSpace demonstration project (©2020, Fraunhofer IEE)

The electrolyser was adapted to read and make use of the schedules in its control system.

Based on wind power forecasts and electricity market prices, an optimal schedule for producing green hydrogen was calculated daily. The optimization model was created in such a way that it prefers to sell the generated energy at the energy market until a certain limit of the energy price is exceeded which then triggers the start of hydrogen production.

A digital platform based on FIWARE technology was used to visualize operational wind data. Applications for anomaly detection in predictive maintenance and for post-construction yield assessment and a virtual met mast have been developed for operation directly with the IDS connectors or within the FIWARE environment.

In addition to the technical demonstrator, further use cases for future applications have been developed for oncoming projects.

20.3.2 Bauhaus.MobilityLab

With the vision "Innovation by experiment," the Bauhaus.MobilityLab (BML-EcoSys) started in July 2020 as a real-world laboratory in the district Brühl of the city Erfurt, Thuringia, Germany. The consortium for this research project, funded by the Federal Ministry for Economic Affairs and Energy (BMWi), consists of the Fraunhofer IOSB, other research institutions, companies, universities, and the Thuringian state capital Erfurt. A complete list of all partners can be found at bauhausmobilitylab.de. Erfurt presents itself as a representative of a typical German city and is therefore ideally suited for a real-world laboratory that includes the sectors of mobility, logistics, and energy.

The real-world laboratory in the district Brühl is in its way unique in Germany. It is the first one that does a fusion between the different sectors of mobility, logistics, and energy. Next to experiments executed as part of the project, BML opens the laboratory for customers. They should be supported in developing and evaluating new data-driven business models by linking data sources available on the platform with own data sources. This is supported by the possibility to integrate existing artificial intelligence (AI) models, either provided by the platform or developed by the customer. The laboratory will not be set up with defined and hard-coded methods of experiments but will allow the user to be able to supply new methods or link existing methods and data in new ways.

The goal of the BML-EcoSys is to bring together several roles and stakeholders:

- Laboratory customers develop or test new services within the real-world laboratory.
- Laboratory users live, work, or visit the district Brühl and use the services that are developed or tested within the real-world laboratory.
- Infrastructure partners supply the necessary basic infrastructure (mobility, electricity, logistics) in the real-world laboratory. In addition to physical

infrastructure, partners can bring in new data sources, which can be shared (based on defined policies) with laboratory customers.
- The laboratory operator takes care of the laboratory and maintains the platform. He also is responsible to support the laboratory customers.

A typical use case of the BML-EcoSys could be a laboratory customer that is developing a new e-mobility service by providing e-bikes for rental. Since the success of such a service depends on various factors, it should be evaluated in a practical experiment.

The BML platform provides access to data, infrastructure, as well as analytic methods to easily realize the service. With the involvement of the people as laboratory customers, the idea can be evaluated in a realistic environment. Data collected during the experiment can be evaluated by analysis methods available on the platform to further improve the service offering. Incentives can be given to users to encourage the usage of the e-bikes, for example, when the forecasted energy price rises above a certain limit and the recharging would be inexpensive in the next hours. Another threshold could be the air quality: if it drops below a certain limit, the customer could motivate the users to use his e-bikes instead of a car to help to make the air quality better. There are an unlimited number of use cases in the fields of energy, mobility, and logistics that are imaginable.

In the project, great importance is attached to the sovereignty of data. Like mentioned before, different organizational actors are part of the BML-EcoSys platforms. Therefore, data from many different providers, sensors, and systems are integrated in a common data sharing platform. This allows to link different data sources and to combine multiple of them during the analysis. Freely sharing data is not possible due to legal, regulatory, or economic reasons. The available data (e.g., current energy usage, traffic, air quality) leads to an advantage in the market. Therefore, actors are usually willing to share data in a controllable and restricted way. In addition, shared data like the mentioned energy usage is subject to privacy restrictions like the GDPR. It is important that the sovereignty of the individual datasets is always guaranteed so that usage can be strictly controlled and misuse can be prohibited. To ensure this, the BML platform relies on the reference architecture of the IDS. Inside the platform data is only available through the IDS. For every data source, this allows to define and enforce usage policies in a fine-grained way. One of the main goals of the BML platform is to provide an open and extensible structure. This allows data providers to offer their data to laboratory customers either free of charge or for a fee and to expand the platform with data from participants for evaluation purposes.

The platform for the real-world laboratory is developed and implemented by the Fraunhofer IOSB. Like already mentioned, key elements of the platform are the sovereignty exchange of the data and the integration of AI components; in addition it provides an execution environment where customers can deploy and evaluate their services. The platform architecture is based on the reference architecture “Open Urban Platform (OUP)” that’s defined in the DIN SPEC 9135. Next to unprocessed—raw data—it’s foreseen to provide pre-processed data—smart data.

This allows project partners and customers to pick the data sources needed to implement their service. Since not all of the available data sources are freely available and since customers might belong to different organizations, IDS connectors are the central point to declare and enforce data usage policies. By this every organization interacting with the BML platform has its own identity in the IDS ecosystem. To offer data on the BML platform, the data can be either copied to the platform or an external system can provide data through the IDS. Analogous data sources on BML can be accessed by external systems through the IDS. Most of the available data source considered in the BML project are time series. To store data in a common and standardized data model as well as to provide a powerful query interface, time series data is stored in the OGC SensorThings API on the platform. This is realized by using FROST®, an open source implementation of the standard, developed by the Fraunhofer IOSB. To enforce usage policies, the FROST® instances aren't directly accessible inside the platform. An IDS connector is placed in front to implement this control. Therefore customers, as well as AI services (integrated using PERMA, a component developed by Fraunhofer IOSB or Kubeflow), can access all data sources only through IDS.

The software solution EMS-EDM-PROPHET® developed by Fraunhofer IOSB-AST for the energy sector offers a broad portfolio of algorithms for time series management, forecasting, and optimization as well as other essential functions in energy data management. Contract structures and network topologies can also be mapped using the software system. EMS-EDM-PROPHET® is the leading software solution in this domain. Therefore, to support the above-described use cases of the energy domain, a tight integration of PROPHET and the BML platform is needed.

Our goal is to develop services for the BML project that are optimized for operation in a cloud environment. For this, we will develop relevant components into a service-oriented or microservice architecture. We will use open-source technologies for the implementation. As a result, the developed services should be characterized by very good availability and security for the BML-EcoSys platform. One example will be an AI-supported method for an automated selection of forecast methods and models in the energy sector. This will be implemented as microservice with an IDS/EDS interface for making the automated selection of the forecast method and the method itself accessible though the BML-EcoSys platform. With this universal IDS/EDS Connector implemented, it's possible to extend the forecast models in near future by more complex AI that will produce even better results. Such AI forecast models then could also fuse data between the different sectors that are available in the laboratory.

Next to the energy data, another example for shared data inside the BML platform is calendar information. To optimize the (AI-)based forecasting models, additional information like public holidays, long weekends, school holidays, or events in the city are necessary. In contrast to other data sources, most of this information is publicly available and no strict usage policies exist. To unify data access within the BML platform calendar information is also available through the IDS.

The Fraunhofer IOSB-AST has already contributed to the further development of the reference implementation of an IDS connector provided by Fraunhofer ISST and

created a connection to the EMS-EDM-PROPHET® software system. This allows IDS-compliant data exchange with other data providers or data consumers.

20.4 Summary and Outlook

The dynamic development as part of the energy transition towards renewables with its large need for data and communication of producers and consumers of electricity as well as grid operators make the energy sector a very promising area for an Energy Data Space.

The prototypes being built in the early research project will spark a first round of early implementation across the energy sector. Specifically pilots and demonstrators will be available for wind energy data energy consumption in buildings and for grid operation. Soon afterwards sector coupling application connecting mobility and heat with the power sector will emerge. With the strong commitment and funding from the European Commission, this will come with the opportunity to scale solutions and applications on the international level within the forming of the European Energy Data Space.

In a second wave of adoption, first movers and early adopters will build up experience with data spaces, and hesitant market participants may build trust into IDS technology. This is the requirement to expand the ecosystem to additional participants and to start the use of central services. The strategy of grid operators may be of critical importance as ICT requirements for grid infrastructure affect all participants in the sector.

After first operational processes in the energy sector have moved into the Energy Data Space, there may follow another wave of exploring new options and business models within the IDS architecture and the evolved IDS components. This will include integration with other data spaces like materials or logistics and the exploitation of the growth of data and applications available.

This scenario may be facilitated by some factors that would be relevant for the acceptance and adoption of an Energy Data Space:

First, participation and coordination in the industry through the relevant industry associations will prove very helpful to agree on initial processes and governance models as well as on suitable information models and vocabulary.

Second, barriers for entry of new participants have to remain low as the energy sector includes a large number of SMEs that may shy away from initial investments. Therefore basic components to onboard to the data space should be available under open-source licenses. This also supports the building of transparency, trust, and a developer community.

Moderating these communities will be a major task for governing institutions such as the IDSA to enable an efficient collaboration within the sector.

The Energy Data Space is an evolving concept to manage the growing complexity and data intensity of the energy sector which is needed to achieve the common goal of a sustainable and renewable energy future.

References

1. Berkhout, V., & Skubowius, E. (2020). *Predictive maintenance für Windenergieanlagen - Energy data space whitepaper.* Dortmund. Retrieved from https://internationaldataspaces.org/download/19022/
2. IDS Metadata Broker API. (2021). Retrieved from https://app.swaggerhub.com/apis/idsa/IDS-Broker/1.3.1
3. International Data Spaces Association. (2021). *Github repository.* Retrieved from https://github.com/International-Data-Spaces-Association
4. Fraunhofer Gesellschaft. (2021). *Webseite international data spaces.* Retrieved from https://www.dataspaces.fraunhofer.de/
5. Garatzogianni, A. (2021). *Projekt-Webseite Platoon.* Technische Informationsbibliothek (TIB). Retrieved from https://platoon-project.eu/
6. European Commission. (2020). *A European strategy for data.* Brussels. Retrieved from https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 21
Mobility Data Space



A Secure Data Space for the Sovereign and Cross-Platform Utilization of Mobility Data

Sebastian Pretzsch, Holger Drees, and Lutz Rittershaus

Abstract To successfully support decision-making or even automatically make decisions of their own, intelligent transport and mobility systems require large amounts of data. Although multitudes of mobility data are already being collected today, the comprehensive processing and exploitation of this data have often been impossible due to technical, legal, or economic reasons. With Mobility Data Space, an open data space is now being created which offers access to real-time traffic data and sensitive mobility data beyond their secure exchange and which links existing data platforms to each other. In the future, it will thus be possible to provide comprehensive mobility data on a national level.

Based on a decentralized system architecture developed by the International Data Spaces Association e. V., the Mobility Data Space offers an ecosystem in which data providers can specify and control the conditions under which their data can be used by third parties. This approach creates data sovereignty as well as trust, and data users can be sure about data origin and quality. By integrating data from the public and private sector via regional and national platforms, the Mobility Data Space will become a digital distribution channel for data-driven business models, providing entirely new options of data acquisition, linking, and exploitation.

Whether data provider, user, developer, or end user—the Mobility Data Space takes all acting parties into consideration and offers:

- Data sovereignty and security along the value chain
- Standardized access to data from both public and private sources
- Space for the emergence of new business models, distribution channels and services, as well as a larger selection of innovative mobility services and applications

S. Pretzsch (✉)
Fraunhofer-Institut für Verkehrs- und Infrastruktursysteme IVI, Dresden, Germany
e-mail: sebastian.pretzsch@ivi.fraunhofer.de

H. Drees · L. Rittershaus
Bundesanstalt für Straßenwesen, Bergisch Gladbach, Germany
e-mail: Drees@bast.de; Rittershaus@bast.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_21

International Data Space
The International Data Space[1] (IDS, until 2019 known as "Industrial Data Space") was designed by the Fraunhofer-Gesellschaft in 2015 with the aim of creating a safe data space for the sovereign management of data assets by enterprises from different fields. Due to the overwhelming feedback to the industry initiative, the International Data Spaces Association e. V.[2] was founded in 2016. It is continuing the development of the International Data Spaces reference architecture and supports the industry in its introduction and implementation.

Fraunhofer coordinates the adaptation of the IDS reference architecture to sector-specific needs through IDS verticalization. The Fraunhofer IVI leads the Mobility Data Space verticalization initiative, supporting the transportation and mobility sector in creating new mobility services using data-based business models in sovereign data ecosystems.

21.1 Mobility Data: The Status Quo

Collecting mobility data is increasingly gaining in importance. It is the only way in which intelligent systems can provide traffic participants and decision-makers with sufficient information to optimize traffic flows, increase safety, and protect the environment.

The interaction of several different traffic participants, providers, and operators requires a trustworthy exchange of data and their interoperability.

In the field of mobility, quite a large amount of data requires protection. Among these are data on traffic infrastructure and real-time data on the current traffic situation. Data from different sources need to be merged either physically or virtually at the point of decision.

Data acquired by the German Federal Government and the Federal States are already being provided to users in a standardized format in the Mobilitäts Daten Marktplatz (MDM, Mobility Data Marketplace) of the Federal Ministry of Transport and Digital Infrastructure (BMVI).

On a regional level, this data is partly available via corresponding platforms. The fact that the data is rarely ever provided in a national context complicates its multi-regional exploitation.

[1] https://www.fraunhofer.de/en/research/lighthouse-projects-fraunhoferinitiatives/international-data-spaces.html.

[2] https://www.internationaldataspaces.org/.

Public transport providers, car sharing providers, and charging station operators are usually reluctant to provide further mobility data to third parties. Reasons for this include the lack of infrastructure (i.e., a National Access Point for data exchange) and the lack of established data formats and interfaces, which as of yet do not apply in certain sectors (car sharing, bike sharing, e-mobility).

Sensitive data, such as passenger flows, which is generated by vehicles or privately owned mobile devices, is being collected and processed by public transport providers, navigation service providers, fleet operators, and mobile communications providers. However, there is virtually no cross-company utilization, processing, and linking of this data due to its sensitivity in terms of data protection, informational autonomy, and protection of trade secrets.

Security and Sovereignty for New Exploitation Options

The Mobility Data Space offers a solution: an open mobility data ecosystem in which data providers can specify and control the conditions under which their data may be used and exploited by third parties. This approach creates data sovereignty and a trusting environment for data providers, and it gives data users assurance about data origin and quality.

Through the assurance of data sovereignty, data that had previously not been usable due to its sensitive nature can now be exploited. The mobility data space will become a digital distribution channel for data-driven business models. The linking of public and private data via regional and national data platforms through a decentralized data space concept provides completely new options for data acquisition, linking, and utilization.

Within the scope of a collaboration between the Fraunhofer-Gesellschaft and ca. 100 enterprises, the International Data Space (IDS) was created as a basis for decentralized data value chains. Ever since its inception, the IDS has been continuously improved by the International Data Spaces Association e. V. For a better combination of resources, the future will see the integration of cloud infrastructures into data spaces and their interlinking through projects such as GAIA-X.

21.2 The Mobility Data Space: Architecture and Components

Beyond the IDS' technical functionalities in the area of secure and sovereign data exchange, the Mobility Data Space aims at making accessible real-time traffic data (e.g., sensor data, traffic light sequences) and sensitive mobility data (e.g., vehicle-generated and smartphone-generated data, movement patterns) as well as connecting local, regional, and national data platforms in order to facilitate the provision of comprehensive mobility data on a national level. Services and applications for data enhancement and exploitation form the basis for a broad mobility data ecosystem.

The Mobility Data Space

- Is based on the open, decentralized system architecture developed by the International Data Spaces Association e. V.
- Guarantees data providers sovereignty over their own data and security along the processing and value chains in the sense of a digital rights management system
- Allows the provision and distribution of sensitive data, as well as the traceability of their use for purposes of billing/payment
- Provides data users (e.g., travel information services) with standardized access to an ecosystem that pools data from public and private sources and services through the connecting of local, regional, and national platforms
- Opens up new business opportunities
  - For developers: data apps for mobility services and applications, including distribution via a data app store
  - For IT service providers: hosting of components and data apps in cloud environments as well as corresponding consulting services
- Offers advantages for end users by fostering the development of novel mobility applications and services through the availability of mobility data sources.

21.2.1 Data Sovereignty Through Usage Control

Participation in a secure data space is possible via a technical connector component that data providers and data users either host themselves or have hosted for them. The data space is established across the networked connectors, meaning that it is not a centralized platform but rather an expandable network of decentralized players (minimum of two). Before being transferred to the target connector, the data to be provided is extended by a set of rules, the so-called "usage policy." The data remains in the target connector and is secure against direct access by the data user. If data users want to work with the data, e.g., for purposes of data analysis or fusion, they must access it within the connector via so-called "data apps."

These apps are capable of integrating further data, e.g., from user databases that are run outside of the connector. A usage control layer within the connector guarantees compliance of the data app with the specified rules, with the effect that only aggregated results will leave the connector. All steps taken during data use and processing within the data space can be recorded. This way, data providers have complete knowledge of all activities relating to their data (Fig. 21.1).

21.2.2 The Mobility Data Space as a Distributed System

Beyond the minimum example, a data space can consist of dozens or even hundreds of participants. This kind of decentralized, distributed system requires a central

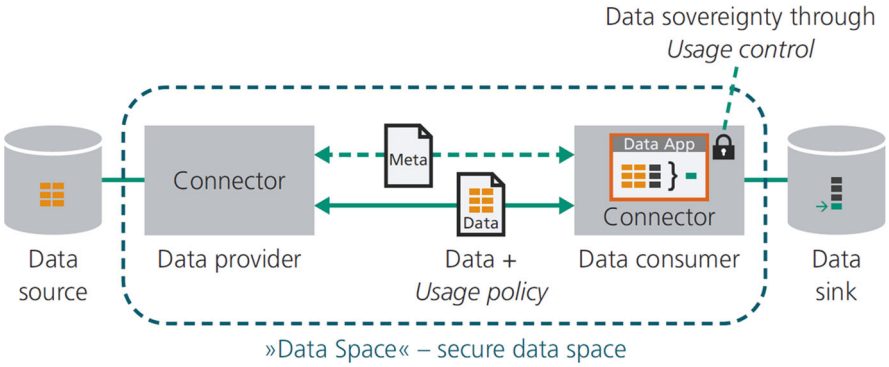


Fig. 21.1 Functional principle of data sovereignty through mechanisms (©2020, Fraunhofer IVI)

directory in which data sources and services are published and which can be searched either manually or automatically by data users. Therefore, existing regional and national mobility data platforms play a special part within the Mobility Data Space. With different operator and business models, one or more central components for the data space can be offered:

- A data marketplace (technically, a metadata directory), for the publication and displaying of data sources and their terms of use. Metadata needs to be provided in a machine-readable format so that devices such as automated vehicles, smartphones, and IoT devices will be able to find and use them autonomously.
- A vocabulary provider that provides the necessary domain knowledge about traffic and mobility data formats (e.g., DATEX II, NeTEx) as well as APIs (e.g., SIRI, TRIAS) in the form of vocabularies and ontologies, thus ensuring the machine readability and interoperability of data.
- An identity provider as a single point of contact that evaluates the trustworthiness of data providers, data users, as well as data and data apps and that also allows secure communication based on the aforementioned evaluations.
- A data app store for the easy registering and marketing of data apps (for the processing of data relating to mobility).
- A clearing house, the system's central logging component, that records transactions made within the distributed system in order to make them available to the relevant parties for purposes of billing and quality analysis at a later point in time.

The connector also allows the exchange of data between data providers and users via the platform. This facilitates the brokering of data through which data users can subscribe to data publications and receive the data provided by the respective data providers in real time. In addition to this brokering task, the connector can execute data apps, for example, to compile the data provided to the platform into new virtual data sources. This way, existing data platforms can be extended to receive sensitive data worth protecting as well as mobility data from data providers and other data platforms and to transfer them in compliance with the usage policy to data apps for enhancement and exploitation.

21.2.3 Design and Operation of Central Components

Due to the important role of the central components in the Mobility Data Space, additional organizational issues must be considered (Fig. 21.2):

- The neutrality of the central components operator is an important prerequisite for a guaranteed discrimination-free exchange of mobility data. This neutrality may be ensured, for example, by a public authority or an association.
- The funding of the central components operation must be stable—not least for creating trust in the concepts of the Mobility Data Space. If operators have to raise user fees in order to cover their costs, the attractiveness of participating in the Mobility Data Space will decrease for all parties. In addition, funding models such as the promotion of data services might impair neutrality.
- There has to be a continuous harmonization of the data formats and models provided by the vocabulary provider. Communication and coordination with the relevant stakeholders are important to identify changing requirements of the data formats and models and to find solutions. Predefined processes can be a way to better include the stakeholders.
- Because license and usage policies are new for many parties acting in the field of mobility, examples and patterns should be offered.

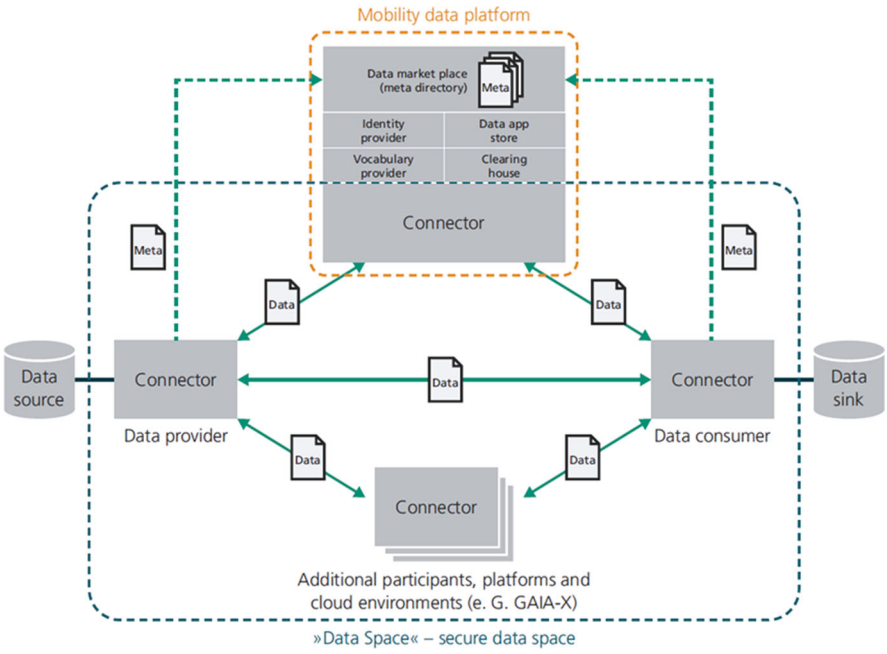


Fig. 21.2 Secure data space (©2020, Fraunhofer IVI)

- Although marketing does not play a key role in data exchange itself, it is an important element in broadening the implementation and knowledge about the Mobility Data Space. Because utilization of the central components lies in the interest of the central components operators, appropriate marketing is necessary.

21.3 The Mobility Data Marketplace (MDM) as Central Platform Within the Mobility Data Space

The Mobility Data Marketplace (MDM) is a platform that already covers some of the concepts of the Mobility Data Space. The Mobility Data Space concepts can enhance the MDM's functionalities, thus increasing its attractiveness. The MDM is known as the central point of contact for road traffic data in Germany. Because it is operated by the Federal Highway Research Institute (BASt), it has a neutral position. This way, data providers can rely on a neutral IT infrastructure that is not influenced by the interests of private economy. Currently, the most important providers of road traffic data are authorities on all levels of public administration, ranging from ministries to small municipalities. With the help of this data and the results of their processing by service providers, traffic participants will receive better information, and both the safety and efficiency on roads will increase.

The MDM offers two core functionalities:

- The MDM has a metadata directory for searching relevant data publications. The directory's entries can be filtered according to various criteria.
- Due to its brokering functionality, the MDM is a data distributor: Through 1:n distribution, data provision is made easier for both data providers and data users. Data providers offer their data publications, and interested data users can then subscribe to them. This means that the MDM is not focused on end users (travelers, users of a mobility app, etc.), but on establishing data exchange in the B2B sector (e.g., infrastructure operators and service providers).

For data exchange via the data distributor, the MDM primarily uses the DATEX II data model. This European standard is commonly used in traffic control centers and is required by law as a basis for the exchange of traffic data. The MDM website provides DATEX II profiles for several data types. With the help of these profiles, data providers can identify the requirements for the individual elements of their data publications, and data users know what to expect from the publications so that they are able to integrate them into their systems.

Some of the MDM's functionalities correspond with the core components of the Mobility Data Space:

- The data marketplace is the core functionality of the MDM. Metadata is searchable via a web-based user interface, but it is not machine-searchable. Also, it is

possible to distribute usage and content data in addition to metadata via the data distributor. Through this type of 1:n distribution system, a large number of subscribers to a certain offer can receive real-time data, while the data provider only has to manage one interface. This way, only the most recent data content is available in the MDM, and the historicization of data does not take place.
- The vocabulary provider functionality is already partly supported by the MDM through the provision of DATEX II profiles.
- While access to the metadata search is free, data providers and data users have to register as users for the MDM. The range of certificates is comparable to that of the identity provider functionality.
- In analogy to the clearing house, transactions are also logged in the MDM. However, a standardized procedure following the IDS concept is not implemented in the MDM.

The MDM is currently not implemented in an IDS-compliant way. The metadata directory is not machine-readable, data distribution is not carried out via a connector, and the mobility vocabulary provider, identity provider, and clearing house components were not created according to the concepts of the Mobility Data Space. Also, a data app store is missing.

Figure 21.3 shows how the MDM, as complemented by an IDS connector, could become a part of the Mobility Data Space, run data apps, and broaden its spectrum of

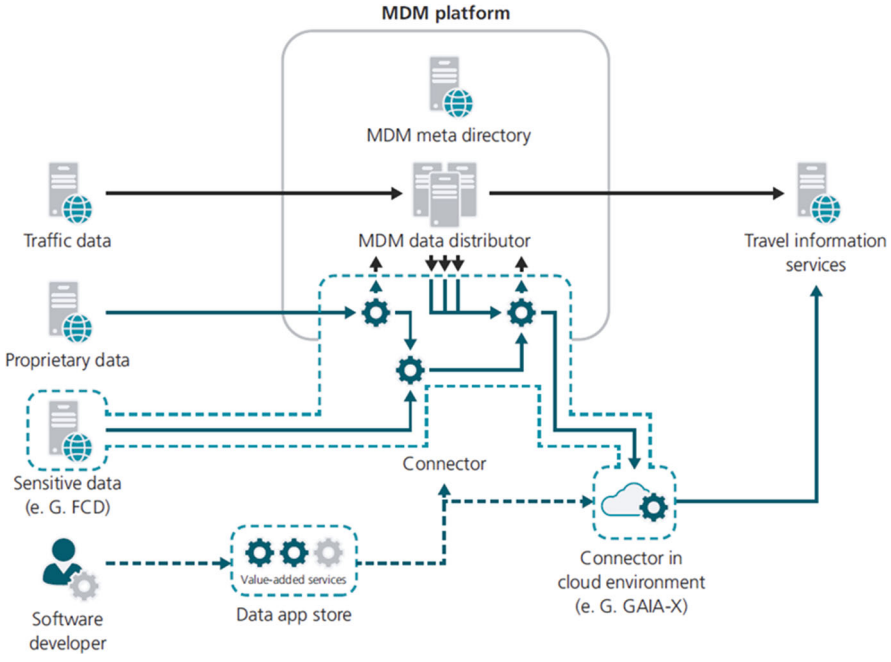


Fig. 21.3 The MDM, extended by an IDS connector, as part of the Mobility Data Space (©2020, Fraunhofer IVI)

services through data processed by those apps. Looking at the necessary organizational aspects of a central platform within the Mobility Data Space, the MDM already considers various aspects relating to the operation of central components in a similar way:

- As a neutral operator, the MDM is trustworthy. Therefore, additional key roles, such as the identity provider, the app store provider, and the vocabulary provider, can also be taken on by the MDM.
- Commercial services and advertising are currently not pursued in the MDM as ways to raise funds, as such practices might impair neutrality. However, should the hosting of data apps that explicitly create added value require extensive resources, fees might be considered.
- The limitation to DATEX II as the only model on the platform could be lifted. In particular, the increased inclusion of mobility data beyond road traffic calls for the adoption of additional data standards. In the future ecosystem, additional standards will be recommended and developed, and conversions between them will be supported. Therefore, it is imperative to harmonize the use of data standards.
- Machine-processible standard licenses offered by the MDM for some frequently occurring cases are conceivable.
- Marketing activities conducted by the MDM are realistic in the future. The MDM does not only wish to be a part of the mobility data ecosystem, but it also wishes to contribute to the onboarding of additional stakeholders. For maximum efficiency in terms of marketing, several important partners within the mobility data ecosystem should undertake joint steps.

21.4 Datenraum Mobilität (DRM): A National Implementation in Germany

In 2020, the German Federal Government has decided to implement and to promote the operation of a federated national Mobility Data Space “Datenraum Mobilität” (DRM), following the decentralized architecture principles of the here presented Mobility Data Space concepts. A large-scale stakeholder and governance process, led by acatech[3] (German Academy of Science and Engineering), has resulted in an extensive stakeholder engagement, supporting DRM by the provision of mobility data and the implementation of DRM-based use cases.

DRM will address the private and public sector equally in order to establish and promote a comprehensive mobility data ecosystem. A very important role will be played by existing data platforms (such as MDM, HERE), since they provide access to already connected participants and their data offers (Fig. 21.4):

[3] https://www.acatech.de/projekt/datenraum-mobilitaet/

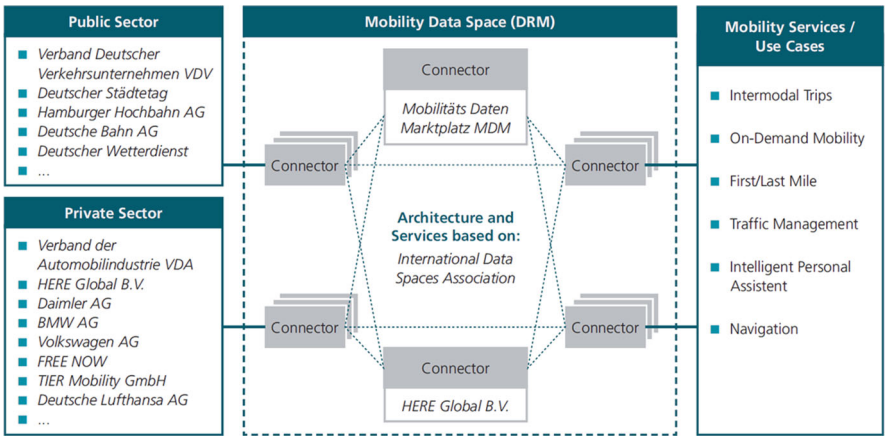


Fig. 21.4 Connecting participants and platforms (©2020, Fraunhofer IVI)

On behalf of the Federal Ministry of Transport and Digital Infrastructure (BMVI), acatech founded[4] the nonprofit organization "DRM Datenraum Mobilität GmbH" in May 2021, together with further supporting public and private shareholders. This entity will bring the DRM in operation and will be responsible for legal and governance aspects.

On a technical level, the DRM will provide the central services that are necessary for the operation of a data space according to IDSA: a data marketplace (technically, a metadata directory), a vocabulary provider, an identity provider, a data app store, and a clearing house (see also Sect. 21.2). The data exchange is established directly between the participants themselves in a distributed manner by using IDSA-compliant connectors. The DRM operator has no touching point with the exchanged data itself as proposed in Sect. 21.2, resulting in an opposite architecture and then a data platform/data lake.

The DRM services are based on reference implementations by Fraunhofer, following the IDSA specifications. The interim operation of DRM is provided by Fraunhofer IVI, until a professional operator takes over in 2022.

21.5 Connecting Data Platforms

The connection of several platforms will result in comprehensive visibility and availability of data sources for data users. Mobility data in particular are generated and used on a regional level, either by communities or by fleet operators in private economy.

[4] https://www.acatech.de/allgemein/vernetzter-verkehr-acatech-gruendet-traegergesellschaft-drm-datenraum-mobilitaet-gmbh-als-non-profit-organisation/

Currently, mobility data platforms are created on a regional level, e.g., by smart city initiatives, in order to pool the local services. Through the integration of these platforms and the data space concept, as well as through the resulting network, the MDM helps to make regional mobility data visible on a national level.

Further national data platforms with different focus topics, such as open data in mCLOUD, and commercial data services, such as geodata, vehicle data, or navigation services, can also be combined into an ecosystem with the help of data space concepts.

Cloud services provide a further level of networking. With the help of their resources, cloud services create scalability for business models in the field of data exploitation and management. It is thus possible to offer customized CPU-intensive prediction models, AI applications, and high volume data analyses, which would be impossible for just one conventional platform.

The use of resources and the resulting costs for cloud computing can be tied to customer demand, which means that they can be planned and calculated. Hosting an IDS connector in a cloud environment is just as secure as hosting it on a platform, the only difference being that a cloud-hosted connector is scalable according to the demand.

In addition to that, cloud environments, just as single platforms, are also often data and service ecosystems. Thus, the data offered by a connector is made available for additional interested parties within the cloud ecosystem.

The GAIA-X initiative, which is currently being promoted by the German federal government, gives an outlook on the prospective network of cloud ecosystems. The technological core of the initiative aims at connecting several European cloud environments with the help of data space concepts to form a connected infrastructure.

21.6 Application Example: “Mobility Service Provider”

The following example illustrates the potential of a mobility data ecosystem as pictured above including the MDM and additional decentralized stakeholders: A mobility service provider wants to offer short trips on dynamic routes. Their business model only works if they can serve a large number of customers per trip and direction.

For routing and achieving optimal travel times, they need traffic state information. This information is gathered by road operators, road administration offices, and environmental agencies through traffic monitoring and provided via the MDM. This is already a daily practice.

In order to achieve the highest possible occupancy rate of their vehicles, the service providers also need mobility data, movement data, and demand data.

Fleet operators (taxis, logistics, public transport) as well as providers of navigation services are already gathering floating car data (FCD) representing individual traveling speeds. This type of data is highly sensitive because it contains personal driving profiles. For this reason, the transfer of raw floating car data to third parties has been impossible so far.

In the depicted scenario, both the data provider and the MDM have an IDS interface (IDS connector). In this data space, the data provider can control how their sensitive floating car data may be processed by the MDM and in what shape they may be transferred from the data space to the data user after processing.

In doing so, data providers can offer their sensitive data for external business processes without the fear of unauthorized data exploitation for other purposes than originally intended.

It is also possible to transfer data directly to users without a central platform, such as the sensitive movement profiles gathered by telecommunications and transportation providers referred to in this example. In this case, the data is processed within the data users' IDS connector for utilization in their business processes and in compliance with the data providers' specifications.

Data processing (traffic data fusion) is realized by a data app whose compliance with the data providers' requirements has been verified by a certification agency. The app is run within the MDM's IDS connector. This app and other data apps may be developed by an independent software developer and offered in an app store. App development may be commissioned by data providers/data users, but it is also possible for stakeholders to develop apps on their own initiative with the aim of implementing a business model.

Data that has been enriched by an app is a potential new data source available to MDM users.

This way, data apps can become the basis for a novel mobility data ecosystem. The IDS's decentralized architecture allows the integration of further IT resources. In the above example (see Fig. 21.5), the architecture is extended by an external cloud environment that runs a more complex data app for the calculation of travel times and predictions (Fig. 21.6).

21.7 A Common Mobility Data Space: Outlook on a European Level

The European Commission has requested the establishment of National Access Points (NAPs) that are a prerequisite for the uniform handling of mobility data across Europe. The legal basis for this can be found in the ITS Directive no. 2010/40/EU. According to the directive, all member states are obligated to offer a platform on which at least the respective states' mobility data metadata description can be

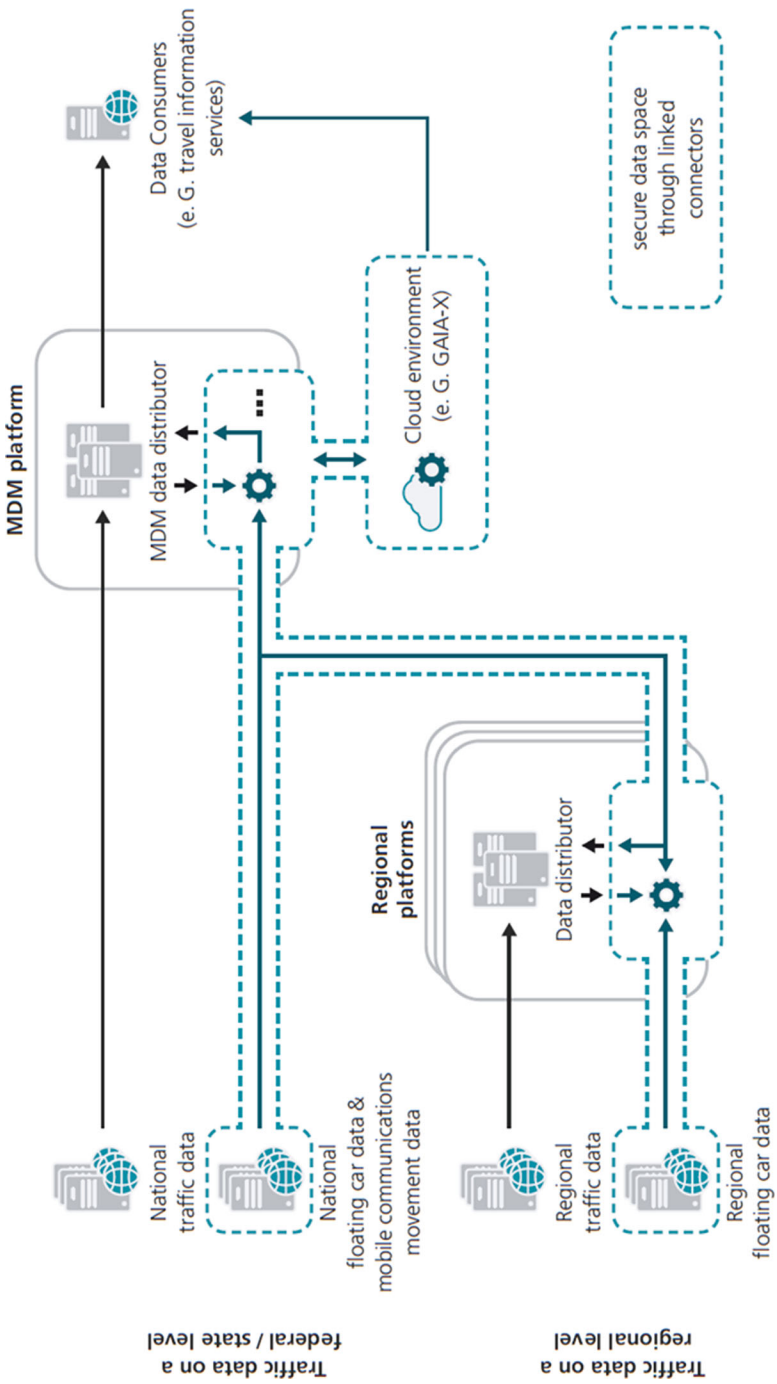


Fig. 21.5 Connecting regional data platforms with MDM and cloud environments (©2020, Fraunhofer IVI)

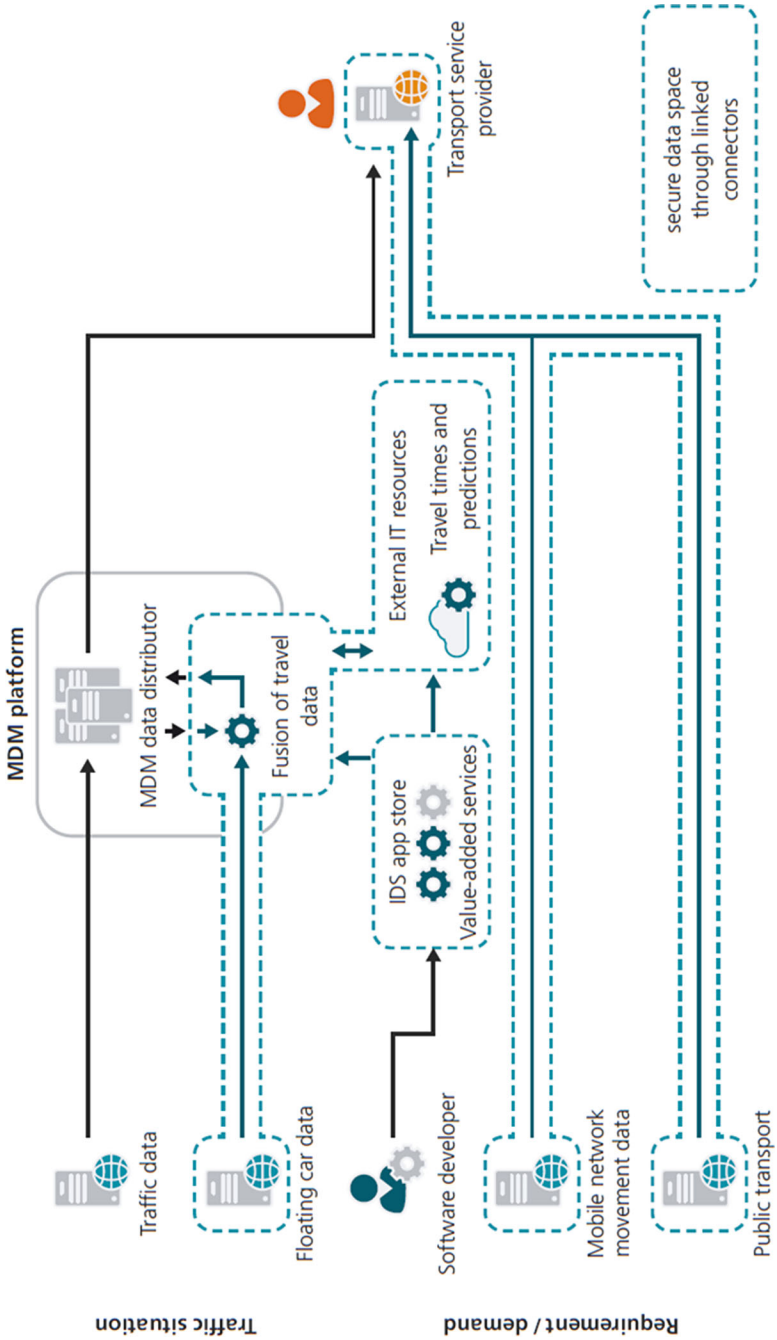


Fig. 21.6 Secure provision of mobility data for external business processes (©2020, Fraunhofer IVI)

published. In addition to the ITS directive, several delegated regulations specify the data providers' obligations to publish mobility data via the following NAP.

Safety-Relevant Traffic Information
According to the delegated regulation no. 2013/886, end users are to be granted free access to general safety-relevant traffic information (SRTI). This means that road operators in particular are obligated to provide existing data, e.g., on road works or exceptional weather conditions. This data is often used by service providers and forwarded to their clients.

Through increasingly connected vehicles, more and more private parties are in possession of safety-relevant information that can help, for example, to detect temporarily slippery roads. Because this data has the potential for commercial exploitation, there are reservations concerning its disclosure. The sharing of data in a secure data space can be a way to reduce these reservations.

Real-Time Traffic Information
The same applies to the provision of real-time traffic information (RTTI) across Europe in compliance with the delegated regulation no. 2015/962.

This regulation calls for the publishing of data on traffic volume and traffic jams, as well as dynamic speed limits and road closures via the NAP. In addition to road operators, this regulation also increasingly affects private parties with access to vehicle data.

Multimodal Travel Information Services
The delegated regulation no. 2017/1926 on the provision of multimodal travel information services (MMTIS) across the EU demands that both static and dynamic as well as historical travel and traffic data are to be published via the NAP by traffic authorities, transport providers, infrastructure operators, and providers of demand-based transportation services.

These multimodal travel planning and information services need to be linkable. This way, Europe-wide services can be created for end users.

Although the aforementioned legislative initiatives obligate private companies to provide data in high volumes, enterprises often fear the disclosure of business secrets and customer data. The sharing of sensitive data in a secure data space such as the Mobility Data Space will help alleviate these fears. Data providers can trust that their data will only be exploited according to the terms and conditions of use and licensing specified by them and that they will be able to control and monitor the usage.

Another obstacle for the utilization of European NAPs for internationally acting enterprises, such as vehicle manufacturers and navigation service providers, is the fact that there still are a large number of platforms in Europe. Ca. 30 NAPs, some of which differ significantly in the way they are implemented, need to be supplied in order to offer services internationally. The further harmonization—or, even better,

the connection of European NAPs through the concepts of the Mobility Data Space—would certainly be widely welcomed.

This can be the first step towards a common European mobility data space as envisioned within the Commission's COM 2020/66 data strategy. On the whole, the Mobility Data Space includes all concepts necessary to "facilitate access, pooling and sharing of data from existing and future transport and mobility databases."

A European Data Strategy
On February 19, 2020, the European Commission published the Communication 2020/66, introducing a European data strategy. This strategy explicitly promotes the creation of Europe-wide data spaces in different sectors including the mobility sector:

"[...] a Common European mobility data space, to position Europe at the forefront of the development of an intelligent transport system, including connected cars as well as other modes of transport. Such data space will facilitate access, pooling, and sharing of data from existing and future transport and mobility databases. [...]"[5].

Currently, it seems likely that this document will influence European legislation regarding the provision of data at National Access Points as well as different funding instruments.

21.8 Implementation Within mFUND Research Projects

The foundations for the development of the Mobility Data Space were laid in two mFUND projects funded by the Federal Ministry of Transport and Digital Infrastructure (BMVI): "Vorstudie-MDM-MDS" (12/2017 to 05/2018) and "MobilityDataSpace" (06/2019 to 05/2022).

Within the "Preliminary Study on Connecting the MDM to the Intended Mobility Data Space" (Vorstudie-MDM-MDS)[6] project, BASt, Fraunhofer IVI, and Fraunhofer IAIS developed potential improvements to the MDM through an integration concept for MDM and IDS components. The concept investigates different multimodal and intermodal mobility scenarios, considers the integration of open data from the mCLOUD, and illustrates potential contributions from the MDM/MDS that can help establish the future National Access Point for multimodal travel information.

[5] https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:52020DC0066

[6] https://www.bmvi.de/SharedDocs/DE/Artikel/DG/mfund-projekte/vorstudie-verknuepfung-des-mdm-mit-mds-mdmd-mds.html

This preliminary study is the basis for the implementation of the intended Mobility Space within the scope of a follow-up research and development project. It answers organizational, functional, and technical questions regarding the development, operation, and use of the Mobility Data Space.

In order to motivate relevant stakeholders to participate in the Mobility Data Space, the study's scientific results were presented at several relevant expert conferences such as the 2018 ITS World Congress, the MDM Conference, and the mFUND Conference, as well as a number of industry symposia.

The preliminary study resulted in a technical and temporal roadmap for the Mobility Data Space that will be updated and implemented within the mFUND "MobilityDataSpace" project.

The project "MobilityDataSpace: Connecting local, regional and national data platforms through data space concepts, as well as enrichment and exploitation as a mobility data ecosystem"[7] aims to initiate the development of the Mobility Data Space, which will establish itself as a mobility data ecosystem by including the Mobility Data Marketplace by BASt as well as additional regional traffic data platforms.

New local traffic data and nation-wide mobility data will be acquired and provided for secure and sovereign processing on platforms extended by data space concepts. By connecting regional platforms with the MDM, it will be possible to provide and exploit regional data on a national level.

Within the project, the MDM and further local platforms will be improved for the support of data-driven services. To achieve this, they will be expanded by a secure and protected execution environment for services and data apps in which mobility data can be provided and processed under guarantee of data sovereignty. This way, sensitive mobility data such as floating car data (FCD) will be exploitable for the first time.

By connecting the MDM with local platforms into a decentralized data space, a federal mobility data ecosystem will be created. With this ecosystem as a basis, complex real-time use cases can help lower environmental impact, optimize traffic flows, and improve multimodal commuter information services (Fig. 21.7).

[7] https://www.bmvi.de/SharedDocs/DE/Artikel/DG/mfund-projekte/mobility-data-space.html

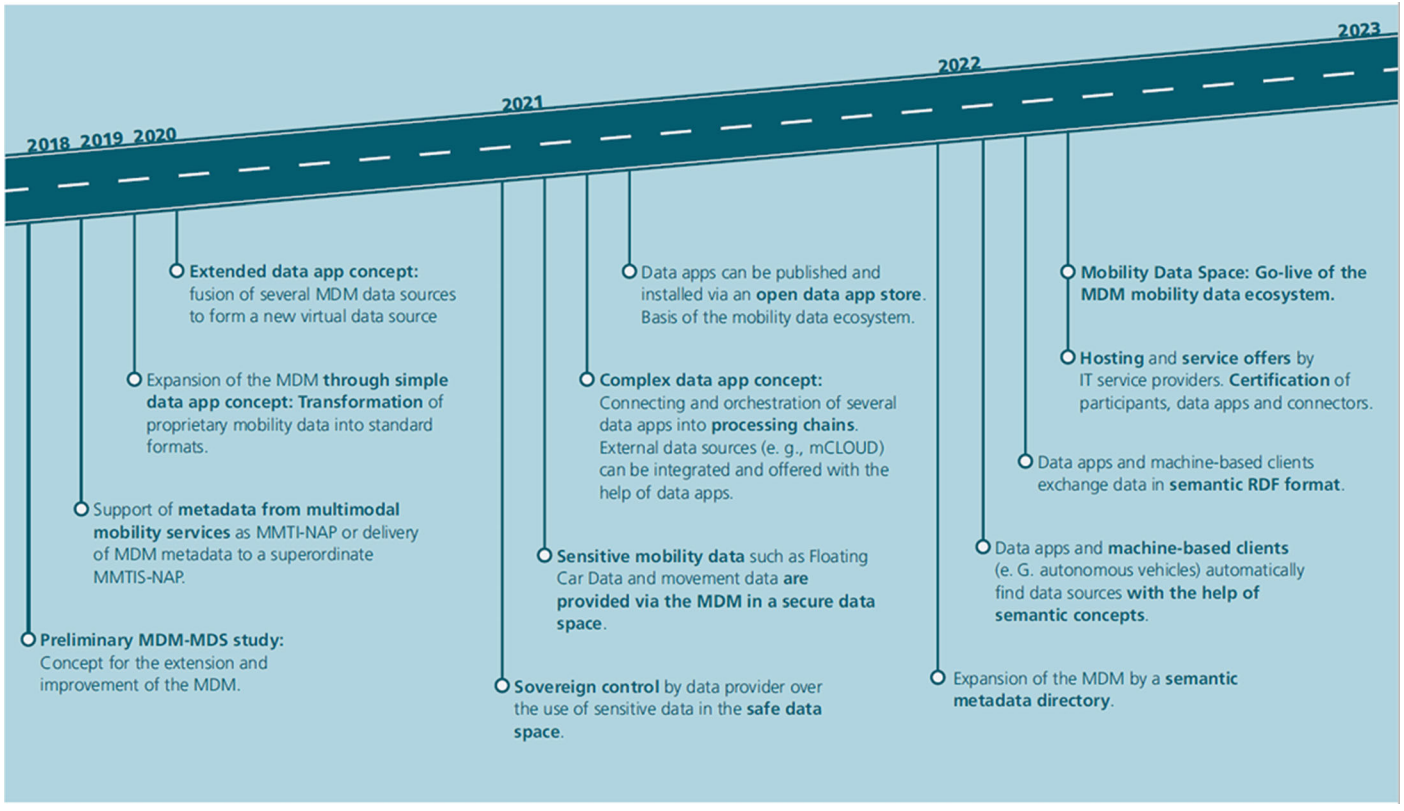


Fig. 21.7 Development roadmap for the Mobility Data Space. (©2020, Fraunhofer IVI)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part IV
Solutions and Applications

Chapter 22
Data Sharing Spaces: The BDVA Perspective



Edward Curry, Tuomo Tuikka, Andreas Metzger, Sonja Zillner, Natalie Bertels, Charlotte Ducuing, Davide Dalle Carbonare, Sergio Gusmeroli, Simon Scerri, Irene López de Vallejo, and Ana García Robles

E. Curry (✉)
Insight SFI Research Centre for Data Analytics, NUI Galway, Ireland
e-mail: edward.curry@insight-centre.org

T. Tuikka
VTT Technical Research Centre of Finland, Oulu, Finland
e-mail: Tuomo.Tuikka@vtt.fi

A. Metzger
Paluno, Universität Duisburg-Essen, Essen, Germany
e-mail: andreas.metzger@paluno.uni-due.de

S. Zillner
Siemens AG, Munich, Germany
e-mail: sonja.zillner@siemens.com

N. Bertels · C. Ducuing
KU Leuven, Leuven, Belgium
e-mail: natalie.bertels@kuleuven.be; charlotte.ducuing@kuleuven.be

D. Dalle Carbonare
Engineering Ingegneria Informatica SpA, Rome, Italy
e-mail: davide.dallecarbonare@eng.it

S. Gusmeroli
Politecnico di Milano, Milan, Italy
e-mail: sergio.gusmeroli@polimi.it

S. Scerri
metaphacts GmbH, Cologne, Germany
e-mail: simon.scerri@metaphacts.com

I. López de Vallejo
DisCO.coop, Bilbao, Spain
e-mail: Irene@disco.coop

A. García Robles
Big Data Value Association (BDVA), Bruxelles, Belgium
e-mail: ana.garcia@core.bdva.eu

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_22

Abstract Using data and Artificial Intelligence, it is possible to answer the big questions, how sustainable the planet is or what impact industry has on climate. The Big Data Value Association (BDVA) believes that Data Sharing Spaces will be a key enabler to this vision. The BDVA community has created a unified perspective on the value of data sharing spaces across the pillars of data, governance, people, organization, and technology, with trust as a central foundation. This chapter details this BDVA perspective, explaining the five pillars needed to create value in data with trust as a central concept, together with the tools and mechanisms for strategic stakeholders to create data sharing spaces jointly. It elaborates the strategic challenges which need to be overcome and sets out our call to action for the community to make this a reality. The chapter also summarizes the initial progress on data platform development, data governance, and Trustworthy AI to make data sharing spaces a reality. Finally, it details an example of a data space in smart manufacturing.

22.1 Introduction

Creating a European-wide Data Space is a very ambitious goal emerging from European stakeholders and visioned through many interventions during the last ten years. The quest for a European Digital Single Market, common Data Strategy, and a new kind of regulative framework based on European values are vehicles for a new kind of data economy enabled by data sharing infrastructures and the motivation to keep pace in the global competition. Data spaces are a must for Europe.

Simultaneously, another major technological breakthrough has happened. Artificial intelligence has become central as technology to facilitate the transformation of businesses exploiting digitalization. Also, in AI, European values are essential for technology development. Principles of privacy, security, and fairness are the basis of solutions where the European citizen, a human being, is center stage of society.

In a digital society, data is needed to improve industry performance and understand how sustainable our society is. Using data and AI, it is possible to answer the big questions, how sustainable the planet is or what impact industry has on climate.

This chapter delineates, starting from Chap. 2, the vision of Big Data Value Association on Data Sharing Spaces, explaining the five pillars needed to create value in data, trust as a central concept, and tools and mechanisms for strategic stakeholders to create data sharing spaces jointly. Section 22.3 elaborates on the strategic challenges which need to be overcome to realize the vision. Section 22.4 sets out our call to action for the community to make this a reality. Section 22.5 summarizes the initial progress on the development of a data platform to support data sharing. Sections 22.6 and 22.7 detail the importance of data governance and trustworthy AI. Section 22.8 details an example of a data space in smart manufacturing. Finally, Section 22.9 concludes the chapter.

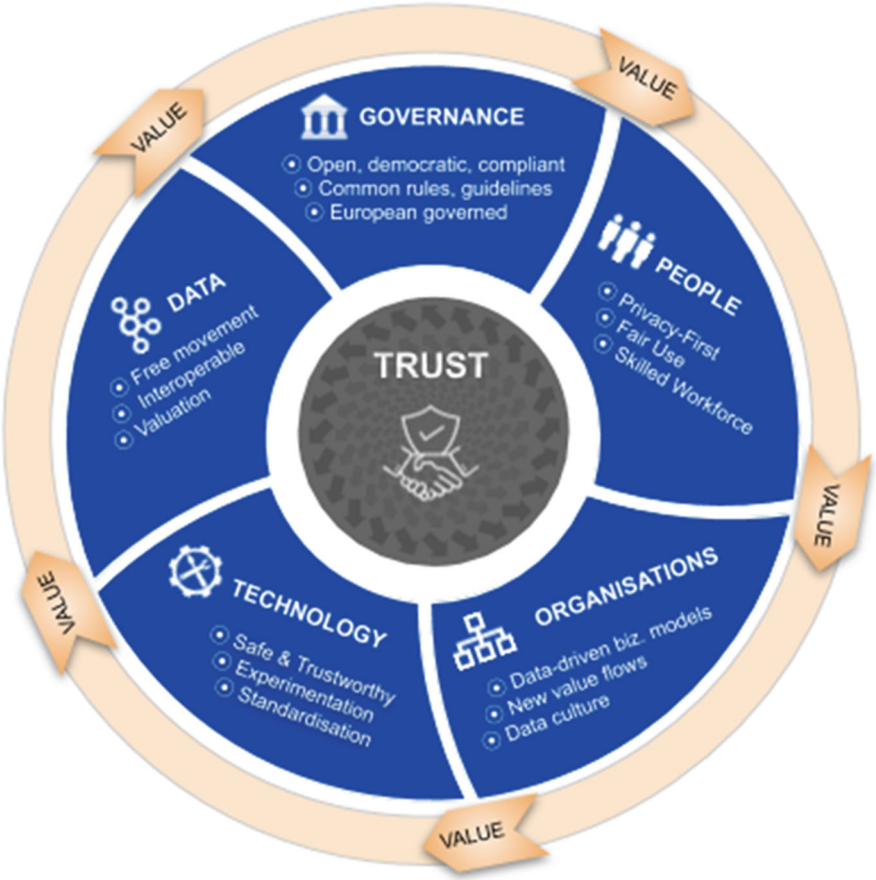


Fig. 22.1 The BDVA Data Sharing Value "Wheel" [1] ©2020, Big Data Value Association. Used under permission from Big Data Value Association

22.2 Vision

The realization of a functioning and frictionless European-governed data sharing space that can successfully generate economic value by broadening data access for AI relies on carefully planned iterative implementation strategies and a timely concerted effort between all relevant stakeholders [1]. As depicted in Fig. 22.1, the success of widespread data sharing activities revolves around the central key concept of trust: in the validity of the data itself and the algorithms operating on it; in the entities governing the data space; in its enabling technologies; as well as in and among its wide variety of users (organizations and private individuals as data producers, consumers, or intermediaries).

To achieve the required levels of trust, each of the following five pillars must meet some of the necessary conditions:

- **Data**—As a touted fifth European fundamental freedom, free movement of data relies on organizational data strategies that embed methodologies for data sharing by design (e.g., interoperability) and clear standard guidelines that help determine the market value of data assets.
- **Governance**—A European-governed data sharing space can inspire trust by adhering to the more advanced European rules, guidelines, and regulations and promote European values. Participation should be equally open to all and subject to transparent and fair rules of conduct.
- **People**—Data sharing needs to guarantee individual privacy and offer fair value or compensation for shared personal data. For Europe to drive data sharing activities, the European workforce needs appropriate reskilling and upskilling to meet the evolving labor market's needs.
- **Organizations**—More organizations (including business, research, and governmental) need to rethink their strategy to fully embrace a data culture that places data at the center of their value proposition, exploring new data-driven business models and exploiting new data value flows.
- **Technology** – Safer experimentation environments are needed to catalyze the maturation of relevant technology behind trustworthy data, data access, and algorithms (privacy, interoperability, security, and quality). Standardization activities need to adjust for faster reaction times to emerging standards and the identification of new ones.

The BDVA recognizes two complementary high-impact opportunities that can materialize as a result of timely interventions to converge data sharing initiatives in Europe and realize its vision:

- Achieve wider access to data to realize the full potential of emerging AI technology by designing and implementing a common, trustworthy, decentralized data space that enables safe and democratic data sharing and boosts the European data economy.
- Achieve a European-governed data space, giving Europe the possibility to assume a prominent position steering international efforts to develop data and AI solutions that reflect and respect European ethical values, including democracy, privacy protection, and equality [2].

The introduced vision for a European-governed data sharing space needs to be built around and in consultation with the same wide array of stakeholders that can exploit its benefits as users. Figure 22.2, designed on the triple helix view of research, development, and innovation production, shows the various roles that the active strategic stakeholders (Industry, Academia, Government) can play in the realization of this vision, the tools they can actively contribute, and the existing potential to achieve different kinds of societal impact (Economic, Technological, Political, and Cultural).

Rather than focusing on specific Business to Business (B2B) scenarios or restricting the vision to specific sectors, we envision a data sharing space that is open to all, thus offering equal opportunities and spanning all societal spheres,

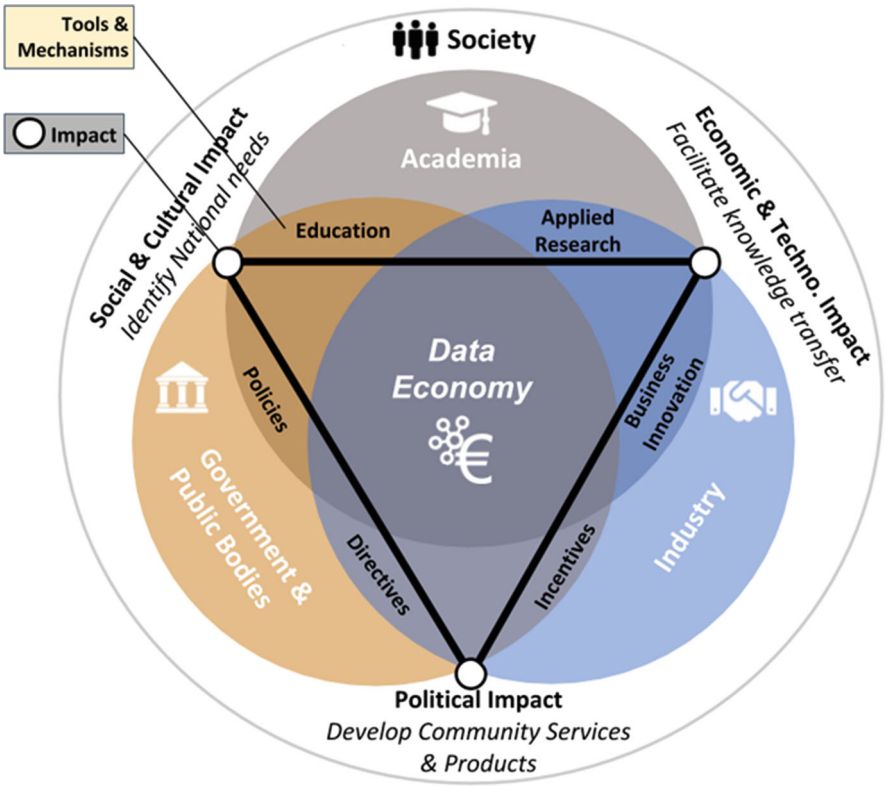


Fig. 22.2 Tools and mechanisms for strategic stakeholders to jointly realize a data sharing space [1] ©2020, Big Data Value Association. Used under permission from Big Data Value Association

including private citizens. Even though the latter are not actors in the realization of the data sharing space, they still play an essential role in data sharing. Although as the main economic driver we retain business at the center of our recommendations, in addition to B2B cases, we also consider Business to Governance and vice versa (B2G, G2B), Business to Science and vice versa (B2S, S2B), as well as Consumer to Business (C2B) opportunities.

22.3 Challenges

The BDVA community has identified the most critical challenges (see Table 22.1) that stand in the way of the expected value generated by the identified opportunities [1]. The challenges can be categorized into two main concerns: interorganizational (lack of suitable data sharing ecosystems) and intraorganizational (issues faced by data producers and consumers, as data sharing participants).

Table 22.1 Challenges for common European Data Sharing Spaces

Technical	**Business and organizational**
Sharing by design	EU values
Digital sovereignty	Global competition
Decentralization	Dynamic ecosystems
Veracity	Dynamic skills
Security	Digital transformation
Privacy protection	Trust
	Valuation standards
Legal compliance	**National and regional**
Data protection	Workforce skills
Free-flowing data	Resistance to change
Privacy preservation	Investment evaluation
Regulatory compliance	EU-wide policies
	Policy compliance

The most pressing interorganizational concern remains the lack of functional and trustworthy data sharing ecosystems that inspire immediate large-scale participation. Primary causes include the lack of robust legal and ethical frameworks and governance models and trusted intermediaries that guarantee data quality, reliability, and fair use. This is compounded by the lack of widespread adherence to emerging best practices and standards (e.g., interoperability, provenance, and quality assurance standards), whose maturity pace also continues to fail expectations. From a technical point of view, data sharing solutions need to address European concerns like ethics-by-design for democratic AI, and the rapid shift towards decentralized mixed-mode data sharing and processing architectures also poses significant scalability challenges.

In terms of intraorganizational concerns, a first significant concern is a difficulty to determine the value of data due to a lack of data valuation standards and assessment tools, compounded by the highly subjective and party-dependent nature of data value and the lack of data sharing foresight exhibited by a majority of producers. The second concern revolves around the difficulty faced by data producers balancing their data's perceived value (after sharing) against risks exposed (upon its sharing) despite adhering to standard guidelines. Specific examples include the perceived loss of control over data (due to the fluid nature of data ownership, which remains hard if not impossible to define legally), the loss of trade secrets due to unintentional exposure or malicious reverse-engineering (in a business landscape that is already very competitive), and the risk of navigating around legal constraint given potential data policies breaches (including GDPR and exposure of private identities).

22.4 Call to Action

BDVA has identified five recommended preconditions for successfully developing, implementing, and adopting a European Data Sharing Space [1]. Following widespread consultation with all involved stakeholders, the recommendations have been translated into 12 concrete actions. These can effectively be implemented alongside the Horizon Europe and Digital Europe programs [3]. This call for action is aligned with the European Commission's latest Data Strategy [4]. The recommended actions are categorized under five independent goals: Convergence, Experimentation, Standardization, Deployment, and Awareness, each of which is targeted towards specific stakeholders in the data sharing ecosystem. The implementation of the five goals should take place within the timeframe shown in Fig. 22.3. Assuming the convergence initiatives that are required over the next three years will yield satisfactory outcomes, deployment efforts can be scaled up with experimentation acting as a further catalyst. Other deployment efforts need to go hand in hand with intensified standardization activities, which are key to a successful European-governed data sharing space. Activities targeted at greater awareness for all end-users can initially target organizations, entities, and individuals that can act as data providers and then extend to all potential consumers as solid progress is achieved. The actions are targeted to specific actors, which map to one or more of the strategic stakeholders in Fig. 22.2.

22.5 Convergence: Data Platform Projects of the Big Data Value PPP

Trusted and secure platforms for secure sharing of "closed" personal and industrial data are key for creating a European data market and data economy. The data platform projects running under the umbrella of the Big Data Value PPP develop integrated technology solutions for data collection, sharing, integration, and exploitation to facilitate the creation of such a European data market and economy [5].

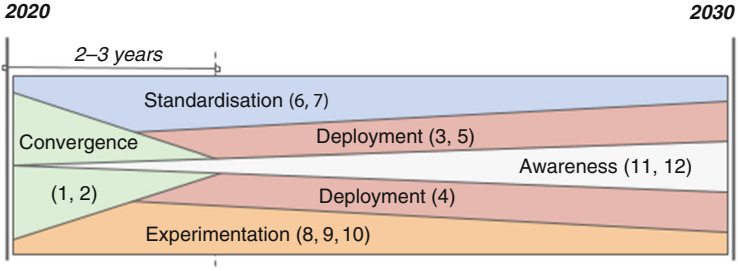


Fig. 22.3 Timeframe for implementing the recommended actions over the next decade [1] ©2020, Big Data Value Association. Used under permission from Big Data Value Association

The data platform projects fall under the following three main types:

- *Personal data platforms* facilitate respecting prevailing legislation and allow data subjects and data owners to remain in control of their data and its subsequent use. Personal data platforms preserve utility for data analysis and allow for the management of privacy versus utility trade-offs and metadata privacy and query privacy.
- *Industrial data platforms* facilitate trusted and secure sharing and trading of proprietary and commercial data assets. Industrial data platforms offer automated and robust controls on compliance (including automated contracting) of legal rights and fair remuneration of data owners.
- *Mixed data platforms* represent a combination of the above two types of data platforms.

22.5.1 The Portfolio of Projects

The portfolio of the Big Data Value PPP covers the data platform projects shown in Table 22.2.[1] This table gives an overview of these projects, the type of data platform they develop, and the domain, respectively, the use cases they address. Each of these projects is briefly summarized below based on open data from https://cordis.europa.eu/.

BD4NRG delivers a reference architecture for Smart Energy, which aligns with the BDVA, IDSA, and FIWARE reference models and architectures to enable B2B multi-party data exchange while providing full interoperability of leading-edge big data technologies with smart grid standards and operational frameworks. BD4NRG delivers an open modular big data analytic toolbox as front-end for one-stop-shop analytics services development by orchestrating legacy and third-party assets (data, computing resources, models, algorithms).

BD4OPEM develops an analytic toolbox to improve existing energy services and create new ones, all available in an open innovation marketplace. The analytic toolbox is based on big data techniques, providing tools for enabling efficient business processes in the energy sector. By extracting more value from available data, a range of innovative services are created in the fields of grid monitoring, operation, and maintenance, network planning, fraud detection, smart houses/buildings/industries energy management, blockchain transactions, and flexibility aggregation for demand-response. The open innovation marketplace ensures secure data flows from data providers to solution providers, compliant with GDPR requirements so that asset management is enhanced, consumer participation in energy balancing is promoted, and new data-driven business models are created through innovative energy services.

[1] Plus the coordination and support action "Data Market Services," an accelerator project aiming at overcoming the barriers of data-centric SMEs and start-ups.

Table 22.2 Portfolio of the Big Data Value PPP covering data platforms

Acronym	Name	Type of platform	Domains/use cases
BD4NRG	Big Data for Next Generation Energy	Industrial	Energy
BD4OPEM	Big Data for OPen innovation Energy Marketplace	Mixed	Energy
DataPorts	A Data Platform for the Cognitive Ports of the Future	Industrial	Transport and logistics
DataVaults	Persistent Personal Data Vaults Empowering a Secure and Privacy-Preserving Data Storage, Analysis, Sharing and Monetisation Platform	Personal	Sports, mobility, healthcare, smart home, tourism
i3-Market	Intelligent, Interoperable, Integrative and deployable open-source marketplace with trusted and secure software tools for incentivising the industry data economy	Industrial	Automotive, manufacturing, healthcare
KRAKEN	Brokerage and market platform for personal data	Personal	Education, health
MOSAICrOWN	Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and Owner control	Personal	Connected vehicles, finance, marketing
MUSKETEER	Machine learning to augment shared knowledge in federated privacy-preserving scenarios	Mixed	Smart manufacturing, healthcare
OpertusMundi	A Single Digital Market for Industrial Geospatial Data Assets	Industrial	Geospatial
PIMCITY	Building the next-generation personal data platforms	Personal	Generic
PLATOON	Digital PLAtform and analytic TOOls for eNergy	Mixed	Energy
Safe-DEED	Safe Data-Enabled Economic Development	Mixed	Marketing, manufacturing
SmashHit	Smart dispatcher for secure and controlled sharing of distributed personal and industrial data	Mixed	Connected cars, smart cities
SYNERGY	Big Energy Data Value Creation within SYNergetic enERGY-as-a-service Applications through trusted multi-party data sharing over an AI big data analytics marketplace.	Industrial	Energy
TheFSM	The Food Safety Market: an SME-powered industrial data platform to boost the competitiveness of European food certification	Industrial	Food supply chain
TRUSTS	Trusted Secure Data Sharing Space	Mixed	Finance, telecom

DataPorts designs, implements, and operates a cognitive ports data platform that (i) connects to the different digital infrastructures currently existing in digitized seaports, enabling the interconnection of a wide variety of systems into an integrated

ecosystem, (ii) sets the policies for trusted and reliable data sharing and trading based on data owners' rules and offering a clear value proposition, and (iii) leverages on the data collected to provide advanced data analytics services based on which the different actors in the port value chain can develop novel AI and cognitive applications.

DataVaults delivers a framework and a platform that has personal data coming from diverse sources in its center and that defines secure, trusted, and privacy-preserving mechanisms allowing individuals to take ownership and control of their data and share them at will, through flexible data sharing and fair compensation schemes with other entities (companies or not). The overall approach rejuvenates the personal data value chain into a multi-sided and multi-tier ecosystem governed and regulated by smart contracts, which safeguard personal data ownership, privacy, and usage and attributes value to those who produce it.

i3-MARKET addresses the growing demand for a single European data market economy by innovating marketplace platforms, demonstrating with industrial implementations that data economy growth is possible. i3-MARKET provides technologies for trustworthy (secure and reliable), data-driven collaboration and federation of existing and new future marketplace platforms, with special attention on industrial data and particularly on sensitive commercial data assets from both SMEs to large industrial corporations.

KRAKEN brings personal data sharing and trading to a level of maturity that does not yet exist by leveraging on (i) the emerging paradigm of self-sovereign identity built upon a stack of distributed ledger technologies (multi-ledger) which ensures future compatibility with different specific blockchain implementations for identity management. KRAKEN provides a decentralized user-centric approach to personal data sharing and incorporates trust and security assurance levels deriving claims from national identity schemas.

MOSAICrOWN enables data sharing and collaborative analytics in multi-owner scenarios in a privacy-preserving way, ensuring proper protection of private/sensitive/confidential information. MOSAICrOWN provides effective and deployable solutions allowing data owners to maintain control on the data sharing process, enabling selective and sanitized disclosure, and providing for efficient and scalable privacy-aware collaborative computations.

MUSKETEER creates a validated, federated, privacy-preserving machine learning platform tested on industrial data that is interoperable, scalable, and efficient enough to be deployed in real use cases. MUSKETEER alleviates data sharing barriers by providing secure, scalable, and privacy-preserving analytics over decentralized datasets using machine learning. Data can continue to be stored in different locations with different privacy constraints but shared securely.

OpertusMundi delivers a trusted, secure, and highly scalable pan-European industrial geospatial data market, acting as a single point for the streamlined and trusted discovery, sharing, trading, remuneration, and use of proprietary and commercial geospatial data assets. The OpertusMundi platform guarantees low cost and flexibility to accommodate the current and emerging needs of data economy stakeholders regardless of their size, domain, and expertise.

PIMCity delivers Personal Information Management Systems (PIMS) that give users back control over their data while creating transparency in the market. PIMCity implements a PIMS development kit (PDK) to commoditize the complexity of creating PIMS. This lowers the barriers for companies and SME to enter the web data market. In addition, PIMCity designs and deploys novel mechanisms to increase users' awareness.

PLATOON facilitates deploying distributed/edge processing and data analytics technologies for optimized real-time energy system management in a simple way for the energy domain expert. Data governance among the different stakeholders for multi-party data exchange, coordination, and cooperation in the energy value chain is guaranteed through IDS-based connectors. The PLATOON architecture and components are valuable for the different stakeholders of the energy sector value chain.

Safe-DEED brings together partners from cryptography, data science, business innovation, and the legal domain aiming to improve security in data sharing, increase trust, and promote privacy-enhancing technologies to conform with global macrotrends and the data economy. It also delivers a set of tools to ease the evaluation of data value in large companies motivating data owners to utilize the protocols developed by Safe-DEED.

smashHit assures trusted and secure sharing of data streams from both personal and industrial platforms, which is required to build sectorial and cross-sectorial services. smashHIT establishes a framework for processing of data owner consent and legal rules and effective contracting, as well as joint security and privacy-preserving mechanisms. The tools of this framework facilitate traceability of the use of data, data fingerprinting, and automatic contracting among the data owners, data providers, service providers, and users.

SYNERGY introduces a novel reference big data architecture and platform that leverages data, primarily or secondarily related to the electricity domain, coming from diverse sources (APIs, historical data, statistics, sensors/IoT, weather, energy markets, and various other open data sources). SYNERGY helps electricity stakeholders to simultaneously enhance their data reach and improve their internal intelligence on electricity-related optimization functions while getting involved in novel data (intelligence) sharing/trading models to shift individual decision-making to a collective intelligence level.

TheFSM delivers an industrial data platform that will significantly boost the way that food certification takes place in Europe. It brings together and builds upon existing innovations from innovative ICT SMEs to deliver a uniquely open and collaborative virtual environment that will facilitate the exchange and connection of data between different food safety actors interested in sharing information critical to certification. TheFSM catalyzes the digital evolution of the quite traditional but very data-intensive business ecosystem that the global food certification market involves.

TRUSTS facilitates trust in the concept of data markets as a whole via its focus on developing a platform based on the experience of two large national projects while allowing the integration and adoption of future platforms. The TRUSTS platform acts independently and as a platform federator while investigating the legal and

ethical aspects that apply to the entire data valorification chain, from data providers to consumers. TRUSTS delivers a fully operational and GDPR-compliant marketplace targeting both personal and industrial use by leveraging existing data marketplaces, such as the IDS.

22.5.2 Cross-Cutting Challenges in Data Platforms

A series of online workshops were organized by the BDVA and the EC under the umbrella of the European Big Data Value PPP to facilitate collaboration and interaction among the data platform projects. In addition to learning from each other about key aspects of the new data platform projects, these workshops facilitated identifying important transversal topics and challenges of common interest.

These challenges include:

- *Federation and interoperability of domain-specific data platforms:* Many of the existing data platform projects focus on distinct vertical domains (such as energy, health, or transport). While such focus is important to deliver data platform services that are meaningful to their users (as there are not too generic and thus decoupled from the actual domain semantics and data types), federation and interoperation among these vertical data platforms will facilitate a further push towards a European Data Market. Among the current data platform projects, I3-MARKET, MUSKETEER, and TRUSTS investigate paths towards achieving such federation and interoperability. Complementing these paths, the EUHUBS4DATA project aims to set up a European federation of Big Data Digital Innovation Hubs (DIHs), with the ambition of becoming a reference instrument for data-driven cross-border experimentation and innovation, and support the growth of European SMEs and start-ups in a global Data Economy.
- *Multi-sided aspects of data platforms:* To facilitate adoption and use of data platforms, their multi-sided aspects must be considered. To facilitate the adoption of data and service users, concerns such as attractiveness, ease of access, etc. are important. To attract data and service providers, concerns such as data protection and privacy, incentivizing mechanisms and strategies to attract data providers to platforms, as well as mechanisms to enforce or better incentivize data quality are important.
- *Commonalities among building blocks:* Many of the data platforms share common concerns with respect to data ingestion, sharing, protection, management, etc. As a result, many of the projects develop building blocks to facilitate these different concerns of a data platform project. Many of the data platform projects focus on domain-specific building blocks or building blocks for certain types of data (e.g., personal vs. industrial). This is an important step to delivering effective platform services. Still, there are domain-independent commonalities among these building blocks, which may be leveraged to identify commodity building blocks that facilitate more efficiently bootstrapping future data platforms as well

as federating data platforms. Identifying such common building blocks may also pave the way towards a common reference architecture for European data platforms.

22.6 The Needs for Data Governance

While originally mainly technical, the meaning of "data governance" has grown in scope over time. There is no consensual definition of "data governance," although the expression is broadly used. From the legal and policy side, governance generally refers to "the high-level management of organisations or countries, as well as the decision-making system and institutions for doing it." Data governance can therefore be defined (very broadly) as a system of rights and responsibilities that determine who can take what actions with what data.

The objective of developing integrated legal, ethical, organizational, and technical frameworks that can facilitate fair, compliant, trustworthy access to and reuse of data can only be reached by interdisciplinary efforts and involvement of a broad range of stakeholder's perspectives. The BDVA community and activities and the projects of the BDV PPP contribute to this R&I topic.

- The need for this forward-looking approach to data governance implies to advance on the following challenging questions: New regulations on substantive rights and organizational aspects are to be expected. However, we already have many legal frameworks dealing—more or less directly—with data and pursuing various legal and policy objectives. How to ensure that the new will interact with the existing legal frameworks is an important challenge that requires research in order to build consistent frameworks/infrastructure from the perspective of "data" and data spaces, to create a clear and fair legal ecosystem.
- To date, there is a tension between horizontal and sector-specific regulation of data. How to position the role of the data spaces in this tension field constitutes a challenge. In any case, data spaces should not result in (re-)creating "data silos" but work towards a genuine Common European Data Space. Data governance mechanisms can open avenues for regulating data while paying due consideration for the context in which they are processed. Based on case-specific studies, the factual and regulatory factors that influence data governance need further research to draw lessons for further regulation. Sandboxes and testbeds could be used to gain more insight into the concrete needs of stakeholders and, therefore, help solve this tension.
- It remains a challenge to safeguard a human-centric and fair approach. The "data control" paradigm should be further experienced in concrete settings to test to what extent it can serve to expand the data economy while protecting individuals with respect to data related to them and preventing commodification of their data. Data holders fear that they would "lose control" upon sharing "their" data. Therefore, empowering data holders (both legal entities and individuals) and

turning them into active economic players concerning "their" data is increasingly viewed by policymakers as a way forward. In this respect, it could also be considered that data do have not only economic value but also societal value, i.e., for research and for fighting against collective challenges.

- There is an important challenge in the facilitation of the creation of best practices of (sectorial) cases of specific (more collaborative) data governance mechanisms: data pools, data spaces, data commons, data trusts, data federations, data altruism, data cooperatives, data marketplace, PIMs, usage rights for co-generated data, etc. Therefore, an experimental research approach is needed to identify the factors for success or failure, e.g., the technology, the nature of data and stakeholders, the objectives assigned to the governance mechanism, and the legal framework.
- The data intermediation landscape is today in its infancy. It is, therefore, crucial to map the emerging data intermediation models (centralized, decentralized cf. innovative business models in H2020 R&I projects) and how they could be influenced by EU legislation (such as the DGA proposal or the upcoming Data Act). Data intermediaries should not be channelled into a given direction, as other (also more collaborative and decentralized) models should be allowed to emerge in line with the data space objectives.
- Technical and business models and architectures and also standards emerge to structure and facilitate data sharing. These initiatives are aimed at bringing trust to data holders and data users, which also requires another step, namely, *solid legal infrastructure*. The *interplay of technical and legal infrastructure* is a critical topic, e.g., legal layers of interoperability, the potential impact of fairness imperatives on the de facto appropriation of data, and the potential dynamic aspects of data that would require an evolutionary approach.

Results of these R&I efforts will create building blocks that stimulate the development of data spaces and the realization of trustworthy technology applications and ecosystems.

22.7 Towards Trustworthiness of Industrial AI

One of the core characteristics that AI-enabled industrial system needs to display is trustworthiness. The purpose of industrial AI is to boost the effectiveness and quality of the services delivered to the client and ensure that no negative impact is brought as a result of deploying AI solutions in critical applications. Building systems that can be trusted is critical to their acceptance.

Although only some critical AI application need high levels of trustworthiness, all applications need to be trustable. Trustworthiness will not be embodied by a single data source or technology but needs to be designed into an AI system, i.e., created by the interaction between all its technology building blocks and the data assets used [6]. Trustworthiness is built on multiple underlying system characteristics, such as reliability, dependability, safety, robustness, transparency, etc. In this context, high

data quality and efficient and transparent means for *data sharing* are key leverages to ensure the trustworthiness of industrial AI.

However, the trustworthiness of Industrial AI involves the simultaneous achievement of objectives that are often in conflict. For instance, one critical challenge stems from the ever-increasing collection and analysis of personal data and the crucial requirement for protecting the privacy of all involved data subjects as well as protecting commercially sensitive data of associated organizations and enterprises. As all means for the trusted and secure sharing add cost and complexity to Industrial AI systems, the optimal trade-offs without adding considerable complexity are significant research challenges to be addressed.

In this context, the BDVA community has identified several R&I challenges [3, 7] that need to be addressed when implementing the trustworthiness of industrial AI while respecting transparent and secure data sharing:

- To protect data for AI, improved privacy-preserving technologies are required. This includes *data protection* in machine learning, protecting the confidentiality and the integrity of training data, learned models and test samples, and means for data protection in dynamic environments (e.g., cloud/fog/Edge) with resource-constraint devices and immutable data stores.
- All *sensing and perception technologies* that help create, access, assess, convert, and aggregate signals representing real-world objects into communicable data assets need to be transparent, traceable, and reliable. In addition, the data processing and management methods need to ensure data privacy, integrity, and accountability. The development of trusted execution environments for edge devices keeps sensitive data within the source to achieve this.
- All *reasoning and decision-making* algorithms need to be transparent, explainable, and complement with efficient means for testing and validating the AI-based solution. This requires quality standards for reference datasets for the continuous testing and validation of the AI component performance in the context of the Industrial AI system, techniques that can work reliably with insufficient and missing data, as well as benchmarks for determining the performance, robustness, reliability, usability, and other quality indicators of industrial AI systems.
- For all industrial AI systems, safe *interaction* in safety-critical and unstructured environments needs to be ensured. This requires the co-development of technology and the development of data-based confidence measures.
- Deploying AI and Data systems often require the integration of diverse technologies ranging from software to hardware. Ensuring trustworthiness requirements, such as reliability, privacy, robustness, safety, dependability, transparency, etc., requires *data-driven methodologies and tools* as well as data-based validation processes and means for verification. This can be achieved by using data to identify hardware and software anomalies and new quality standards and methodologies to verify and "by-design" approaches.

22.8 Example: Smart Manufacturing Data Space

Data sharing spaces are developing for different sectors (i.e., Healthcare) and different sources of data (i.e., Internet of Things/Smart Environments [8]). The manufacturing sector, taking into account both discrete product and continuous process industries, is considering Data Spaces from a threefold perspective: availability of FAIR, high-value datasets, adoption of advanced AI-based Industrial Data Platforms, and deployment of governance rules to respect the business and security models of all the stakeholders [9]. On the one side, there is a continuously growing amount of data produced at all stages (e.g., factory, product, supply chain), while on the other side, analytical skills and tools are getting more and more advanced, mainly thanks to the adoption of Artificial Intelligence and data-intensive applications. The third challenging aspect is the governance of the match between the demand (e.g., manufacturing companies) and the offer (e.g., ICT service providers) to maximize the benefits obtained by the concrete application of AI technologies in real business cases while respecting security and confidentiality constraints. This match relies on the availability of a secure and trusted framework that enables the interaction among different subjects and the exchange of information, knowledge, and data among them, for example, via Data Sovereignty governance models.

Several Business Cases are available in the Smart Manufacturing Industry ecosystem. One of them is a use case developed in the MUSKETEER project, one of the research actions listed in Table 22.1. The case considers an industrial company (robot manufacturer) opportunity to set up a new business service based on the distributed knowledge available from the different customers they have. The scenario considers that a number of those customers may be interested in developing and deploying, on their own, an anomaly detection process for a defined robot-enabled operation (e.g., a welding action). Each one of those customers will collect data from the system, in the production line, and part of that data will be used to train a Machine Learning model able to detect if the welding operations are performing appropriately or not. In this case, the knowledge included in the ML model (the accuracy) is obtained considering just the cases that occurred in a single customer site. Assuming that different customers are operating the same kind of robots for the same kind of activity, they may face and record different correct and incorrect behaviors. Consequently, they can collect different training sets. By sharing such knowledge in a Manufacturing secure Data Space, a new "Fleet Management" application can be developed, which can consider a broader and more complete variety of customers behaviors and configurations.

The MUSKETEER platform federates the Data Spaces of the robot manufacturer and its clients. Starting from local ML models trained at customer sites, the robot manufacturer can aggregate them in a more accurate and high-value Data Space. From a service economy point of view, by implementing such a Manufacturing Data Space, the robot manufacturer can provide an added value service to all the clients in full respect of Data Sovereignty principles. The same paradigm is flexible enough to

be applied to other kinds of analysis (predictive maintenance) or other domains (healthcare images).

22.9 Conclusions

In this chapter, we presented a wealth of work done on data platforms in the form of European research projects, which educate us about European governed data sharing spaces. These concrete activities provide a corpus of knowledge, which can be used to derive best practices and ways to design data spaces.

Convergence of solutions, however, is necessary in order to build compatible data spaces. Therefore, while data space implementations are emerging, it is necessary to consider how different approaches and systems in respective domains are interoperable and if their data access solution is scalable. In general, European software producers must eventually align their offerings, and data space implementations need to follow similar design principles.

As we have seen, currently, many of the approaches are domain-oriented, but several cross-cutting challenges exist: developing domain-independent building blocks or how to federate among data sources. Cross-domain and cross-border data spaces are an opportunity that still requires further work. A cross-border use case successfully boosting the value network of European businesses would eventually show the potential of European data spaces. There is no easy way out. We need experimenting, piloting, and understanding how to navigate in the construction of European data spaces.

Acknowledgments We would like to thank the contributors from the BDVA Task Force on Data Sharing Spaces, including Simon Scerri (Fraunhofer), Tuomo Tuikka (VTT), Irene López de Vallejo (BluSpecs), Martina Barbero (BDVA), Arne Berre (SINTEF), Davide dalle Carbonare (Engineering), Oscar Corcho (UPM), Edward Curry (Insight Centre for Data Analytics), Valerio Frascolla (Intel), Ana García Robles (BDVA), Robert Ginthör (Know-Center), Sergio Gusmeroli (POLIMI), Allan Hanbury (TU Wien), Jim Kenneally (INTEL), Antonio Kung (TRIALOG), Till Christopher Lech (SINTEF), Antonis Litke (NTUA), Irene López de Vallejo (BluSpecs), Brian Quinn (INTEL), Dumitru Roman (SINTEF), Simon Scerri (Fraunhofer), Harald Schöning (Software AG), Tjerk Timan (TNO), Tuomo Tuikka (VTT), Theodora Varvarigou (NTUA), Ray Walshe (DCU), and Walter Weigel (HUAWEI). We would like to acknowledge the comments from BDVA members and external communities received.

References

1. Scerri, S., Tuikka, T., & Lopez de Vallejoan, I. (Eds.). (2020). *Towards a European data sharing space*. Report. Big Data Value Association.
2. Digital Single Market. (2019). Draft ethics guidelines for trustworthy AI | Digital Single Market. In *High-level expert group on artificial intelligence* (Issue December).

3. Zillner, S., Bisset, D., Milano, M., Curry, E., Hahn, T., Lafrenz, R., Liepert, B., Robles, A. G., Smeulders, A., & O'Sullivan, B. (2020). *Strategic research, innovation and deployment agenda - AI, data and robotics partnership*. Third Release (Third). BDVA, euRobotics, ELLIS, EurAI and CLAIRE.
4. Communication: A European strategy for data. (2020). https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf
5. Zillner, S., Curry, E., Metzger, A., Auer, S., & Seidl, R. (Eds.). (2017). *European big data value strategic research & innovation agenda*. Big Data Value Association. http://www.edwardcurry.org/publications/BDVA_SRIA_v4_Ed1.1.pdf
6. Curry, E., & Sheth, A. (2018). Next-generation smart environments: From system of systems to data ecosystems. *IEEE Intelligent Systems, 33*(3), 69–76. https://doi.org/10.1109/MIS.2018.033001418
7. Zillner, S., Gomez, J. A., García Robles, A., & Curry, E. (Eds.). (2018). *Data-driven artificial intelligence for European economic competitiveness and societal progress: BDVA position statement*. BDVA. www.bdva.eu/sites/default/files/AI-Position-Statement-BDVA-Final-12112018.pdf
8. Curry, E. (2020). *Real-time linked dataspaces*. Springer International Publishing. https://doi.org/10.1007/978-3-030-29665-0
9. Gusmeroli, S., & Dalle Carbonare, D. (2020). *Big data challenges in smart manufacturing industry*. BDVA.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 23
Data Platform Solutions



Fabrice Tocco and Laurent Lafaye

Abstract Private and public organizations hoard troves of data yet remain unable to unlock its full business potential. Data exchange platforms powered by adapted technology and driven by data exchange strategies act as catalysts to develop data ecosystems and data spaces, accelerate data circulation, and liberate its value.

Data spaces are powerful business, innovation, and societal enablers, whose growth and success rely on their ability to foster and develop trust.

Data exchange platforms contribute a lot to building trust as they provide the required tools and automation to data acquirers, data providers, and data exchange services providers to operate at scale within secure and compliant environments.

New and upcoming European regulatory frameworks also contribute to raising trust as they foster a harmonized data ecosystem across member states and define the rules of engagement between businesses, governments, and individuals engaged in data sharing and exchange.

Additionally, data exchange environments must provide traceability at all levels of the data transactions, which is particularly needed in increasingly complex data ecosystems.

Finally, in order to provide the flexibility required to answer the needs of complex, distributed, and heterogeneous environments, different models of data exchange governance are necessary.

23.1 Data Circulation: The Catalyst of Economic Value Creation

Data needs to circulate in order to liberate its full value potential. What does it mean?

Data has become a core asset in the economy, fostering new industries, processes, and products and creating significant competitive advantages.

F. Tocco · L. Lafaye (✉)
Place Louis Pradel, Lyon, France
e-mail: tocco@dawex.com; lafaye@dawex.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_23

Studies forecast the worldwide data valorization to reach $500B by 2022 and $708B by 2025 [1]. In the European Union, the data economy will soon contribute between 1.9% and 4% of GDP [2]. A Boston Consulting Group study shows that by generating just 1% of incremental revenue through data monetization, organizations can see their earnings increase by 10% and company valuation rise by more than 25% [3].

The need for data is skyrocketing, as is the volume being produced. Private and public organizations can take advantage of this mega-trend and develop data-driven strategies that are:

- **Impacting their corporate performance and competitiveness,** by optimizing business processes, creating new data-driven products and services, developing and training AI models, and generating new revenue streams through innovative data productization and distribution strategies,
- **Impacting the society as a whole**, by addressing more global objectives such as data altruism or leveraging data for decarbonation and other environmental use cases.

Innovation is greatly accelerated when data can be circulated, exchanged, and shared between organizations.

However, until now, the most useful data to achieve these goals are mostly industrial data, which has until now remained poorly accessible or not accessible at all, due to the lack of trustworthy frameworks allowing stakeholders—external to the organization that produced or collected the data—to access it.

The development of data ecosystems and data spaces will unlock the full potential of data.

Key requirements to be successful include:

- Achieving a high level of trust among all stakeholders involved in the exchange of data
- Building data spaces on the basis of solid and trusted data exchange environments
- Regulating data ecosystems through data exchange regulatory frameworks
- Providing traceability at all levels of data transactions
- Creating optimal conditions for data governance within complex, distributed, and heterogeneous environments

23.2 Trust as the Cornerstone of Data Exchanges

Data spaces' main challenge is to enable the fair and trusted circulation of data. Trust is the cornerstone of data exchanges, which must be achieved at both the operational level of the interactions between data providers and data consumers, and the more strategic and ecosystemic levels.

Trust is an essential component in conditioning the development of data spaces. It can be expressed on several levels:

- **Trust in the source of the data:** data acquirers engaging in data transactions want to be sure they can trust data providers, that they are acting in good faith. Trust will depend on many factors, including the reputation of the organization providing the data, its ability to deliver over time, the visibility offered on the origin of the data being exchanged, etc.
- **Trust in the acquirer and/or user of the data**: data providers are expecting data acquirers to act according to the agreed terms and conditions for the use of the data being exchanged. Trust typically increases if the data provider already knows the data acquirer or if it can access sufficient details about the acquiring organization's profile and identity before engaging in data transactions.
- **Trust in the use made of the data**: the use of clear and detailed legally binding contractual agreements, which can be based on open data licenses or commercial licenses, as well as the guarantee that the data transactions comply with regulations in force, contribute a lot to increasing trust in the use of the data.
- **Trust in the data exchange service providers:** Data providers and data acquirers increasingly engage and conduct data transactions via structured data exchanges, data marketplaces, or data hubs that act as trusted third parties facilitating data exchanges and offering all the guarantees of security and compliance to regulations. Those entities, also named *Data Intermediation Services* providers in the EU Data Governance Act, play an essential role in the building of trust.

23.2.1 A Solid Data Exchange Environment at the Heart of the Data Value Chain

A data exchange platform is materialized by the technology layer that provides tools and automation needed by data acquirers, data providers, and data exchange services providers to deploy their data exchange strategy.

Private corporations and public entities use data exchange platforms to:

- **Become more data-driven** and use data to build better products and services, and improve their business processes
- **Unlock the full value of their data** by distributing them (monetization or/and for free), in a controlled way, to other organizations
- **Position their organization strategically in emerging data ecosystems** by providing data exchange services to other organizations

Data exchange platforms are used by data providers and data acquirers to conduct data transactions smoothly and at scale. They can rely on automated features and capabilities for:

- Streamlining data discovery so that organizations spend less time searching for data and more time synthesizing valuable insights

- Facilitating the packaging of data products with the description in clear terms of the conditions of use of the data and the licensing terms
- Handling the technical exchange of the data through various file-based or API-based mechanisms

Data exchange platforms also include all the tools needed for data exchange service providers to orchestrate, industrialize, and automate the governance of the platform, and grow its usage and the number of business use cases leveraging the data exchanged on the platform. The platform supports and covers the main activities of the orchestrator, also called operator of the platform, and acts as the trusted third party, including:

- The administration of the data exchange platform and all its participants, namely the orchestrator's teams, data acquirers, and data providers, such as the registration, vetting, access rights, etc.
- The monitoring of the activity on the platform with automatic review of meta-information used by data providers to describe data offerings
- The stimulation of participants to exchange data using dedicated tools for automatically matching data supply and demand using algorithms, combined with notification systems
- The management of the business models applied to the use of the data exchange platform by the participant, which can combine free, transactional, and/or subscription-based approaches
- The production of various reports and metrics to conduct analysis on activity, behaviors, and trends, but also to enable the orchestrators to meet their obligations regarding interoperability and exercise of individual rights, that may be required by regulations.

Data exchange platforms play a critical role in the building of data spaces by clearly separating the roles and responsibilities of those who make data circulate from those who collect, store, and transform data (Fig. 23.1).

23.2.2 Regulating Ecosystems: Compliance at the Heart of Data Exchange

Institutions, governments, and regulators worldwide are playing a crucial role in boosting a sustainable data economy. By regulating data access and use, they contribute to **raising trust** among all types of organizations engaged in data sharing and data exchange activities (Fig. 23.2).

Europe is taking leadership in establishing data regulations. Since the General Data Protection Regulation (GDPR [4]) in 2018, which has inspired many similar other regulations around the world (the California Consumer Privacy Act (CCPA [5]), the Lei Geral de Proteção de Dados (LGPD [6]) in Brazil, or the Act on Protection of Personal Information (APPI [7]) in Japan, new European

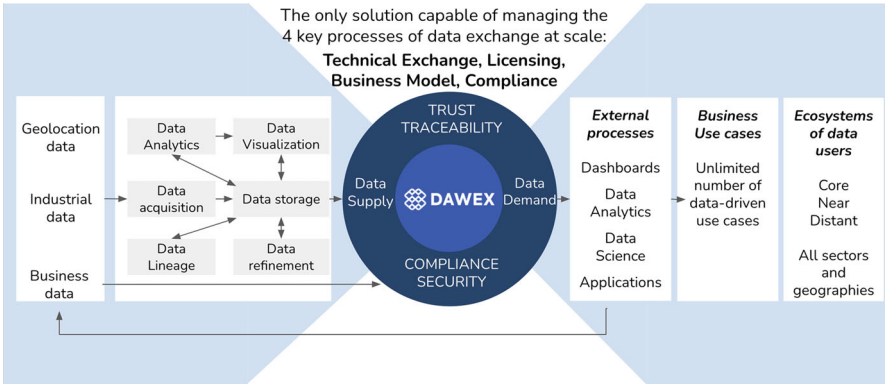


Fig. 23.1 Data exchange platforms at the heart of the data value chain (©2021, Dawex)

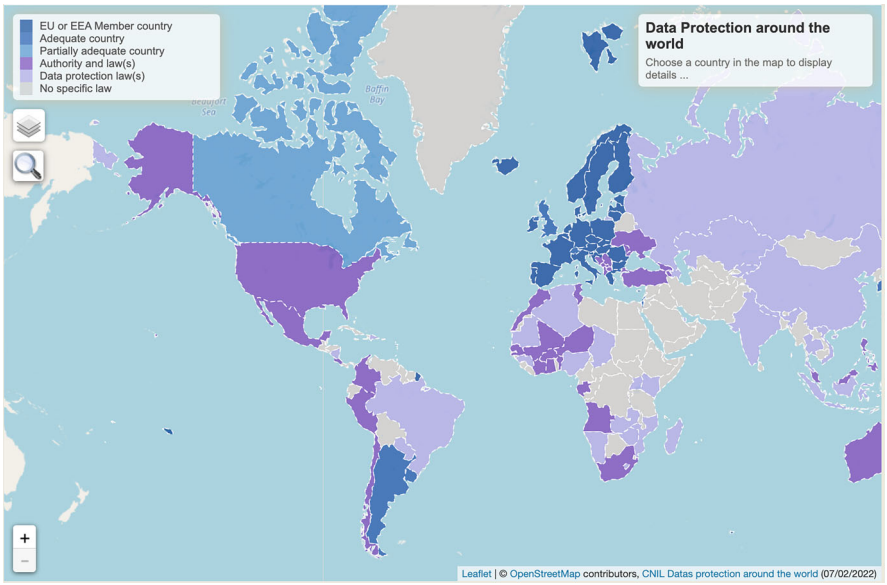


Fig. 23.2 Data protection around the world (©CNIL, 2022)

regulations have been proposed by the European Commission extending their scope beyond personal data to cover all types of data. In particular, industrial data has been identified by the European Commission as the next challenge in the creation of a single market for data that will allow it to flow freely within the EU and across sectors for the benefit of businesses, researchers, and public administrations. The President of the Commission, Ursula von der Leyen, declared in her political

guidelines for the 2019–2024 Commission that Europe must "balance the flow and use of data while preserving high privacy, security, safety and ethical standards."

The ensuing Commission Work Programme 2020 outlined several strategic objectives, including the **European Strategy for Data**, which was adopted in February 2020. The data strategy aims at building a genuine single market for data and at making Europe a global leader in the data-agile economy.

As part of the European Strategy for Data, the recently adopted proposal for a **Data Governance Act** aims to facilitate the voluntary sharing of data by individuals and businesses, harmonizes conditions for the use of certain public sector data, and **defines the requirements applicable to data intermediation services**. In particular, it stipulates among other conditions that data intermediation services providers may not use the data for which they provide data intermediation services for other purposes than to put them at the disposal of data users and shall provide data intermediation services through a separate legal person.

Complementing the Data Governance Act, the **Data Act**, whose proposal has been published by the European Commission in February 2022, aims at ensuring fairness in the allocation of value across the data economy. Future regulation will:

- Facilitate the access to and use of data by businesses by increasing legal certainty in the sharing or exchange of data, especially IoT data
- Provide for the public bodies' use of data held by enterprises in certain public emergencies
- Facilitate switching between cloud and edge services
- Provide for safeguards against unlawful data access by non-EU/EEA governments
- Provide for the development of interoperability standards for data to be reused between sectors
- Also additional initiatives are driven in Europe through the Gaia-X association looking at developing common requirements for a European data infrastructure, defining a reference architecture, and providing a secure, federated system that meets the highest standards of digital sovereignty while promoting innovation. Gaia-X also positioned data exchange as one of the four key pillars defining Gaia-X Federation Services.

23.3 The Need for Traceability

Traceability refers to the completeness of the information about every step in a process chain. It refers to the capability of an application to track and trace the state of objects, discover information regarding its past states and potentially estimate future states [8].

As data is being utilized in the context of specific projects or initiatives, knowing what data acquirers and data providers intend to do with the data really matters. If participants in a data ecosystem are seeking to leverage data circulation for the

common good, as part of a progressive project with a societal impact, or want to participate in the data economy, data exchange environments must provide them with traceability capabilities to reassure all participants that the data is being used in accordance with the initial agreement. This means data traceability is not only a cornerstone of trust but a crucial element to guarantee the business viability of all collaborative data projects.

In order to entirely fulfill their promise, data exchange environments must provide complete traceability of data to follow its pathway. Providing participants with traceability will generate trust in the exchange and in the environment. This is a key capability for any expanding data ecosystem. It enables participants to provide regulatory authorities with information on the source, nature, and development of all exchanges with the data ecosystem, which becomes increasingly time consuming and complex as ecosystems grow in size and activity. Traceability also provides huge data management benefits to participants engaging in numerous data exchanges with several different organizations, under different licenses. Data providers and data acquirers are empowered to optimize the orchestration of their data sharing and exchanges.

Data exchange platform environments can provide multiple capabilities to implement traceability, such as negotiation tools to find all interactions between data providers and data acquirers, the negotiation status, messages exchanged, or the date of all contacts. Functions to follow all legal documents related to the data transaction and the status of all data exchanges can also go a long way to provide increased visibility.

Multiple examples can be found to illustrate the urgent need to multiply traceability capabilities in all instances of data sharing and data exchanges, some of the most telling in the financial services sector. Stock exchanges are especially illustrative.

Financial markets are operated based on trust. Stock or commodities exchanges play a central role in creating trust, and traceability of all transactions that are taking place on those exchanges is vital for markets to operate properly, also allowing authorities to verify that trades are complying with regulations.

In the same way, proper traceability capabilities integrated into a complete and compliant data exchange framework will contribute to more efficient, trusted data exchanges, impacting directly the growth of such exchanges.

23.4 Data Exchange Governance: Toward Hybrid Data Exchanges Platforms

A data exchange platform is first and foremost a place where data supply and demand meet, securely and with confidence: the latter wanting to use or exploit data owned by the former.

Thus, specific conditions are essential to build a trustworthy digital ground for exchanging or accessing data such as:

- **Access conditions:** these conditions allow only legitimate participants to share or acquire data.
- **Usage rules:** enable control of the data for data providers through licenses that specify the criteria for use of the data or even the rights to sub-license the data, therefore providing legal coverage to limit misuse of the data.
- **Traceability of data circulation:** the origin of data must be traced, as well as all its circulation.
- **Data exchange governance:** this requirement is a means of governing the exchange of data, particularly with a view to the right of audit by regulators or authorities, by having mechanisms for data circulation, traceability of data exchanges, and data use.
- **Economic model:** this requirement reflects the commercial objective of the data exchange platform, which is to generate revenue for itself via its services. This objective is achieved through commissions earned on the activity taking place on the data exchange platform or through other additional means such as subscriptions to the platform.
- **Data Sovereignty:** the platform must have mechanisms that allow the data provider to control who can access its datasets and manage permissions, usage restrictions, and associated licenses.
- **Secure data exchange:** this is a requirement related to the most fundamental aspect of the data marketplace: data exchange. These exchanges must be carried out in the most secure way, as the data exchanged has a high commercial and strategic value. The disclosure of this data would reduce its value and lead to commercial losses, competitive movements, and reputational and regulatory impacts.
- **Compliance and privacy:** for data-driven innovation to thrive, it is essential that stakeholders share their data. To achieve this, the trust and security framework is key to both reassuring and providing technical guarantees of the security implemented.

The objective in addressing these challenges would, therefore, be to extend the traditional technological capabilities of data exchange platforms to develop a new type of solution that creates the optimal conditions for data governance within complex, distributed, and heterogeneous environments. Numerous technologies are already making it possible to envisage answers to these problems: deeptech, and particularly IoT and edge computing, for example. Nevertheless, many challenges exist, from monitoring millions of heterogeneous devices to secure data exchange between them.

Depending on the use cases, different modes can be provided:

- Managed mode: in this mode, the orchestrator acts as a trusted third party for the exchange of data. Data flows through the data exchange platform.

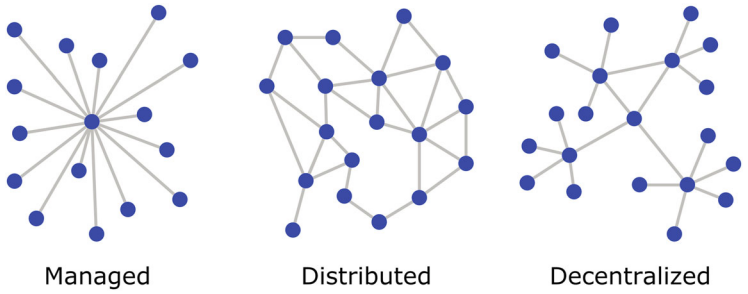


Fig. 23.3 Illustration of managed, distributed, and decentralized modes (©2021, Dawex)

- Distributed mode: in this mode, the data provider provides access to its data to the data acquirer when the transaction is finalized. Data flows peer-to-peer.
- Decentralized mode: in this mode, part of the business logic is delegated on nodes, through the use of agent or of smart contracts for example (Fig. 23.3).

These three modes do not oppose each other but complement each other, correspond to different use cases, and require specific technological responses. Data exchange platforms implementing these capabilities are called hybrid data exchange platforms.

Distributed ledger technologies (DLTs), such as blockchains, are powerful tools to reach decentralization needs. DLTs offer thus disruptive alternatives for these platforms in a peer-to-peer data exchange approach. The auditability of the information facilitates monitoring and allows anomalies to be identified. The blockchain also helps equipment interoperability by providing a reliable communication layer. In cases of large-scale use such as in smart cities, the blockchain and the IoT will work together.

To meet the requirements of distribution, decentralization, heterogeneity of actors, interoperability, protection of personal data, end-to-end encryption, and data governance requires finding the right balance between integrating distributed and managed technologies.

23.4.1 Innovative Nature of the Hybrid Approach

Indeed, there are different types of blockchains (public, private, pseudonyms, or with strong means of identification), using different consensus algorithms (proof of work, proof of stake, proof of participation, trusted execution environments), specific cryptographic means (including but not limited to asymmetric cryptography, zero disclosure proof of knowledge), different types of tokens (including but not limited to utilitarian, monetary, fungible or non-fungible), or different computer languages to write smart contracts. We can mention Ethereum, Hyperledger, Quorum, or

ZCash technologies. These technologies are mature and implemented by many different economic actors or different consortia such as the European EBP/EBSI initiatives.

While these blockchain technologies are mature, their large-scale implementations beyond the PoC (Proof of Concept) or PoB (Proof of Business) stage are still relatively limited, facing many bottlenecks such as interoperability, environmental efficiency, privacy, and user experience, among others. The creation of data marketplaces that are both distributed and encrypted from end to end may be required only when the nature of the players involved are very heterogeneous. Thus, in addition to offering interfaces and functionalities that facilitate data exchanges, this new operating mode must offer maximum efficiency, trust, traceability, and security. These conditions (efficiency, transparency, traceability, security) are essential to create an environment of trust that is indispensable to the commitment of stakeholders in the exchange of data and the establishment of a transaction between these players.

Combined with great flexibility supporting multiple business models, the hybrid approach allows managed, distributed, and decentralized use cases, from data sharing among the organization's business divisions to data exchange with external partners, leveraging free models, subscription-based access models, or pay-as-you-go models that charge a fee on each data transaction.

References

1. Transparency Market Research. (2018, April). *Data monetization market expected to reach US$ 708.86 Bn by 2025*. Transparency Market Research. https://www.transparencymarketresearch.com/pressrelease/data-monetization-market.htm
2. President Jean-Claude Juncker's State of the Union Address 2017. http://europa.eu/rapid/press-release_SPEECH-17-3165_en.htm
3. BCG. (2018). *How IoT data ecosystems will transform B2B competition*. https://www.bcg.com/fr-fr/publications/2018/how-internet-of-things-iot-data-ecosystems-transform-b2b-competition.aspx
4. The GDPR. https://eur-lex.europa.eu/eli/reg/2016/679/oj
5. The CCPA. https://oag.ca.gov/privacy/ccpa
6. The LGPD. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm
7. The APPI. https://www.cas.go.jp/jp/seisaku/hourei/data/APPI.pdf
8. Integrated approach to traceability data management. http://ceur-ws.org/Vol-963/paper4.pdf

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 24
FIWARE for Data Spaces



Ulrich Ahle and Juan Jose Hierro

Abstract This chapter describes how smart applications from multiple domains can participate in the creation of data spaces based on FIWARE software building blocks. Smart applications participating in such data spaces share digital twin data in real time using a common standard API like NGSI-LD and relying on standard data models. Each smart solution contributes to build a complete digital twin data representation of the real world sharing their data. At the same time, they can exploit data shared by other applications. Relying on FIWARE Data Marketplace components, smart applications can publish data under concrete terms and conditions which include pricing or data usage/access policies.

A federated cloud infrastructure and mechanisms supporting data sovereignty are necessary to create data spaces. However, additional elements have to be added to ease the creation of data value chains and the materialization of a data economy. Standard APIs, combined with standard data models, are crucial to support effective data exchange enabling loose coupling between parties as well as reusability and replaceability of data resources and applications. Similarly, data spaces need to incorporate mechanisms for publication, discovery, and trading of data resources. These are elements that FIWARE implements, and they can be combined with IDSA architecture elements like the IDS Connector to create data spaces supporting trusted and effective data sharing.

The GAIA-X project, started in 2020, is aimed at creating a federated form of data infrastructure in Europe which strengthens the ability to both access and share data securely and confidently. FIWARE is bringing mature technologies, compatible with IDS and CEF Building Blocks, which will accelerate the delivery of GAIA-X to the market.

U. Ahle (✉) · J. J. Hierro
FIWARE Foundation, e.V., Berlin, Germany
e-mail: ulrich.ahle@fiware.org; juanjose.hierro@fiware.org

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_24

24.1 Introduction to FIWARE

FIWARE[1] was created with the ultimate goal of creating an open sustainable ecosystem around public, royalty-free, and implementation-driven software platform standards easing the development of smart solutions and supporting organizations in their transition into smart organizations. From a technical perspective, FIWARE brings a curated framework of open-source software components which can be assembled together and combined with other third-party platform components to build platforms easing the development of smart solutions and smart organizations in multiple application domains: cities, manufacturing, utilities, agrifood, etc. Since the creation of the FIWARE Foundation in late 2016, a vibrant FIWARE Community has been formed with a true worldwide dimension, comprising more than 90 member organizations,[2] including large corporations, SMEs, technology centers and universities, and hundreds of individual members (Fig. 24.1). Parallel to this growth, the number of organizations adopting FIWARE has not stopped growing.

Any software architecture "powered by FIWARE'' gravitates around management of a digital twin data representation of the real world. This digital twin data representation is built based on information gathered from many different sources, including sensors, cameras, information systems, social networks, end users through mobile devices, etc. It is constantly maintained and accessible in near real time ("right time" is the term also often used, reflecting that the interval between the instants of time at which some data is gathered and made accessible is enough short to allow a proper reaction). Applications constantly process and analyze this data (not only current values but also history generated over time) in order to automate certain tasks or bring support to smart decisions by end users. The collection of all digital twins modelling the real world that is managed is also referred to as *Context*, and the data associated with attributes of digital twins is also referred to as *context information*.

In FIWARE, a digital twin is an entity which digitally represents a real-world physical asset (e.g., a bus in a city, a milling machine in a factory) or a concept (e.g., a weather forecast, a product order). Each digital twin:

- Is universally identified with a URI (Universal Resource Identifier).
- Belongs to a well-known type (e.g., the Bus type, of the Room type) also universally identified by an URI.
- Is characterized by several attributes which in turn are classified as:
  - Properties holding data (e.g., the "current speed" of a Bus, or "max temperature" in a Room).
  - Relationships, each holding a URI identifying a third digital twin entity the given entity is linked to (e.g., the concrete Building where a concrete Room is located).

[1] https://www.fiware.org/

[2] https://www.fiware.org/foundation/members/

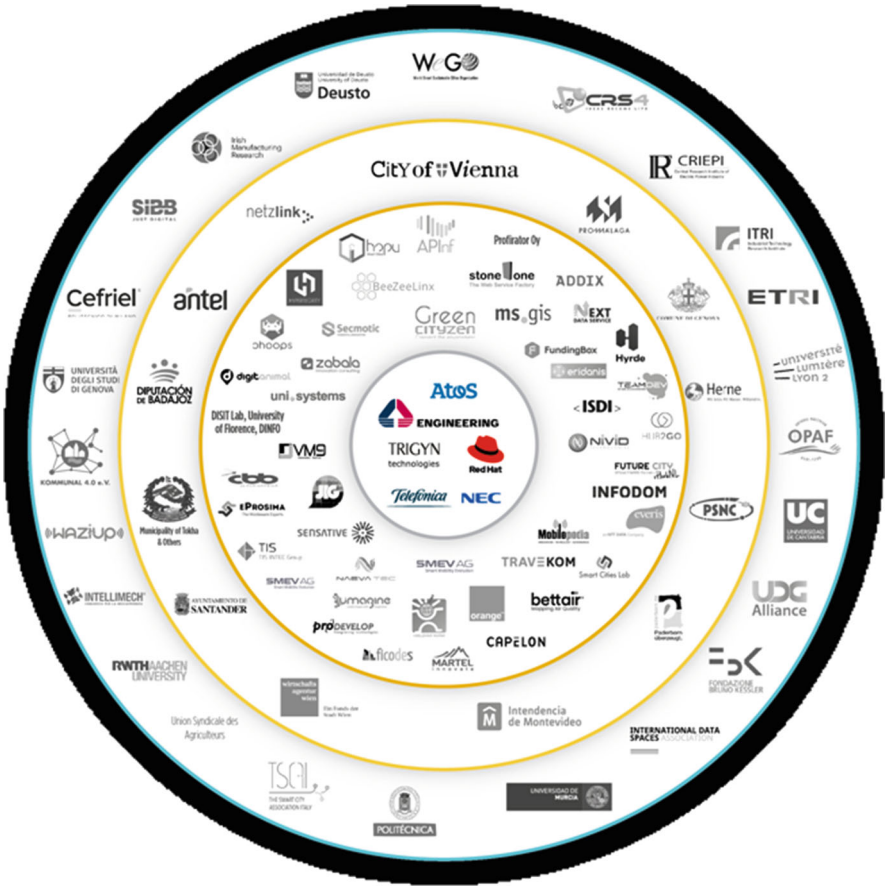


Fig. 24.1 Current member organizations in the FIWARE community (©2021, FIWARE)

Attributes of a digital twin may vary ranging from attributes that are quite static (e.g., the "license plate" of a Bus) to attributes that change very dynamically (e.g., the "speed" or "number of passengers" in a Bus) to attributes which still change but not that often (e.g., the "driver" in a Bus which may change twice a day). Very important, the attributes of a digital twin are not only limited to observable data but also inferred data. Thus, for example, the digital twin of a Street may not only have an attribute "current traffic", which may be automatically measured through sensors or cameras, but an attribute "forecasted traffic in 30 minutes" which might be calculated based on AI algorithms that keep the value of this attribute updated based on current traffic data, other relevant data impacting traffic (e.g., current weather observed and forecasted, schedule of events nearby, etc.), and historical information about traffic in the given street. Therefore, the digital twin data representation of the world that is managed in a "powered by FIWARE" architecture is expected to contain all the information needed by smart applications, not only

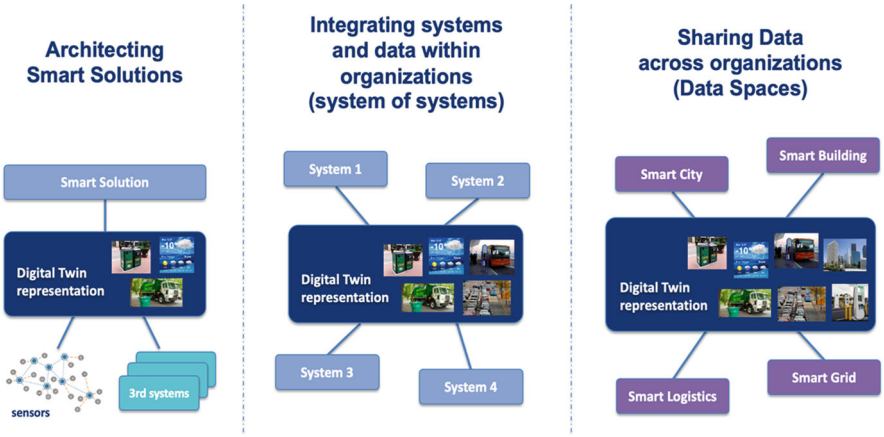


Fig. 24.2 Levels of integration supported following a digital twin approach (©2021, FIWARE)

measurable data but also other augmented insights and knowledge acquired over time.

A digital twin approach provides the basis for data integration at different levels, as illustrated in Fig. 24.2:

- Within a *vertical smart solution*, solving how main building blocks within the architecture can be integrated.
- Within a *smart organization*, bringing support to the integration of the different systems within a smart organization following a system of systems approach.
- Within a *smart data space*, establishing the basic "common lingua" that systems linked to the different organizations speak and understand.

Two critical elements need to be standardized in order to support an effective data integration at these three levels: the API to get access to digital twin data and the data models describing the attributes and semantics associated with the different types of digital twins being considered. The FIWARE Community has driven and continues to drive standardization at both fronts:

- The *NGSI API* provides a simple yet powerful RESTful API for getting access to context/digital twin data. NGSI API specifications have evolved over time driven by feedback from developers and implementation experiences. A first mature version of the API is the *NGSIv2 API*, which was defined by members of the FIWARE Community and is currently used in many systems in production within multiple sectors. Evolution of the API has taken place within the *ETSI CIM ISG*[3] (Context Information Management Industry Specification Group), where members of the FIWARE Community and the FIWARE Foundation have led the definition of an evolved version of the API, known as the *NGSI-LD API*, whose

[3] https://www.etsi.org/committee/cim

specifications were first published by ETSI in 2019 and continue to evolve.[4] The NGSI-LD API is used as the data integration API and is implemented by the core component of any "powered by FIWARE" architectures: the so-called Context Broker component. Different alternative open-source implementations of a Context Broker are available within the FIWARE Community, namely, the Orion-LD, Scorpio, and Stellio products.

- The *Smart Data Models initiative,*[5] launched by the FIWARE Foundation, provides a library of Data Models described in JSON/JSON-LD format which are compatible, respectively, with the NGSIv2/NGSI-LD APIs as well as any other RESTful interfaces compliant with the Open API specification. Data Models published under the initiative are compatible with schema.org and comply with other existing de facto sectoral standards when they exist. They solve one major issue developers are facing like it is the fact that a given data model specification may be mapped into JSON/JSON-LD in many different ways, all of them valid. Thanks to the Smart Data Models initiative, developers can rely on concrete mappings into JSON/JSON-LD, compatible with the NGSIv2/NGSI-LD APIs, which are made available within this library, avoiding interoperability problems derived from alternative mappings. Since its creation, more than 500 data models have been published, and the number of organizations contributing data model descriptions is constantly growing. Relevant organizations like TM Forum[6] or IUDX[7] are joining forces with the FIWARE Foundation bringing support to an open governance model for the initiative, following best open-source practices.

Building around the FIWARE Context Broker, which is the core mandatory component in any "powered by FIWARE" architecture, a rich set of complementary open-source components are available listed as part of the FIWARE Catalogue.[8] These components can be classified in the following categories or chapters:

- Components easing development of interfaces with the Internet of Things, Robotic, and third-party systems.
- Components supporting context/digital twin data processing and monitoring or the connection with data processing engines (e.g., Apache Spark, Apache Flink), monitoring tools (e.g., Grafana), and analysis platforms (e.g., Apache Superset).
- Components covering aspects related to Identity and Access Management (IAM) as well as Publication and Monetization of Data (including data accessible via APIs like NGSI-LD).

[4] https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.03.01_60/gs_CIM009v010301p.pdf

[5] Github: https://github.com/smart-data-models; website: https://smartdatamodels.org/

[6] https://www.tmforum.org/

[7] https://www.iudx.org.in/

[8] https://github.com/FIWARE/catalogue

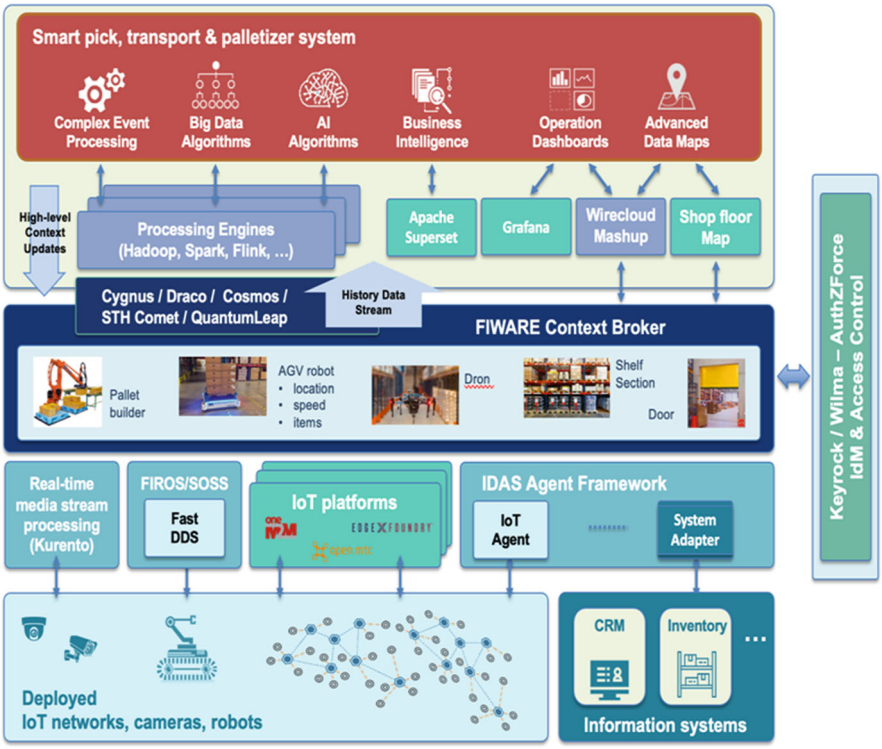


Fig. 24.3 Smart solution for picking and palletizing products from a warehouse (©2021, FIWARE)

Figure 24.3 depicts the reference architecture of a vertical smart solution powered by FIWARE. The concrete example corresponds to a smart solution for picking and palletizing products from a warehouse using robots. This reference architecture is structured in essentially three layers:

- A Context Broker component is at the core of the architecture, keeping a digital twin representation of the real-world objects and concepts relevant to the specific problem tackled: AGV robots, palletizer robots, shelf sections where products are stored in the warehouse, automatic doors AGV robots have to pass, operators in the shopfloor, items of stored products, orders generated from the CRM system, etc.
- Southbound to the Context Broker, the NGSI IoT Agents, available as part of the FIWARE IDAS framework, are used for connections to robotic systems exporting the OPC-UA IoT protocol or to specific sensors or actuators, used, for example, to detect items in shelf sections or to be able to open the shop floor doors. They perform the necessary conversions between IoT protocols and NGSI. In addition, System Adapters developed based on the IDAS Agent library cope with the connection to the CRM and the Warehouse Inventory Management

system that the solution has to interface with. FIWARE components like FIROS/SOSS allow, on the other hand, to perform the adaptation to robotic systems based on ROS/ROS2. Last but not least, the FIWARE component Kurento is able to process the video streams of cameras deployed in the shop floor, which are helpful to detect potential obstacles or risky situations.

- Northbound to the Context Broker, a number of tools are targeted to support real-time big data processing of the streams of history data generated as context/digital twin information evolves over time. A combination of third-party open-source components (Apache Superset, Grafana) and FIWARE components (e.g., Wirecloud) is shown in the picture targeted to support the creation of operational dashboards and advanced data maps for monitoring processes. A number of FIWARE Data Connectors (Cygnus, Draco, Cosmos, STH Comet, QuantumLeap) are available as part of FIWARE to facilitate transference of historic context/digital twin information to these tools.

Transversal to all these layers, a number of FIWARE components support Identity and Access Management. They control the flow of data across the different layers. With regard to the access to the Context Broker, they enforce the policies establishing what users can update, query, or subscribe to changes on context/digital twin data. Note that the flow of data is not only south to north in the picture. Northbound applications can perform updates on context data, which in turn will trigger changes in the devices, robots, or systems that are connected southbound.

An important point to highlight is that FIWARE is not about taking it all or nothing. You are not forced to use all the complementary FIWARE components mentioned above, but you are free to use other third-party platform components as well to design the hybrid platform of your choice. Thus, for example, you may opt to use a concrete IoT platform instead of IDAS IoT Agents to interface with sensors and actuators as reflected in the picture. As long as it uses the FIWARE Context Broker technology to manage context information, your platform can be labeled as "Powered by FIWARE" and solutions build on top as well.

Figure 24.4 depicts the reference architecture of a smart city powered by FIWARE. Again, the Context Broker component is at the core of the architecture, holding a digital twin representation of the real-world objects and concepts and describing what is going on in the city: streets, waste bins and containers, waste trucks, buses, electric vehicle chargers, buildings, events, citizen claims, etc. The different vertical smart solutions deployed in the city (e.g., Air Quality Monitoring, Smart Traffic Management, Smart Parking, Smart Waste Management) are connected to the Context Broker contributing that information they manage which is relevant for creating a holistic Context/Digital Twin representation of the whole City, thereby breaking the information silos. Some of these vertical smart solutions may be powered by FIWARE (e.g., Traffic Control and Waste Management systems in the figure) in which case their interface with the global city-level Context Broker does not require any adaptation. Others may not be powered by FIWARE, but this doesn't represent a major problem because creation of NGSI system adapters which translate from whatever API those systems export to NGSI-LD has proven not to be

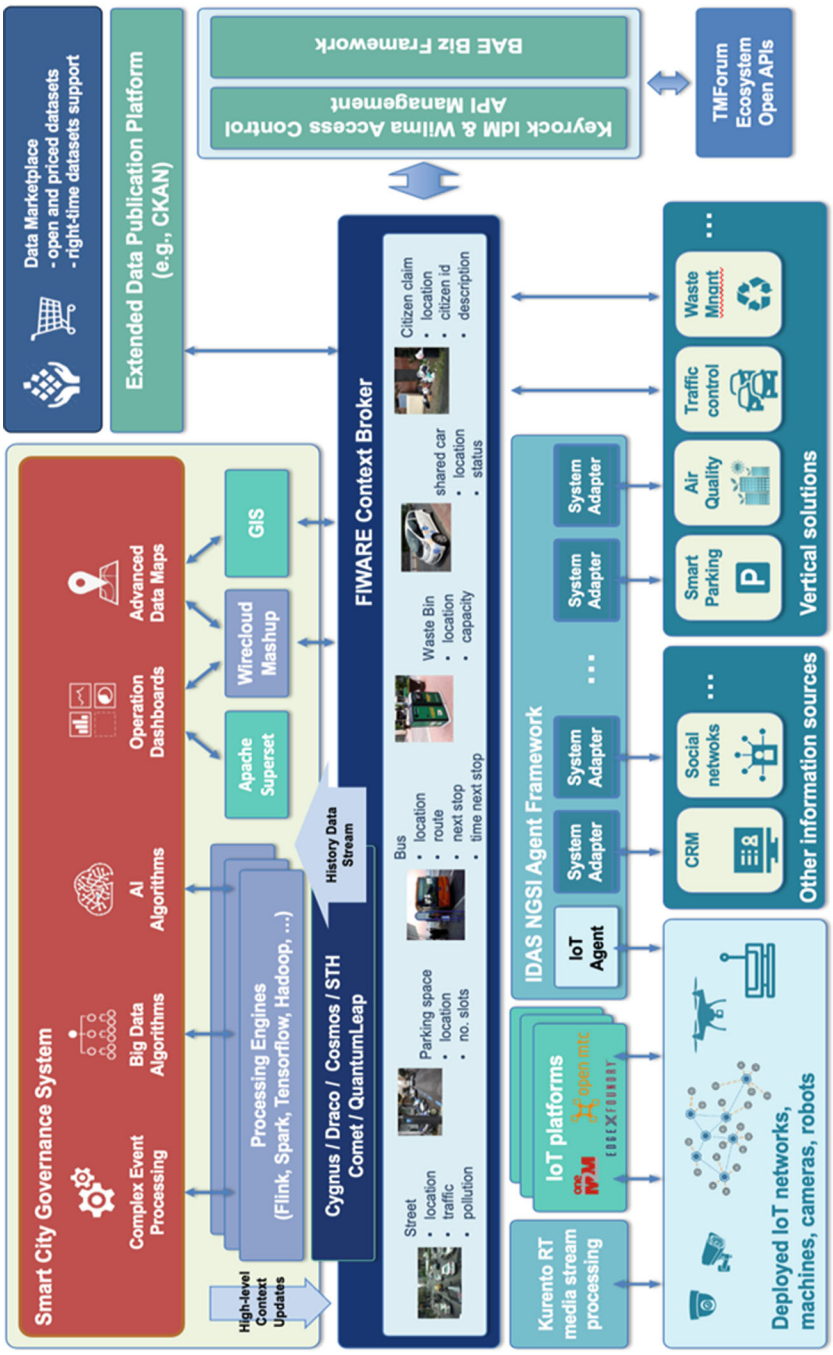


Fig. 24.4 Smart city reference architecture (©2021, FIWARE)

difficult. Last but not least, the City may deploy sensor/camera infrastructures through which valuable data is extracted.

Exploiting the complete Context/Digital Twin representation of the City, the Smart City Governance System (or City Operation Center) can be developed. Real-time big data processing tools can be used relying on data coming from multiple sources, extracting more valuable insights for the support of decisions. Similarly, monitoring tools can leverage in this holistic Context/Digital Twin representation of the City.

So far, we have described in more detail how FIWARE components can be used to achieve the two first levels of data integration using a Digital Twin approach (see Fig. 24.2). Within the next section, we will elaborate on how FIWARE also helps achieve the third level of data integration, linked to the creation of Data Spaces as a natural extension of the first two. Actually, in Fig. 24.4, part of the Context/Digital Twin data representation of the city can be made available to data spaces, published through data marketplaces.

24.2 FIWARE and Data Spaces

A *data space* can be defined as a decentralized data ecosystem built around commonly agreed building blocks enabling an effective and trusted sharing of data among participants. From a technical perspective, a number of *technology building blocks* are required ensuring:

- *Data interoperability*—Data spaces should provide a solid framework for an efficient exchange of data among participants, supporting full decoupling of data providers and consumers. This requires the adoption of a "common lingua" every participant uses, materialized in the adoption of common APIs for the data exchange, and the definition of common data models. Common mechanisms for traceability of data exchange transactions and data provenance are also required.
- *Data sovereignty and trust*—Data spaces should bring technical means for guaranteeing that participants in a data space can trust each other and exercise sovereignty over data they share. This requires the adoption of common standards for managing the identity of participants, the verification of their truthfulness, and the enforcement of policies agreed upon data access and usage control.
- *Data value creation*—Data spaces should provide support for the creation of multi-sided markets where participants can generate value out of sharing data (i.e., creating data value chains). This requires the adoption of common mechanisms enabling the definition of terms and conditions (including pricing) linked to data offerings, the publication and discovery of such offerings, and the management of all the necessary steps supporting the life cycle of contracts that are established when a given participant acquires the rights to access and use data.

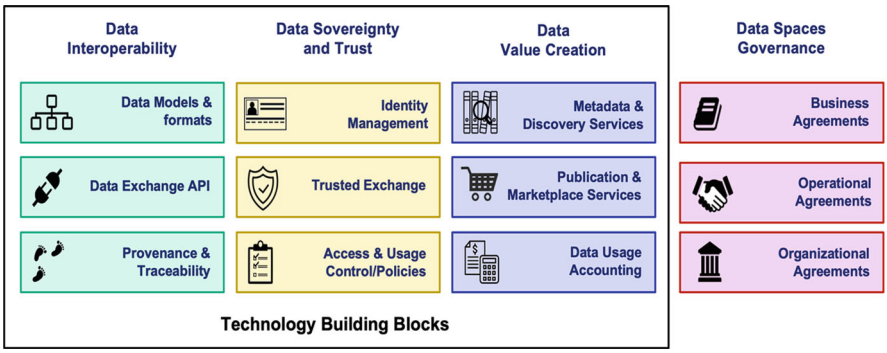


Fig. 24.5 Building blocks in a data space (©2021, FIWARE)

Besides the adoption of a common technology foundation, data spaces also require *governance*, that is, the adoption of a number of business, operational, and organizational agreements among participants. Business agreements, for example, specify what kind of terms and conditions can regulate the sharing of data between participants and the legal framework supporting contracts established through the data space. Operational agreements, on the other hand, regulate policies that have to be enforced during data space operation like, for example, compliance with GDPR (General Data Protection Regulation) or the second Payment Services Directive (PSD2) in the finance sector. They may also comprise the definition of tools that operators of cloud infrastructures or global services supporting data spaces must implement enabling auditing of certain processes or the adoption of cybersecurity practices. Last but not least, organizational agreements establish the governance bodies (very much like ICANN for the Internet). They deal with the identification of concrete specifications that products implementing technology building blocks in a data space should comply with, as well as the business and operational agreements to be adopted. The complete taxonomy of building blocks required for creating data spaces is illustrated in Fig. 24.5.

Sharing of data within a given data space should not be limited to a single domain. This would severely limit the creation of new innovative services since individuals and organizations usually act in multiple domains at the same time and many opportunities will flourish when data generated within organizations operating in a certain domain (e.g., management of traffic in cities) is shared for its exploitation in processes relevant to other domains (continuing with the example, logistics). Therefore, technology building blocks for data spaces must be domain-agnostic. On the other hand, they should rely on open standards, allowing multiple infrastructure and global service providers to emerge and support data spaces, without getting locked in any particular provider. Given this, while making things work in living labs and pilots is relatively easy, the main challenge toward definition of successful data spaces is the decision of what concrete standards and design principles are adopted, since they have to be accepted by all participants.

The following sections elaborate on the different components FIWARE brings materializing the different technical building blocks required for creation of data spaces.

24.2.1 Data Interoperability

Data providers joining data spaces must be able to publish data resources at well-defined endpoints knowing that data consumers, unknown by them a priori, will know how to retrieve and consume data through those endpoints. Data consumers, on the other hand, must know how data available through endpoints they discover can be consumed. This is a key principle which was observed in the design of the World Wide Web: content providers publish web pages on web servers (endpoints) knowing that web browsers will be able to connect to them and retrieve web pages whose content they can render and display to end users. It means that all participants in data spaces should "speak the same language," which translates into adopting domain-agnostic common APIs and security schemas for data exchange (the way of constructing sentences) together with data models represented in data formats compatible with those APIs (the vocabulary used in constructed sentences).

The *NGSI API* is domain-agnostic. Actually, many different systems have been developed using NGSI in domains such as smart cities, smart manufacturing, smart energy, smart water, smart agrifood, smart ports, or smart health, to mention a few. This facilitates data sharing because each system participating in a data space will be publishing data that simply enriches a digital twin data representation of the world that the rest of the systems connecting to the data space will know how to access. Systems participating into the data space don't know a priori what other systems may consume the data they publish (although they will be able to set up concrete terms and conditions for accessing/using data as we will explain in the next section).

Figure 24.6 illustrates how different systems participating in data spaces "powered by FIWARE" will exchange data. Context Broker servers are the endpoints through which systems connected to the data space publish digital twin data, very much like web servers publish html content on the World Wide Web. Those systems can in turn connect to Context Broker servers in order to obtain information they need. Note that data spaces powered by FIWARE enable near-real-time (right-time) exchange of digital twin data which is fundamental in the design of innovative value chains demanding a very dynamic exchange of data among participants. Just think about scenarios like a city managing traffic lights in streets close to a given train station in order to facilitate that travelers arriving and taking a taxi can leave faster to their destinations. NGSI-LD brings very simple therefore easy-to-use operations for creating, updating, and consuming context/digital twin data but also more powerful operations like sophisticated queries, including geo-queries, or the subscription to get notified on changes of digital twin entities. On the other hand, data spaces "powered by FIWARE" can also support the exchange of large files

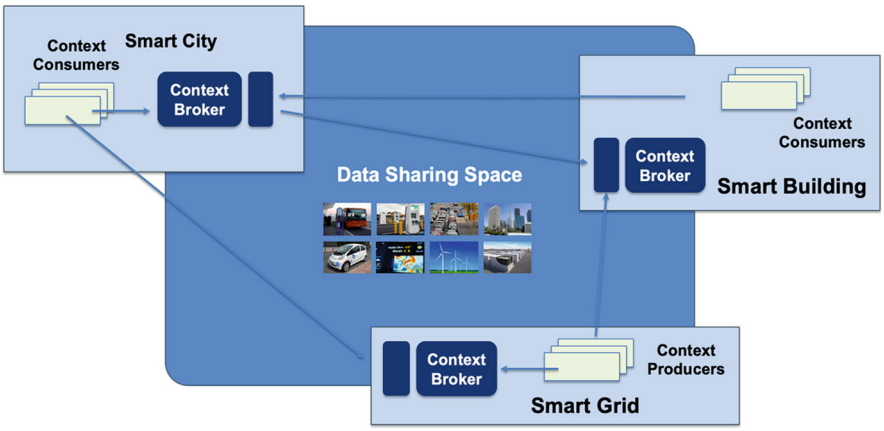


Fig. 24.6 Data exchange in a data space "powered by FIWARE" (©2021, FIWARE)

using standard file transfer protocols, since this kind of file transferences may be required for certain scenarios like training of AI algorithms.

Note that systems participating in data spaces "powered by FIWARE" do not need to be themselves "powered by FIWARE." Systems which have not been architected using FIWARE can still use the NGSI API to share data they produce and consume data they need in the form of data associated with attributes of digital twin entities which represent that part of the world they deal with. This can be done directly by the systems or through NGSI system adapters which have been programmed to perform a conversion between NGSI and the API that the system natively supports for managing data.

From a theoretical perspective, systems connected to a data space should be able to share data using the API they prefer. The specification (information model) of each API could be published as some kind of manifesto that certain components, integrated as part of the platform that systems should use to connect to the data space, can dynamically process in order to perform an automated adaptation from/to the APIs. However, such an approach faces important challenges. In the first place, such an approach has only been demonstrated in very simple scenarios and involving very simple APIs. Thus, the ability to exploit the kind of sophisticated features NGSI-LD would support for accessing digital twin data will be rather limited. On the other hand, creating a "common lingua" has proven to work for creating the kind of ecosystems we pursue: we shall ask ourselves if the World Wide Web had experienced the speed on adoption reached if HTTP and HTML hadn't been adopted as "common lingua" for web servers and browsers and each web server had the ability to choose a different protocol (as opposed to HTTP) or a different document format (as opposed to HTML). NGSI-LD has the advantages of being an open standard (defined by ETSI), has a strong open-source community behind (FIWARE), and, quite relevant within Europe, it will pave the way for alignment with developments in the Connecting Europe Facility (CEF) program. Creation of system adapters that

transform from specific APIs a given system may still require to use from/to NGSI-LD has proven to be not a complex task.

As mentioned before, the NGSI-LD API is domain-agnostic; therefore, it is designed to work for any type of digital twin. Consequently, achieving full interoperability requires also the adoption of common data models to be represented in formats compatible with the APi. Here, the *Smart Data Models initiative,*[9] already introduced in the previous section, reaches great relevance. It brings a powerful resource for developers who can rely on the way data model specifications are mapped into concrete JSON and JSON-LD structures under the initiative, compatible with NGSIv2 and NGSI-LD, respectively.

Figure 24.7 illustrates how resources are organized within the Smart Data Models initiative on GitHub. Data models are grouped into "subjects" (weather, parking, aquaculture, etc.) which in turn are referred from repositories associated with the multiple application domains being considered (Smart Cities, Smart Agrifood, Smart Manufacturing, Smart Water, Smart Energy, etc.). Note that there are subjects which are very specific to a given application domain (e.g., "street lighting" with regards to smart cities and communities), while others may be relevant to multiple domains (e.g., "weather" that is relevant to almost every domain or "sewage" that is relevant to the Smart Cities and Smart Water domains). Published data models are open to contributions and royalty-free.

An open governance model has been defined for the Smart Data Models initiative defining the life cycle of data models comprising incubation of brand new data models as well as curation of data models via harmonization of different contributions. Processes and procedures for management of the different activities follow best practices from open-source communities, guided by principles of transparency and meritocracy.

Completing the picture of building blocks for Data Interoperability, FIWARE brings components which provide the means for tracing and tracking in the process of data provision and data consumption/use. It provides the basis for a number of important functions, from identification of the provenance of data to audit-proof logging of NGSI-LD transactions. For those data spaces with strong requirements on transparency and certification, FIWARE brings components (i.e., Canis Major) that ease recording of transactions into different *Distributed Ledgers/Blockchains*.

24.2.2 Data Sovereignty and Trust

Data spaces must provide means for guaranteeing organizations joining data spaces that they can trust the other participants and that they will be able to exercise sovereignty on their data. That requires the definition of common building blocks,

[9] https://github.com/smart-data-models

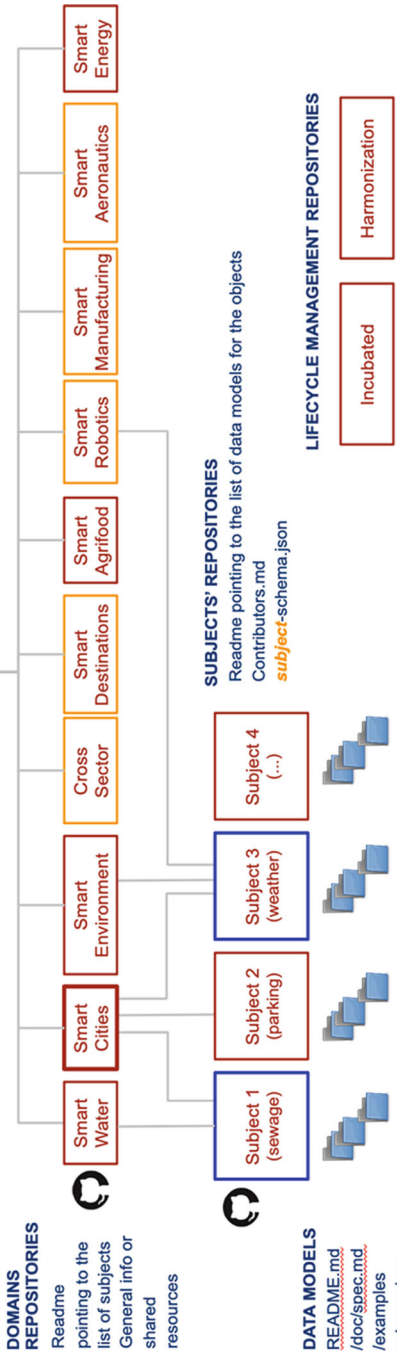


Fig. 24.7 Smart data models organization on GitHub (©2021, FIWARE)

based on mature security standards that will be used by all participants in the data space.

A first fundamental building block to support within data spaces has to do with *Identity Management (IM)*. This building block allows identification, authentication, and authorization of organizations, individuals, machines, and other actors participating in a data space. While this building block can be implemented on the basis of third-party open-source technologies like KeyCloak,[10] just to mention one popular example, FIWARE brings the *Keyrock* component which supports *OpenIdConnect,*[11] *SAML 2.0,*[12] and *OAuth2*[13] standards. Quite relevant for data spaces deployed in Europe, Keyrock also resolves integration with *eIDAS,*[14] a building block provided by the European Commission that enables the mutual recognition of national electronic identification schemes (eID) across borders, allowing European citizens to use their national eIDs when accessing online services from other European countries.

A second fundamental building block will be the one which facilitates trusted data exchange among participants, providing certainty that participants involved in the data exchange are who they claim to be and that they comply with defined rules/agreements. Trust refers to the fact that data providers and data consumers can rely on the identity of the members of the data ecosystem and beyond that between different security domains. Here, *IDS Connector technology,* as described in the *IDS Reference Architecture Model (RAM),*[15] emerges as a solid basis, and the FIWARE Community has incubated an open-source implementation of this technology that has already been tested how it can integrate with the rest of core FIWARE components.

Figure 24.8 illustrates how an AI service provider application and an AI service consumer application, each hosted in different organizations and participating in a common data space "powered by FIWARE," may interact with each other. It shows the role of components that are part of the incubated FIWARE IDS Connector implementation and their integration with other FIWARE components. The AI service provided can be an advanced traffic prediction service, while the AI service consumer application might be a traffic management system running in a given city. The AI service consumer wishes to incorporate near-real-time traffic predictions based on current traffic levels measured through IoT sensors deployed in the streets. The Context Broker component running on the consumer application sends notifications based on updates on attribute "current traffic" of a given street. These updates are notified to the AI service where the value is processed in real-time applying AI algorithms (e.g., using Apache Spark). The predicted traffic in about 30 minutes is

[10] https://www.keycloak.org/

[11] https://openid.net/connect/

[12] http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-tech-overview-2.0.html

[13] https://oauth.net/2/

[14] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eID

[15] https://internationaldataspaces.org/download/16630/

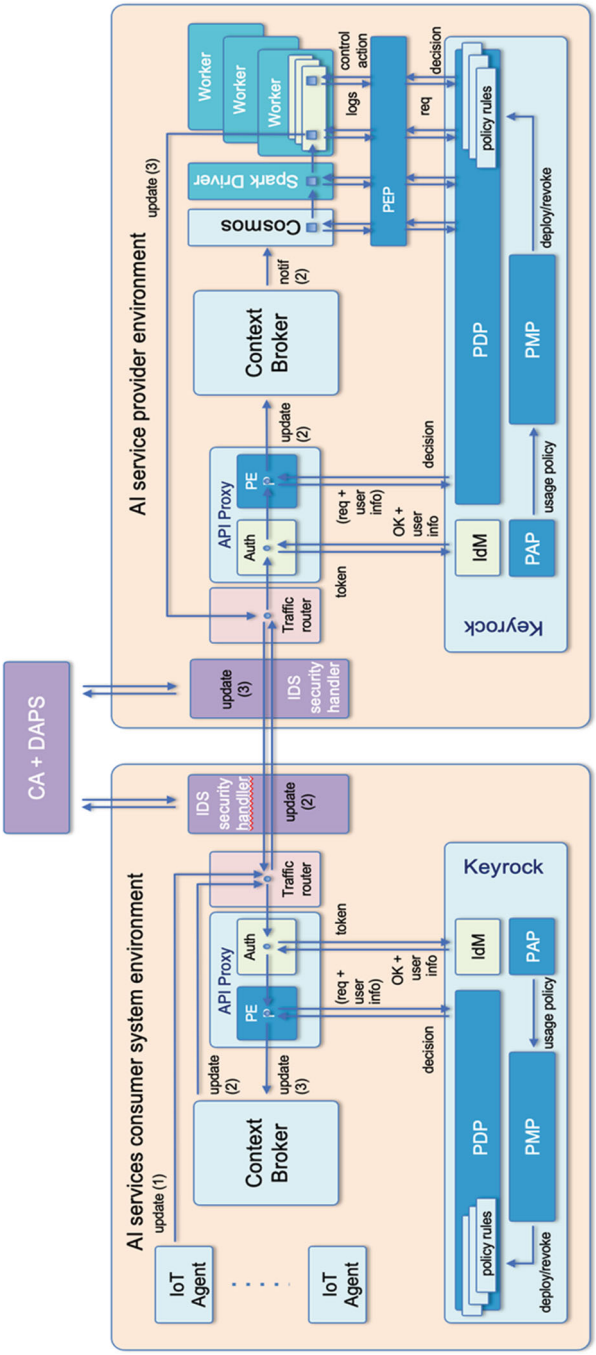


Fig. 24.8 Integration of FIWARE IAM and IDS Connector functionalities (©2021, FIWARE)

calculated (generating, e.g., values "low," "medium," or "high"), and the value of attribute "traffic in 30 minutes" of the given street is updated if changes occur.

In FIWARE, the implementation of IDS Connectors comes closely linked to the deployment of Kubernetes clusters. In this respect, the implementation of the concept of IDS connectors would not be limited to the implementation of the procedures used to exchange data among IDS connectors but also the monitoring and control of what is deployed in the associated clusters (e.g., guaranteeing that only certified components/applications are deployed) as well as the control of the flow of data within the cluster, ensuring enforcement of data access/usage control policies. For this last purpose, the usage of products supporting service mesh concepts like Istio[16] is essential. In Fig. 24.8, such components are identified as "traffic routers."

In the IDS RAM, identity and access/usage control is currently performed at organization level. The *Certification Authority (CA)* and *Dynamic Attribute Provisioning Service (DAPS)* are in charge of verifying that organizations exchanging data are known and trustworthy. The IDS Connectors ensure that data exchanges are authorized at organization level. In Fig. 24.8, for example, the FIWARE IDS Connector running at the AI service consumer side ensures that when an IoT agent updates context/digital twin attributes on the local Context Broker (1) forwarding of updates (2) are sent out only to authorized organizations like the AI service provider in this example. The FIWARE IDS Connector running at the service provider side will in turn verify that the organization from which updates are received is a trusted party and is authorized to make such updates to be then forwarded through the FIWARE Cosmos component to Spark where AI algorithms for traffic prediction purposes run. Similarly, when updates on the Context Broker deployed at the consumer side are invoked from the AI algorithms (3), the FIWARE IDS connector at the AI service provider side verifies that only requests authorized to be transmitted will be sent out. The FIWARE IDS connector at the AI consumer side verifies that the update request received comes from a trusted and authorized organization. All these verifications take place relying on the global CA + DAPS services of the data space.

On top of this authentication and data access/usage control procedures, performed at organization level via IDS Connectors, FIWARE security components enable an additional and finer-grained authentication and access control at the end user level. Therefore, coming back to Fig. 24.8, notifications received at the AI service provider side, same as update requests handled at the AI service consumer application side, carry additional security tokens (namely, JSON Web Tokens, JWT) linked to users on behalf of which those notifications/requests have been issued. The API proxy components, both at the consumer and on the provider side, validate those tokens and obtain the info associated with the corresponding users relying on standard Identity Provider services supported by the FIWARE Keyrock component. To

[16] https://istio.io/

manage access control, a standard XACML[17] process is implemented. The API proxy plays the role of the Policy Enforcement Point (PEP), and the FIWARE *AuthZForce* component, also alternatively Keyrock, implements the Policy Decision Point (PDP) functionality. That means that the proxy forwards user info of the particular user together with information about the concrete notification/request to the PDP which then checks whether users with those credentials are entitled to perform the given operation. FIWARE brings alternative implementations of the API proxy, namely, *Wilma*, *API Umbrella,* and *CoatRack*, all of them compatible with Keyrock or any third-party product implementing OpenID Connect, OAuth2, and XACML standards. Keyrock also implements Policy Administration Point (PAP) and Policy Management Point (PMP) standard XACML functions.

24.2.3 Data Value Creation

Loose coupling of participants is a fundamental principle in data spaces. Data providers and consumers do not necessarily know about each other. Therefore, it becomes essential to incorporate building blocks enabling the management of *data resources as true assets* with a business value, assets which can be published, discovered, and, eventually, traded, this way boosting the creation of multi-side markets where innovative services can be created.

FIWARE Business Application Ecosystem (BAE) components enable creation of *Marketplace services* which participants in data spaces can rely on for publishing their offerings around data assets they own. Different types of data assets can be defined via plugins, but three kinds of standard data asset types are supported by default, namely, static data files, right-time data resources provided via NGSI-LD at well-defined endpoints, and data processing services, which typically have associated well-defined endpoints for providing input data and publish results, both in right time, using NGSI-LD. Marketplace services are accessible through a portal or via APIs. Users, either end users through the portal or applications via APIs, of the Marketplace service can perform the following main actions:

- Define new data asset types via plug-ins.
- Register offerings around defined data asset types which typically means providing description of the asset, data models behind, endpoints (URLs) through with the data asset will be accessible, and, very important, terms and conditions defined around the data asset, including SLAs, legal clauses, access control policies users of the asset have to comply with, and associated pricing schema, which may be based on:
  - Free access (open data): User pay no money.
  - One-time payments: Users pay only once.

[17] https://en.wikipedia.org/wiki/XACML

  - Recurring payments: Users pay periodically (monthly, yearly, etc.) for getting access to some data. In addition, users will be able to cancel the subscription, but they won't be able to access data anymore.
  - Usage payment: Users pay per use. Their payments are based on the amount of information consumed.
- Ability to navigate and search/discover existing offerings based on selected criteria.

The following is a list of backend components and APIs associated with the FIWARE BAE Marketplace:

- Backend implementing standard TM Forum APIs supporting configuration of the marketplace:
  - Catalog Management API.
  - Product Ordering Management API.
  - Product Inventory Management API.
  - Party Management API.
  - Customer Management API.
  - Billing Management API.
  - Usage Management API.
- Rating, Charging, and Billing backend.
- Revenue Settlement and Sharing System.
- Authentication, API Orchestrator, and Web portal.

Figure 24.9 shows the FIWARE Data Marketplace portal deployed and configured for the i4Trust project.[18]

FIWARE also comprises components for publication of data resources linked to data assets around which offerings are managed through the FIWARE Data Marketplace. For this purpose, we have the Idra publication platform and have developed as well extensions to the CKAN open data platform, which is an open-data publication platform widely adopted in the market. These extensions support enhanced data management capabilities and integration with FIWARE technologies including NGSI. In particular, data publication and discovery features provided by CKAN have been enhanced with the following features:

- Right-time (near-real-time) time data publication—thanks to this extension, CKAN is not limited to list data resources linked to static files as part of its catalogue but also data resources linked to NGSI-LD requests served by Context Broker components deployed in a data space. This brings the ability to discover data resources relying on DCAT capabilities publication platforms support.

[18] https://i4trust.org/

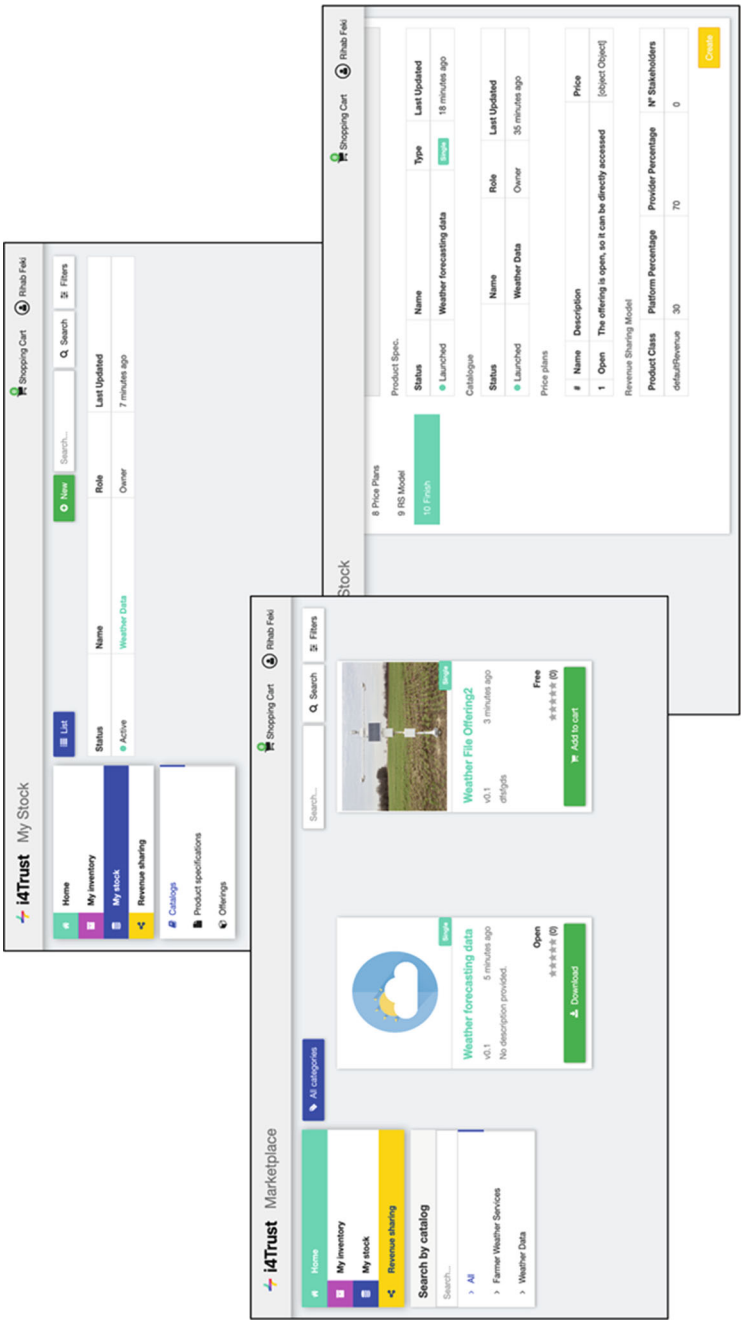


Fig. 24.9 FIWARE data marketplace portal (©2021, FIWARE)

- Identity management, authentication, and access control functions based on Keyrock components—therefore supporting OpenId Connect, OAuth2, and XACML standards adopted at overall data space level.
- Publication of priced data resources—thanks to this extension, it is possible to mark data resources listed as part of the catalogue as linked to offerings visible in the Data Marketplace. Users can therefore click on those data resources and navigate to the Marketplace to proceed with the acquisition of access rights.
- Enhanced Data Visualization—thanks to this extension WireCloud (Configurable Dash-boards component) to create adaptive and incremental data visualization. This way, farmers (sellers) can decide the way they want their data to be shown.

24.3 Toward Development of European Data Spaces

In February 2020, the European Commission announced the European Strategy for Data,[19] aiming at creating a single market for data to be shared and exchanged across sectors efficiently and securely within the EU. Behind this endeavor stands the Commission's goal to get ahead with the European data economy in a way that fits European values of self-determination, privacy, and fair competition. For this to achieve, the rules of accessing and using data must be fair, clear, and practicable. This is especially important as the European data economy continues to grow rapidly—from 301 billion euros (2.4% of GDP) in 2018 to an estimated 829 billion euros (5.8% of GDP) by 2025.[20]

The centerpiece of the European Data Strategy is the concept of "data spaces," for which the Commission defined nine initial domains, all driven by sector-specific requirements. Actually, the Commission promotes the development of European data spaces for strategic economic sectors and public-interest domains, starting with the following nine: industrial (manufacturing), green deal, mobility, health, financial, energy, agriculture, public administration, and skills.

While data spaces stimulate higher availability of data pools, technical tools, and infrastructure addressing domain-specific challenges and legislations, the EU Strategy for Data acknowledges that these data spaces should be interconnected and that this challenge requires specific attention. But Europe doesn't need to start from scratch—data sharing and exchange within specific domains and sectors is already happening in existing initiatives. However, each of these initiatives follows its own approach, and therefore they are not always interoperable.

Just like other electronic infrastructures (e.g., the Internet), data spaces basically are sector-agnostic, with many requirements and functions being similar or even

[19] https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

[20] https://datalandscape.eu/sites/default/files/report/D2.9_EDM_Final_study_report_16.06.2020_IDC_pdf.pdf

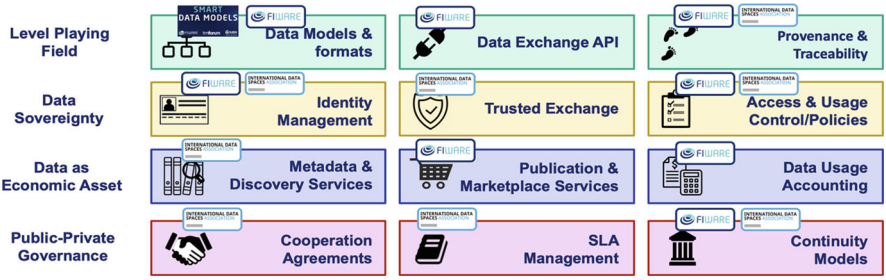


Fig. 24.10 IDSA and FIWARE positioning to create data spaces (©2021, FIWARE) (Labeling of a building block with the IDSA logo means that IDSA has produced or is working on specifications. Labeling with the FIWARE logo means that the FIWARE Community has produced open-source implementation of components and, eventually, is driving standardization activities. When both logos are present, an opportunity of collaboration has been identified linked to evolution of the specifications following an open-source implementation approach)

identical across different sectors and data spaces. Therefore, creating the basis for data spaces primarily is not so much a technological challenge, as there are plenty of technical solutions and standards available. The main challenge toward interoperable data spaces is to agree on standards and design principles that are accepted by all participants. While making it work in pilot applications, proof of concepts, and living labs is relatively easy, the real challenge lies in the convergence of interoperability to new norms and allowing for mass adoption and scalability. Alignment with the CEF (Connecting Europe Facility) Digital program[21] would be highly desirable, since this program is precisely devoted to bring the building blocks for creating Digital Service Infrastructures in Europe.

GAIA-X[22] is aimed at creating a federated form of data infrastructure in Europe which strengthens the ability to both access and share data securely and confidently. Since its creation, the initiative has raised a lot of awareness because of the opportunity it brings to join forces and, leveraging existing assets, accelerate delivery of solutions to the market.

As explained earlier, IDSA is providing several architecture elements which are necessary to create data spaces with trust and data sovereignty being some of the most important ones. These are necessary but not sufficient to create real living data spaces which are accepted and adapted on the market. Standard APIs, standard data models, and marketplace functionalities are also necessary, and this is where the core strength of the FIWARE ecosystems lies. The complementarities of both initiatives are illustrated in Fig. 24.10. As the FIWARE Context Broker technology was adopted in 2019 as one of the Building Blocks of the Connecting Europe Facility (CEF) program and has proven to integrate smoothly with other relevant CEF Building Blocks for the creation of Data Spaces (e.g., eIDAS, EBSI), this would

[21] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home
[22] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

bring an opportunity of alignment with this important program. As a result, the close partnership between FIWARE and IDSA is guaranteeing the aligned contribution of both ecosystems to the success of data spaces based on GAIA-X.

24.4 Conclusions

GAIA-X provides the opportunity to create a federated data infrastructure which is developed and established in a coordinated way, combining technological, functional, operational, and legal processes. This will strengthen the ability of Europe to both access and share data securely and confidently. Thus, it should be carried by the community of public and private stakeholders—and not by an individual keystone company, as we know it today from leading platform providers. FIWARE is bringing today mature technologies, compatible with IDS and CEF Building Blocks, which may accelerate the delivery of GAIA-X to the market. In addition, the FIWARE ecosystem has proven the ability to create standards in Europe which are accepted and adapted in the meantime on a global scale. Solution providers and end users in the Americas, Africa, India, or Asia are building their smart solutions based on FIWARE standards and technologies. With this, FIWARE became the global open-source-based standard, e.g., for the creation of Smart Cities. GAIA-X has the same potential to create solutions and to define standards based on a European value system which will be used also on a global scale.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 25
Sovereign Cloud Technologies for Scalable Data Spaces



Wieland Holfelder, Andreas Mayer, and Thomas Baumgart

Abstract The cloud has changed the way we consume technology either as individual users or in a business context. However, cloud computing can only transform organizations, create innovation, or provide the ability to scale digital business models if there is trust in the cloud and if the data that is being generated, processed, exchanged, and stored in the cloud has the appropriate safeguards. Therefore, sovereignty and control over data and its protection are paramount. Data spaces provide organizations with additional capabilities to govern strict data usage rules over the whole life cycle of information sharing with others and enable new use cases and new business models where data can be securely shared among a defined set of collaborators and with clear and enforceable usage rights attached to create new value. Open and sovereign cloud technologies will provide the necessary transparency, control, and the highest levels of privacy and security that are required to fully leverage the potential of such data spaces. Digital sovereignty, however, still means many things to many people. So to make it more concrete, in this article, we will look at digital sovereignty across three layers: data sovereignty, operational sovereignty, and software sovereignty. With these layers, we will create a spectrum of solutions that enable scalable data spaces that will be critical for the digital transformation of the European economy.

25.1 Introduction: Toward Open Clouds

Sovereignty in the digital world has emerged into one of the most discussed topics in Europe in the past years, and these discussions have recently materialized in the Franco-German initiated Gaia-X project, with 22 founding members, which incorporated over 200 international day 1 members at its official launch. This shows that there is a strong interest in the industry.

W. Holfelder (✉) · A. Mayer · T. Baumgart
Google Germany GmbH, Munich, Germany
e-mail: holfelder@google.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_25

Data spaces are one central building block that will allow organizations to share data within a well-defined, policy-controlled trust boundary and therefore provide the cornerstone for a flourishing ecosystem of novel use cases and business opportunities.

Google Cloud is committed to Europe, and our strategy is very much aligned with the objectives of Gaia-X: openness, security, trust, transparency, and federation. This is reflected in our partnerships with European companies, our focus on security and trust, and in how we address sovereignty demands in our products and solutions.

From our perspective, two pillars are of utmost importance to meet the requirements of European organizations, and these are providing technological excellence, including accessibility to emerging technologies, and at the same time having a deep understanding of their priorities when it comes to digitally transforming their businesses. These need to be put into balance with the workload-specific sovereignty requirements of the particular organization.

When we divide the "X" in Gaia-X into two halves, two important developments become visible that will guide the individual digitization journeys of European organizations. First of all, there is the "Next Generation of trustworthy Data Infrastructure" as envisioned by the lower half, and second there is the "Data Economy" which is constituted by the upper half and will allow the implementation of new data-driven use cases and business models. These are guard-railed by core services like the federated catalogue, the identity and trust layer, as well as by the policy rules and architectures of standards, and interoperability standards, to ensure that related systems are in line with European values.

25.1.1 Openness

We strongly believe in open-source software (OSS) and open clouds [1], because these provide users additional degrees of freedom when it comes to "vendor lock in" and relieve them of their dependence from a single provider. Experience has shown that even in the best of cases, enterprise IT can be rigid, complex, and expensive.

Conversations with European customers that have a strong on-premises IT footprint indicated that they would like to take advantage of the innovative capabilities that public cloud offers, but are worried of ending up in "just another lock-in."

Current industry trends indicate that multi-cloud and hybrid-cloud solutions are the future—not just for big multinational corporations but also for small- and medium-sized businesses. Technology users shall be able to build, port, and deploy their applications across platforms—cloud or on-premises. Open source, portable workloads, and open APIs are cornerstones of this approach. Instead of tethering customers to proprietary technology stacks, products, and solutions, providers should leverage open-source technologies, where it makes sense. An example is Google Cloud Anthos, which allows customers to manage their applications across different clouds, including our competitors, on-premises data centers, and the edge.

Some OSS projects have already proved that changing how industries work does not necessarily require a commercial interest. Kubernetes, Istio, and TensorFlow are just a few very successful examples of OSS initiated and open sourced by Google that gained wide adoption globally. The additional benefits of openness in the cloud ecosystem are giving cloud users greater control and flexibility while enabling healthy competition and unlocking new partnerships.

Initiatives like Gaia-X should act as an enabler for wider cross-organizational collaboration while at the same time allowing organizations to make sovereign choices about the technology stacks that they want to adapt. To allow accelerated growth, organizations should be able to pivot to global platforms if that is where they get access to the latest innovations in certain cases. This shall however happen under well-defined rules and in accordance with European values. The work on the policy rules and architecture of standards for Gaia-X has begun to leverage existing widely supported European frameworks, e.g., for portability and interoperability, which we welcome.

25.1.2 Security and Trust

Cloud users expect their providers to offer the highest level of security, and Google Cloud heavily invests in this domain since its inception. One external proof that our hard work is respected in the industry is the "The Forrester Wave™: Infrastructure as a Service (IaaS) Platform Native Security," Q4 2020 report that listed Google Cloud as a leader.

While security can be a strong differentiating factor, it is still something that users "just expect" to be present. But even with best in class security, no one will decide to use a specific technology or platform if the second ingredient is missing—and this ingredient is trust.

Users need trust in their providers and their behavior. They must be sure that the providers respect their free choice and don't add artificial barriers that would make it harder for them to switch data or software systems. We agree that customers should have the strongest level of control over their data that is stored and processed in cloud environments. In order to increase the transparency even further, customers need evidence through third-party audits, certification, and attestations and through additional transparency reports.

International Data Spaces (IDS) as the enabler of the Data Spaces Economy within Gaia-X has the potential to foster the uptake in cloud usage in general.

In order for this endeavor to be successful, the compliance rules for Gaia-X shall be founded on widely accepted certifications and attestations that customers and service providers already successfully work with like ISO, BSI C5, and others. As we are contributing to international standardization bodies ourselves, we believe that it is necessary to also consider the standards work from other global regions. This would lead to wider acceptance and would help establish and strengthen the perception of this European project internationally.

Our commitment goes beyond certifications, attestations, and reports. We provide guidance, documentations, and legal commitments to help our customers align with laws, regulations, and other frameworks. Therefore, we actively contribute to the implementation of a Cloud Code of Conduct based on the General Data Protection Regulation (GDPR) and are advocating toward a more modern approach to data security policies.

25.1.3 The Pillars of Digital Sovereignty

In order to practically address digital sovereignty [2] requirements in the Cloud, we see three important layers that need to be put into consideration, which are illustrated in Fig. 25.1.

25.1.3.1 Data Sovereignty

Customers need to be provided with a mechanism to prevent providers from accessing their data, unless the customers explicitly approve the access for specific provider behavior because they think the access requests are necessary.

Examples of such customer controls include storing and managing encryption keys outside the cloud, giving customers the power to only grant access to these keys based on detailed access justifications, and protecting data-in-use. Such controls allow the customer to be the ultimate arbiter of access to their data.

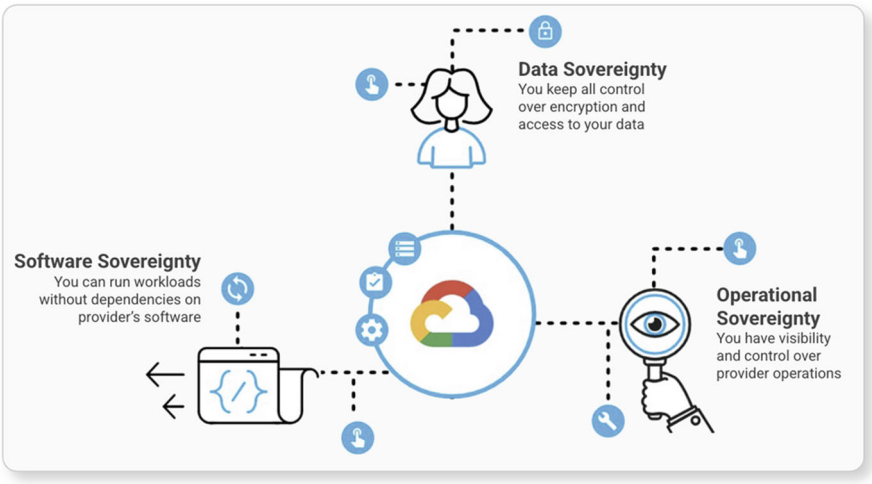


Fig. 25.1 The three pillars of sovereignty. © 2020, Google

25.1.3.2 Operational Sovereignty

Depending on the industry that an organization operates in, there might be a requirement for further controls. With these capabilities, the customer benefits from the scale of a multi-tenant environment while preserving control similar to a traditional on-premises environment.

Examples of such controls include restricting the deployment of new resources to specific provider regions and limiting support personnel access based on predefined attributes such as citizenship or a particular geographic location.

25.1.3.3 Software Sovereignty

The demand to control the availability of workloads and to run them wherever an organization wants, without being dependent on or locked-in to a single cloud provider, has increased steadily. This includes discussions around the ability to survive events that require them to quickly change where their workloads are deployed and what level of outside connection is allowed.

This is only possible when two requirements are met, both of which simplify workload management and mitigate concentration risks: first, when users have access to platforms that embrace open APIs and services and, second, when they have access to technologies that support the deployment of applications across many platforms, in a full range of configurations including multi-cloud, hybrid, and on-premises, using orchestration tooling.

Examples of these controls are platforms that allow users to manage workloads across providers and orchestration tooling that allows them to create a single API that can be backed by applications running on different providers, including proprietary cloud-based and open-source alternatives.

25.1.4 Partnering with European Companies

Having the critical skill set in Europe and increasing the amount of technology knowledge has to be an important element of any successful European sovereignty strategy. As outlined above, Google is committed to open source. By partnering with European companies, their employees gain access to open source and state-of-the-art technologies, which will not only provide for more independence but also allow them to increase their knowledge and own market value while at the same time increasing the organizations' footprint in the local markets.

We are also fully committed to contribute to the goals and values of Gaia-X and to the European Digital Data Ecosystem by providing best in class technology to our customers in Europe:

- By investing further in capital, engineering, and go-to market resources in Europe
- By implementing the highest data privacy, data residency, and security standards
- By working with policymakers and partners to meet their specific requirements

We will bring our expertise to the table, but we also want to listen and learn from our European customers and partners around the table.

We envision these kinds of partnerships also in the context of the wider data spaces ecosystem, which would allow the amplification of knowledge transfer to an even border ecosystem and not just to the bilateral partners.

25.2 Technological Evolution from Cloud Native Applications to Data Spaces

The evolution of data spaces was a consequence of a longer technological evolution.

Today, organizations have the desire to reduce their on-premise infrastructure investments and to modernize their IT stacks to focus on their business differentiators, by leveraging innovative, highly scalable cloud services that provide strong security, competitive pricing, and access to continuous, fast-paced innovation. This trend became possible by the development of new technologies, several of them are open source, which allowed IT teams to rethink how they could best possibly support their business needs while at the same time increasing the efficiency of their infrastructure deployments.

25.2.1 Containerization: A New Paradigm

Containers are the first important milestone that we will highlight. The term container is taken from the principle of shipping containers. As long as we do not talk about special substances, the harbor infrastructure doesn't care about what is inside a container, be it a luxury car or ten refrigerators. The handling of the container itself, due to its uniform interface, stays exactly the same. Containers allow software developers to package their source code with all the required dependencies in a self-contained way, so that they can be deployed and tested consistently across different environments like a developer's laptop, test, system integration, or production. This containerization addresses risks like version mismatches in shared system libraries that can lead to potentially large incidents.

A notable technology at the beginning of this development is Docker, which provides a set of services that use operating system (OS)-level virtualization to deliver software in containers. Over time, the tooling ecosystem was massively extended to simplify working with containers and other solutions entered the market.

25.2.2 Container Orchestration: Kubernetes (k8s)

Once containers were used by a broader community, the next challenge was to run them efficiently on top of distributed infrastructure. One solution to this challenge, which quickly emerged to a broadly accepted standard, is Kubernetes, the open-source container orchestration system, which was initially announced by Google in mid-2014 and was heavily influenced by Google Borg's [3] design.

Kubernetes is the core of several managed cloud services that aim to provide an excellent user experience, low operations overhead, and a high degree of automation capabilities for a seamless user experience. Container orchestration solutions nowadays allow for the implementation of highly efficient, scalable, and secure solutions that provide a strong foundation for all kinds of enterprise applications. The ability to use the underlying infrastructure in an efficient, customizable manner, for example, by leveraging bin packing, is just one of the benefits that this new model introduced.

25.2.3 Service Mesh: Istio

As large-scale containerized applications, consisting of hundreds or even thousands of containers, were built, the next challenge was to effectively connect these services with each other and to control their interaction behavior.

As containers are often described as microservices, the solution to this challenge that found quick adoption was termed Service Mesh, and one of the leading open-source implementations is Istio.

The Service Mesh uses so-called "sidecar" proxies that run alongside each service. This means that for each service, an individual sidecar proxy is instantiated, and the connections between the different sidecars form the mesh. Some important functionalities that a Service Mesh can provide are service discovery, load balancing, failure recovery, metrics and logging, monitoring, A/B-testing, canary rollouts, rate limiting, access control, and end-to-end authentication. These can help developers to focus on solving business problems and service specifics rather than the surrounding ecosystem features and requirements.

25.2.4 Data Mesh

The new Service Mesh principles influenced another important step in the progression toward data spaces. The term data mesh [4] was introduced in 2019 and describes an enterprise architecture that is a symbiosis between the agility of a service mesh combined with product management practices, platform thinking, and self-service around data stores.

The high-level idea is to manage data as if it was a product, including a team that is responsible for this "data product." Besides the quality and availability aspects, the team also needs to ensure that data can be easily discovered and consumed by the surrounding ecosystem, which might be a single organization.

Data discovery, which is one of the key features, is made possible by mechanisms including metadata or machine-readable self-descriptions which describe the underlying data in terms of quality, usability, domain specificity, and further aspects. This helps business users identify information that fits their needs.

In order to ensure that data can be consumed frictionlessly, supplementing technologies, for example, GraphQL, could be added to the mix, but they might introduce new challenges and therefore need to be carefully evaluated before considering them as major building blocks of a technology strategy.

25.2.5 Data Space

While the concepts that were discussed in the previous sections enable broad and industry-agnostic data processing use cases for companies of all sizes, one important element is missing to enable cross-organizational collaboration on data, and this is a model of trust.

Once data leaves the control span of a given organization, the data provider can no longer execute governance on the data and has to rely on the consuming parties to follow agreements, licenses, and contracts. It would be beneficial to have a technology that can help keep control over shared data, by extending the control span to the destinations that request data, which would essentially allow the creation of a trust boundary on top of a virtual overlay network, with multiple participants and distributed locations—a data space.

This is where the Reference Architecture Model of the International Data Spaces Association (IDSA) comes into play, as it provides a solution that can fill this gap, in order to create a vivid and pluralistic data ecosystem that gets widely adopted.

Within a data space, data can be enriched with policies that are enforced by the platform components like the Connector and allow the data provider to control the life cycle of the data and decide how the data can be used. The governance model can also be amended with contractual agreements for cases where the technology alone cannot provide a strong enough foundation of trust.

25.3 Interoperability as a Key Enabler for Hybrid and Large-Scale Data Spaces

While cloud computing established the basis for efficient networks of functional services, "big data" and "data lakes" also had a strong influence on the evolution of data spaces. "Big data" found its first prominent implementation in 2005. The Hadoop Data Management Framework was developed by Yahoo, based on Google's Map Reduce [5] paper.

25.3.1 Big Data and Data Lakes: An Early Generation of Data Spaces

The key idea behind big data is to consolidate data in one location in order to extract value out of it. These consolidated data repositories, and their surrounding ecosystem, are often described as data lakes. This concept has several commonalities with the data spaces idea. One is to bring together information from existing data stores.

The issue with this goal is that over time, the amount and velocity of data have increased and that at the same time, data is often distributed among independent data silos that don't share common governance mechanisms or data models. This makes the integration of all these disparate data sources an extremely complex endeavor. The growing demand for digitalization and cross-enterprise service offerings introduced additional challenges, and often organizations don't even know what data they have, where it originates from, or what the quality of the data is. To get a handle on the data that is present in an organization, data warehouses tried to evolve into "Enterprise Information Stores" which introduced new challenges as the increased velocity and volume of data led to inefficiencies with regard to data ingestion and transformation.

Big data quickly raised strong expectations as the promise of increasing the scope of usable data within an organization was very appealing. The ability to concentrate data in a single location and to join datasets to enable new use cases like machine learning was understood as a strong opportunity to build new products and to generate additional information and organizational wisdom. Some of the concepts underlying data spaces are quite similar to what big data tried to solve. Both aim to generate more value from shared data, by introducing common or joint use cases through evolution and innovation; what has changed is the scope.

One thing that happens in organizations from time to time is that a prototype will silently evolve into a productive solution. The transition from Extract Transform Load (ETL) to Extract Load Transform (ELT) in the data lake context provides an ideal candidate for such a silent transition as storing data from all the different sources "as-is" in a single location sounds like an easy exercise.

Such a transition usually doesn't follow well-informed and purpose-designed architectural decisions, including systems and information architecture, and the idea

that having data in a central location would magically solve all the challenges when it comes to data ingestion and consumption, real-time use cases, streaming and batch scenarios, the integration with other enterprise systems like analytics platforms, or the different requirements in process-driven and transactional use cases, all at the same time, was meant to fail. It is therefore no wonder that different analyst reports in the 2015 to 2017 years came to the conclusion that 60 to 85% of big data projects actually failed.

In alignment with the concepts of data warehouses and data marts, so-called data ponds were introduced to support the data lake architecture. Data gets ingested into the data lake and is then processed to fit a special purpose, and the "cleansed data" is eventually persisted in a data pond. This evolution introduces its own challenges, but on a high-level abstraction, the "new" architecture reminds of a classic enterprise data warehouse solution, which it originally aimed to replace.

While the "built for purpose" subsets of the data lake provided a good starting point, it still couldn't address specific consumption use cases that required things like high Create-Read-Update-Delete (CRUD) performance or the modeling of specific data relationships in the context of a specific knowledge domain. To solve these challenges, a "polyglot" data environment was required that included different types of data stores like relational, document, or graph databases, columnar stores, and cubes. The target state was once again reminiscent of the classic data warehouse architecture.

Eventually, many data lake project owners came to the conclusion that the original data warehouse architecture did incorporate a lot of good concepts and that data lakes were better used as a complement than a replacement. While the scope of a data lake typically includes data sources and consumers within an organization, data spaces envision a much broader ecosystem of cross-organizational collaboration. And while this vision offers immense opportunities, it also introduces nontrivial challenges at the technical and organizational level.

The wider discussion around data meshes started an evolution that not only complemented classic data warehouse architectures but transformed them into distributed networks of Information Management (IM) factories, supporting specialized ingestion and consumption use cases and collaborating with each other to holistically solve larger-scoped IM scenarios.

Concepts like cloud and International Data Spaces (IDS) try to provide answers to some of the aforementioned challenges by integrating technology and architecture solutions to provide highly scalable infrastructure for data meshes while at the same time tackling the challenge of data governance throughout the whole life cycle of the data.

25.3.2 *Gravitation and Expansion: The "Yin and Yang" of a Successful Data Spaces Strategy*

An important learning with regard to data spaces, which particularly needs to be considered in cross-industrial initiatives like Gaia-X, is that a too strong focus on standardization might introduce negative side effects as well. One learning from the data lake era is that polyglot solutions can add a high degree of value to businesses—some of these solutions might however be proprietary to some degree.

While the added value through the consolidation of data into a common data store is well understood, the potential of dedicated, use case specific implementations of data processing capabilities in the context of complex use cases should not be underestimated.

Examples of such implementations can range from specific hardware requirements like edge components or sensor arrays, and use case specific functional and nonfunctional requirements like scalability or specific security patterns, to highly optimized cloud native data warehouse services, which apply sophisticated cloud architecture and operation patterns to provide massive performance. In some cases, these implementations even overcome limitations of classic IT systems like ingestion and consumption efforts in DWH environments or, with a somehow relaxed set of constraints, the CAP theorem [6].

Figure 25.2 illustrates the concept of a data space ecosystem as it is envisioned within initiatives like Gaia-X. Participants share or consume data provided via so-called data services. These services can either provide access to the data itself

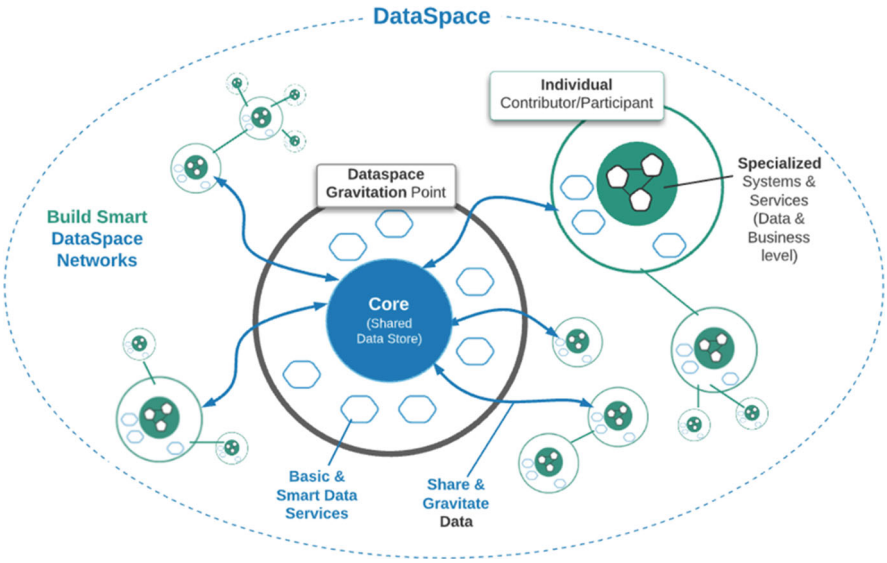


Fig. 25.2 Data space (Network) Vision, incl. Gravitation core, (Smart) data services, and specialized contributors. © 2021, Google

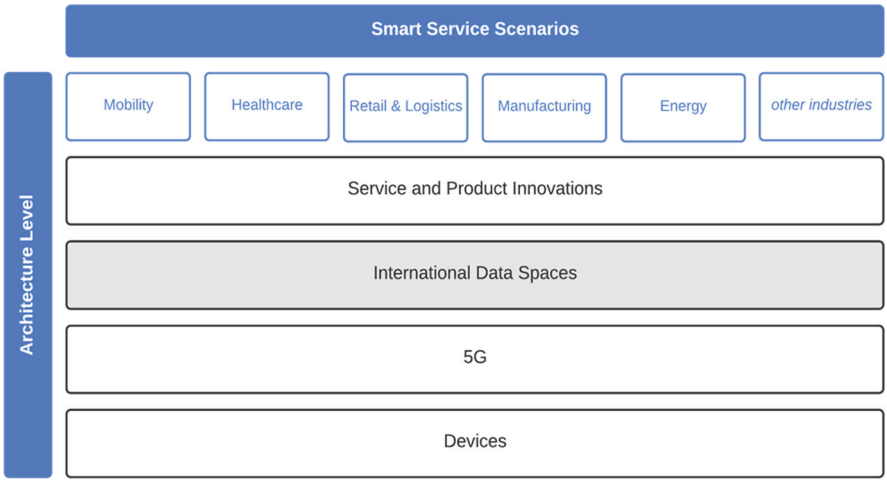


Fig. 25.3 Typical data space enterprise architecture stack based on the IDS Reference Architecture Model 3.0 [7], © 2021, Google

or provide applications that encapsulate the access, or offer data specific operator implementations to standardize or simplify the exchange. By embedding these services into more sophisticated data ecosystems, targeting domain-specific use cases like mobility, healthcare, or logistics, so-called smart services are introduced.

Traversing up the typical data space enterprise architecture stack (Fig. 25.3), these smart services can either provide more advanced data management and operation services or even represent complex use cases by providing domain-specific data models and business-focused interfaces. The ability to combine services at the business level presents one of the foundations to form domain-wide use cases that go beyond simple data exchange scenarios.

While the approach to combine internal and external IDS connectors into data space network clusters eliminates the need to have central data stores, having gravitational elements within a data space scenario does introduce many benefits. Besides potentially reducing network traffic, it gives the opportunity to consolidate data, similar to a classic data warehouse, allowing to combine and more holistically analyze domain and data space data to derive common information and wisdom out of it. Typical data warehouse scenarios usually promote a clear separation between data sources and data consumers; data spaces should introduce bidirectional communication as a core design goal. Feeding back the results provided by analyzing shared data can help with the incremental evolution of data spaces and its driving domain use cases as a whole. The combination of gravitation and expansion within a domain's data space ecosystem is the key to building sophisticated use cases of high maturity.

Smart services that provide custom-tailored data processing and analytics capabilities will play an integral part in the upcoming data spaces, and fundamental platform services like Kubernetes can provide the foundation for them. The provided

layers of abstraction allow the realization of portable solutions that can be seamlessly moved between different platforms, without the need of a complete rewrite. This portability increases the sovereignty of the organizations that deploy and operate such services.

An issue when agreeing on a least common denominator in the context of technologies and platforms can however be that functionality suffers. As mentioned earlier, several highly specialized and sophisticated cloud-based hyper-scale solutions can hardly be mapped to this least common denominator. If "portability under every circumstance" is defined as a core architecture principle, this could lead to a situation where organizations are no longer able to leverage bleeding edge technology that would provide them with a strong business advantage.

In "Don't get locked up into avoiding lock-in" [8], Gregor Hohpe provides a well-balanced view on the different lock-in types and why "lock-in isn't an all-or-nothing affair." This article explains why common attributes like lock-in or coupling aren't binary, debunks common myths, such as using open-source software magically eliminating lock-in, and points out that it is difficult to not be locked into anything, which is why some amount of lock-in might be acceptable in order to unlock the innovation potential of a target architecture to its fullest.

A vast amount of the motivators driving customers to invest into data space initiatives and associated use cases is the high demand for wider adoption of digital solutions within their specific business domain. In order to achieve this goal, they should have access to the best solutions in the market, so that they can focus on their business goal.

25.3.3 Portability and Interoperability: A Perfect Complement

While portability of data and services is important, organizations need to find the right balance and understand trade-offs between efficient digitization and innovation speed on the one hand and strong sovereignty requirements on the other hand. Therefore, another aspect needs to be added to the spectrum, which is already well-known in the enterprise world, and this is interoperability.

Portability can still be a core architecture principle to target for, but we pledge for a slightly more flexible approach that implicitly allows the use of complementary solutions that might not be fully portable as long as they provide a high degree of interoperability. This line of thinking is related to the free choice of the best possible solution for a given task.

When we envision a full stack of large enterprise IT, including all the legacy that was created over decades, an approach that would demand to reimplement everything to fit the defined framework would not be applicable, and as most organizations already use certain technologies, including the required investment in technology, licenses, and skills, they want to be able to continue using solutions

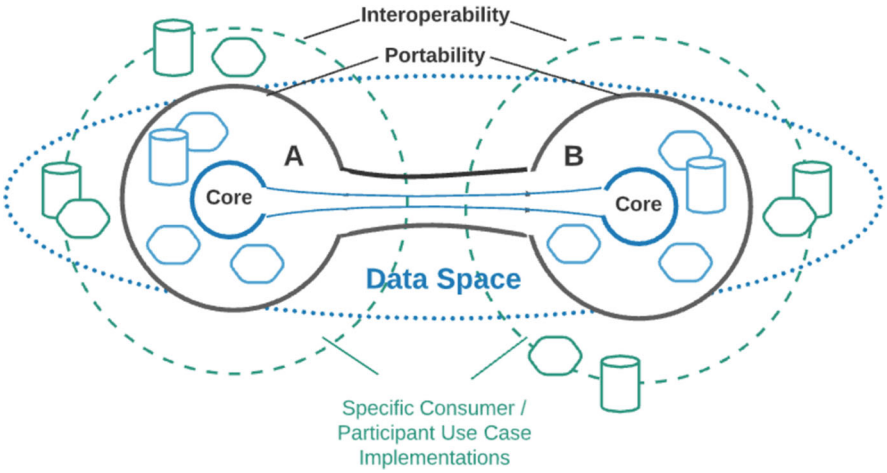


Fig. 25.4 Portability and interoperability perimeters in context of "Polyglot" data spaces. © 2021, Google

that have proven to be valuable to achieving their business mission. These companies can still see a benefit in participating in data spaces, be it Gaia-X-compliant European data spaces or international data spaces.

While there are many ways to achieve this "best of both worlds" goal, Fig. 25.4 illustrates one high-level conceptual idea of how this could be approached. The idea is to bring together portability and interoperability in the context of data spaces-driven use cases.

Figure 25.4 shows a data space with associated two "Gravitation Centers," A and B, consisting of a core, typically represented by a shared data store and associated (Smart) data services. The latter are targeted to offer core processing functionalities associated with a certain data space type and can bring their own complementary data store if needed. Users of the data space access the provided services via defined interfaces and connect them with their very specific use case implementations. While it is important for services in the gravitation center to provide a high level of standardization and portability, the peripheral customer-specific parts of a use case implementation should be less restrictive in this sense. Portability should be interpreted in a sense that not only allows the exchange of data but data space-specific semantics as well as compliance with defined API sets, security measures, and protocols.

As a data space grows, it becomes harder to apply centralized governance; therefore, it is important to introduce additional distributed governance mechanisms. IDS Connector implementations can help govern data space participant endpoints and help solve the problem of decreasing trust for participants that are not part of the core services ecosystem (Fig. 25.5). This ensures that only authorized participants can access a service and that defined policies are enforced at all times, regardless of

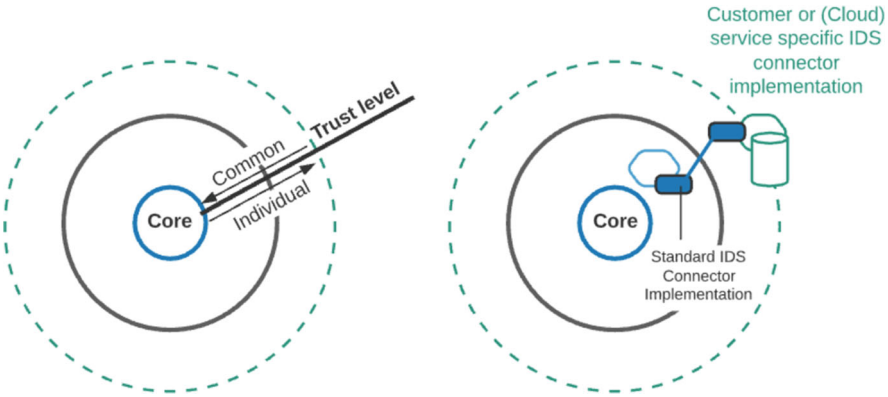


Fig. 25.5 Trust perimeters in the context of a data space gravitation center and IDS. © 2021, Google

the deployment location of the participating service, be it in a Cloud or On-Premises or at the Edge.

25.3.4 Interoperability via Solutions-Specific Connector Implementations

IDS provides a reference architecture model that addresses the aforementioned requirements. The ability to manage policy-driven governance of data and data services (access and usage control) does not only provide the level of functionality needed, but it does this in a transparent and standardized way.

In order to participate in a data space collaboration scenario with customer or cloud-specific components while still complying with the required data sovereignty demands, customer or solutions-specific IDS connector implementations have to be provided.

In this context, there might be scenarios where the use of a specific solution makes sense, e.g., a cloud-based data warehouse with extended machine learning capabilities that allows business users to leverage the power of machine learning through a well-known interface like the sequential query language (SQL).

Regardless of whether such an adapter would be implemented by a customer, a partner, or a provider, a proper evaluation in alignment with regard to the suitability and demand of the underlying use case needs to be performed.

Figure 25.6 illustrates how such a connector construct could look like. Based on a reference connector implementation that would need to be fully portable, the target system-specific adapters would “only” need to comply with interoperability requirements.

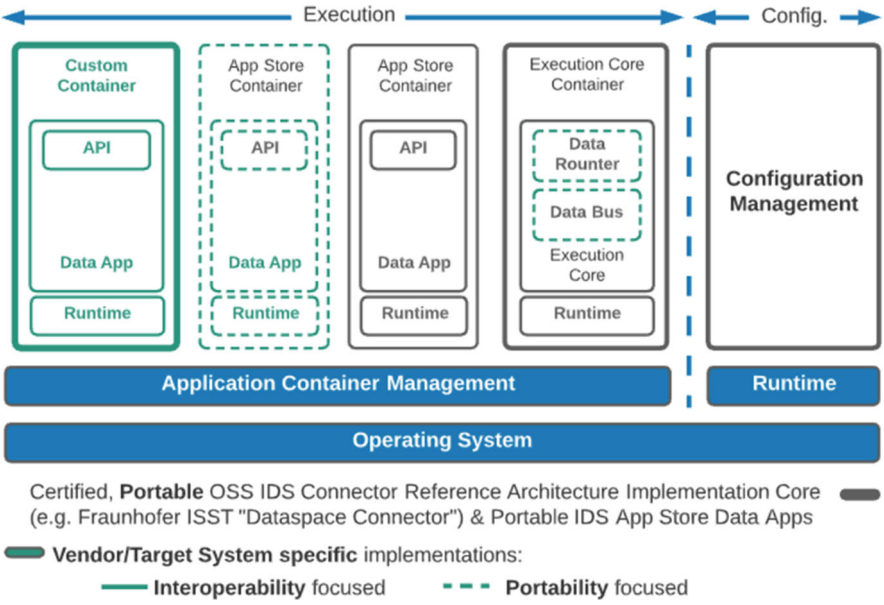


Fig. 25.6 Portability and interoperability in the context of the IDS Connector Reference Architecture. © 2021, Google

How such peripheral services would be embedded in a complete IDS governance environment, including IDS Broker and Clearing House services, that in some cases, especially when adding none open-source on-premise systems into the picture, might not fully comply with all sovereignty aspects targeted for a certain ecosystem, e.g., in context of Gaia-X, still needs to be discussed. We are confident that the value that these extended data space scenarios can introduce to the implementation of efficient digital use cases and innovation cycles will motivate the necessary steps to be taken.

25.4 Future Outlook

Over the course of the next few years, we expect organizations to develop new and innovative use cases and products that embed cross-organizational data sharing as a core principle.

This will lead to a higher demand in network and data center capacity, which will, given the climate change challenge and sustainability objectives, make efficient data center operations and Green IT commitments even more important. Energy-efficient cloud infrastructure is however not the only relevant challenge. The idea of implementing digital use cases by building large application and service composites from reusable services and artifacts shares several commonalities with the visions

behind service-oriented architectures (SOA). Therefore, it makes sense to review why it encountered substantial challenges in order to evaluate potential learnings.

One fundamental issue was finding the right balance between a service-based ecosystem with many independent participants and a governance structure to maintain, mature, and refine it. Any service ecosystem will sooner or later face issues if it doesn't implement a proper governance process that helps control the evolution of the system as a whole. While a central governance instance might not be a contemporary approach anymore, providing tools and best practices to allow participants within such an ecosystem to collaborate and align their service life cycle and evolution management to common goals and principles becomes a mandatory capability for long-term success.

Given the expected dynamics of use cases and distributed innovation power across the different industries and domains, it is very likely that several of such ecosystems will arise. In order to drive and accelerate innovation and digital progress across all areas of society, it will be important to have a connecting element between these ecosystems. How such mechanisms and governing elements can be established is under active development and still work in progress. It will be interesting to see how the different use cases will impact these developments. Initiatives like the International Data Spaces Association (IDSA) and Gaia-X could further emerge into such supporting entities, evolving with the growing ecosystem of data and applications, based on a network of secure and reliable interconnected services.

It will be important that the different stakeholders, including international cloud providers, collaborate in strong alignment to ensure that the best possible outcomes are reached.

References

1. Kurian, T. (2020). *Empowering customers and the ecosystem with an open cloud.* [Blog] Inside Google Cloud. Accessed April 06, 2021, from https://cloud.google.com/blog/topics/inside-google-cloud/empowering-customers-and-the-ecosystem-with-an-open-cloud
2. Kurian, T. (2020). *Engaging in a European dialogue on customer controls and open cloud solutions*. [Blog] Inside Google Cloud. Accessed April 06, 2021, from https://cloud.google.com/blog/products/identity-security/how-google-cloud-is-addressing-data-sovereignty-in-europe-2020
3. Verma, A., Pedrosa, L., Korupolu, M. R., Oppenheimer, D., Tune, E., & Wilkes, J. (2015). Large-scale cluster management at Google with Borg. In *Proceedings of the European Conference on Computer Systems (EuroSys)*.
4. Dehghani, Z. (2019). *How to move beyond a monolithic data lake to a distributed data mesh.* [online] martinfowler.com. Accessed February 22, 2021, from https://martinfowler.com/articles/data-monolith-to-mesh.html

5. Dean, J. & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation* (pp. 137–150).
6. Brewer, E. (2017). *Spanner, TrueTime & The CAP Theorem.* Accessed March 15, 2021, from https://research.google/pubs/pub45855.pdf
7. Boris Otto, I., et al (2019). *IDS Reference Architecture Model 3.0*. Accessed March 15, 2021, from https://internationaldataspaces.org/use/reference-architecture/
8. Hohpe, G. (2019). *Don't get locked up in avoiding lock-in.* [online] martinfowler.com. Accessed April 06, 2021, from https://martinfowler.com/articles/oss-lockin.html

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 26
Data Space Based on Mass Customization Model



Lucheng Chen, Haiqin Xie, Wen Yang, and Lin Xiao

Abstract The industrial Internet has brought the biggest economic opportunity since the mobile Internet. Haier started the exploration of intelligentization, networking, and informationalization as early as 2005 and gradually launched an industrial Internet platform COSMOPlat with independent intellectual property rights. In COSMOPlat, traditional mass manufacturing model is replaced by mass customization model (MCM), in which production is driven by users' order rather than inventory. What the model achieved is not simply automation but also real intelligent manufacturing with high efficiency driven by high precision. With users' participation in the whole process, manufacturing is precisely made according to the users' dynamic needs, which will better meet actual requirements.

The article provides an overview of the mass customization model of COSMOPlat and how it facilitates enterprises for digital transformation and upgrade. Also indicated are the successful use cases in application.

26.1 Big Data Analysis and Mass Customization Model (MCM)

26.1.1 Introduction of Haier COSMOPlat

COSMOPlat is an industrial Internet platform initiated and developed by Haier based on its advanced business philosophy and rich experience in the manufacturing industry. COSMO means the God of Chaos and symbolizes the search for new developments out of Chaos. As a platform with independent intellectual property rights in China, COSMOPlat designs a user-oriented platform with the model of mass customization that opens to users for full-process engagement.

Technology is the driving force of development and change; in COSMOPlat, advanced intelligent manufacturing, network engineering, IoT, digitalization

L. Chen · H. Xie (✉) · W. Yang · L. Xiao
Haier COSMO IoT Ecosystem Technology Co., Ltd., Qingdao, Shandong, China
e-mail: chenlc@haier.com; xiehaiqin@haier.com; yangwen1@haier.com; xiaolin@haier.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_26

engineering, big data, and artificial intelligence are further developed to provide technical support.

Relying on the technological innovation system, COSMOPlat is mainly engaged in industrial Internet platform construction and operation, industrial intelligent technology research and application, intelligent factory construction, as well as hardware and software integration and fundamentally subverts the traditional industrial system and industrial structure, enabling the overall intelligent transformation and upgrading of enterprises.

With the philosophy of mutual evolution and value sharing, COSMOPlat strives to build an open, co-creation, and win-win ecosystem and embraces innovations in mass customization and information technology fusion in different industries. COSMOPlat is becoming a platform that empowers interactions and value sharing, facilitates new industries and new ecology, and supports entrepreneurs and innovations.

Starting from the mass customization model, an industrial Internet ecosystem with COSMOPlat is rapidly extending and empowering to different industries and fields, and through the replication of mass customization solutions, COSMOPlat not only enables Haier's own digitalization but also opens up the whole industry chain ecology around users' needs, realizes cross-industry and cross-regional ecological empowerment, and helps enterprises realize rapid transformation and upgrading. At present, COSMOPlat has built 7 centers and benefited 15 industries in more than 20 countries worldwide.

As typical use cases of COSMOPlat's mass customization model, two Haier interconnected factories, namely, Haier Central Air Conditioning Interconnected Factory and Haier Refrigerator Interconnected Factory in Shenyang, have been rated as global "Lighthouse Factory" by the World Economic Forum, making Haier the only enterprise who owns two "Lighthouse Factories."

26.1.2 Overview of MCM in Haier COSMOPlat

The MCM is a user-oriented approach initiated by COSMOPlat to provide customized and scenario solution driven by user's needs, building a new industrial ecology for stakeholders to achieve win-win cooperation. In this approach, users are no longer just buyers, but rather participants that engage in the whole process of the manufacturing and designers who can put their ideas in building their desired products.

In MCM, based on new technologies including AI, blockchain, big data, and 5G, an open ecosystem for resource-gathering and sharing is built. This "tropical rainforest ecosystem" in the era of IoT enables a user interaction platform for enterprises to get access to user resources, as well as a win-win and value-added platform for SMEs and ecological resources to jointly create value and share value [1]. MCM provides an overall solution covering intelligent manufacturing,

interactive customization, open innovation, smart services, precision marketing, modular procurement, as well as intelligent logistics.

Based on the principle of "co-building with large enterprises and resource sharing with SMEs" [2], it promotes transformation and innovation for large traditional industries and brings digital upgrade for SMEs.

By the user community established on the platforms, a large amount of user demand data has been accumulated, autonomic machine learning is combined to process the user demand data, and neural network is used to learn the characteristics of user's demand, enabling the machine itself for the prediction of user's demand. As a result, we can not only have a clear mind of the user's demand but also help them find the services that best meet their needs. In this sense, quick respond to either the user or enterprise's need is realized.

With MCM, the production efficiency has been improved by 60%, and the non-warehousing rate of the product has reached 75%. At the same time, COSMOPlat is recognized as "the most professional junction between enterprises and intelligent manufacturing resources." While serving its own interconnected factories, it also provides solutions and value-added services for the transformation and upgrading of manufacturing enterprises, so as to continuously improve their mass customization capabilities and offer the best experience for users.

In the implementation of MCM, an interconnected factory system with direct connection to users has been built by COSMOPlat. Users' order information can arrive at the factory in the shortest time, and personalized products are customized with high precision and high efficiency.

26.1.3 MCM and Haier Interconnected Factory

Departure from the downtown of Qingdao, it takes less than 20 minutes' drive to Haier central air conditioning interconnected factory (Fig. 26.1) located in the Sino-German Industrial Park.

The outside appearance of this factory shows nothing extraordinary; while entering you will find a totally different world. Walking into the production workshop, there are two robots waving their orange manipulators to work on the production line in an orderly manner. Technologies including IoT interconnection, AR cooperation, and AI detection empower the robots with intelligent operation capability, which not only liberates manpower but also improves production efficiency and reduces error rate.

Machine intelligence is only the foundation. The interconnection between equipment and production process and the links between manufacturers and users are the key for intelligent manufacturing in the era of industrial Internet. In traditional factories, overstocking always remains a headache for many manufacturing enterprises. In the interconnected factory, the central air conditioner on the production line has already had an "owner" long before it was "born"; thus, non-warehousing



Fig. 26.1 Haier COSMOPlat interconnected factory based on MCM

rate of products almost reached 100%. Among Haier's 15 existing interconnected factories, the average product non-warehousing rate is 75%.

In the interconnected factory, all elements are interconnected with users. Walking around the factory, "screens" can be seen everywhere. Product specification selection, start customization, and order confirmation are all shown on the big screen; orders from all over the world are constantly updating on the screen, and each order is made based on the user's needs. Mass customization platform, intelligent cloud service platform, energy-saving cloud service platform, field operation platform—the abstract industrial Internet platform is thus made visualized on various interfaces to show how different production factors interact and create value.

By interconnecting with the massive data collected by various sensors in the factory, predictive maintenance, intelligent scheduling, and quality intelligent monitoring of core equipment can be realized.

In September 2018, the factory stood out from more than 1000 factories and was rated as one of the first nine "Lighthouse Factories" in the world by the World Economic Forum, which was the only Chinese factory in the list. In September 2020, Hair MCM and interconnected factory won 2020 ROI-EFESO Industry 4.0 Award, becoming the first Chinese enterprise to receive this honor.

26.1.4 From Mass Production to Mass Customization

For high-precision product innovation and high efficient intelligent production, Haier has built seven application platforms based on a variety of intelligent technologies. The seven platforms, covering user interaction, marketing, R&D, procurement, manufacturing, logistics, and service, are broadly explored and applied so as to realize the large-scale customized application service centered on users' personalized needs.

The customized platform for user interaction (HaierDIY) is a customized experience platform for user community interaction. Based on the platform, users can put forward demands, ideas, and comments and conduct online interactions with the home appliances, which become the source and driver for product innovation.

The precision marketing platform is based on CRM and user community resources. Through research on big data, existing user data and the user data collected by the third party are analyzed and studied. At the same time, cluster analysis is applied to form user portrait and build label management, so as to realize precision marketing.

The open R&D platform, which consists of three key components, namely, Haier Open Partnership Ecosystem (HOPE), HID iterative R&D platform, and collaborative development platform, is built to meet various requirements in different stages of R&D technology requirement interaction, product design and implementation, and collaborative development interaction of module manufacturers.

The modular procurement platform is developed on the module vendors' collaborative procurement platform. It is a module vendor resource service and aggregation platform built for the needs of zero distance interaction between module vendors, resources, and users, enabling module vendor on-demand design and modular supply. The procurement system adopts a distributed architecture, in which user requirements are publicly sent to global module vendors, and the system will automatically and accurately match the right vendor and push the process forward.

The intelligent production platform is equipped with 12 key software modules for intelligent production, including OES, APS, MES, MMS, EMS, TPM, SCADA, EAM, SPC, WMS, etc.; it enables intelligent production scheduling, real-time monitoring, accurate distribution, planning, and power optimization. Through its deployment, factory mass customization is supported, and million-class customization can be achieved, eliminating the distance between factories, users, and resources.

The smart logistics platform consists of smart operation and visualization, covering platform reservation management, smart logistics TMS, distribution collaboration platform, logistics trajectory visualization and intelligent vehicle management platform, etc. Eight software products are provided including VOM system, reservation platform, smart vehicle distribution system (TMS), HLES3.0, WMS, hub distribution and mobile app system, etc., with the aim to bring users the best one-stop service experience.

The smart service platform creates a new home appliance service mode, and it meets the user's demand for timely maintenance of home appliances. Users can input information of the home appliance by one key on the platform and create and upload the exclusive file of the home appliance, which can completely replace the traditional paper warranty card, and the information will never be lost.

Based on the platforms, the following new features and highlights have been realized:

- Interaction: autonomous learning, demand forecasting, swarm intelligence.
- Design: system intelligent design, collaborative R&D, virtual reality verification and optimization.
- Experience: precise procurement, decision optimization, and deep mining of cooperative resource side.
- Pre-sale: user portrait, sales forecast; feature analysis, marketing decision.
- Manufacturing: intelligent equipment, predictive maintenance, intelligent production scheduling, intelligent quality monitoring.
- Iteration: user image, sales forecast, feature analysis, marketing decision.
- Human-computer interaction, community interaction, and intelligent life scene are upgraded iteratively.

26.2 Data Service and Data Flow of COSMOPlat

26.2.1 Current Data Service

On one hand, Haier COSMOPlat connects with users through interaction with interaction sub-platform and service sub-platform. On the other hand, through the IoT sub-platform and manufacturing enterprise instrumentation, industrial equipment, industrial control system, host computer connection, and the protocol conversion function, the unified MQTT Protocol will upload data to the cloud for interaction, design, procurement, sales, production, logistics, and service.

IT provides micro-services based on Springcloud and Springboot architecture and softwares such as MES, PLM, CRM, etc. to microenterprises by SaaS subscription. They are deployed to multi-tenant-based containers and are flexibly scalable; besides, capabilities including load balancing and disaster recovery backup are also possible as new features.

IT manages various capabilities in the COSMOPlat main platform, including:

- IoT layer: sub-platform of IoT cloud device management, sub-platform of edge computing, smart gateway boxes integrated with industrial protocols.
- IaaS layer: container data center.
- PaaS layer: sub-platform PaaS cloud management, providing cloud-based public/private cloud service; developing sub-platform developer.
- SaaS layer: tenant center, payment center, order center, logistics center, etc.

- BaaS (Business and Best Practice) layer: mechanism model sub-platform, big data BI sub-platform, big data transaction sub-platform, AI sub-platform, blockchain sub-platform, food security code sub-platform (provide end-to-end production, wholesale, distribution, retail, and transportation tracking of frozen food), and COSMOPlat Tiangong OS (a unified portal that enables users to access multiple industrial apps/micro-services).

26.2.2 Data Flow Based on COSMOPlat

The data generated in the business interfaces (portals, apps, public accounts, etc.) of microenterprises are uploaded in real time to the data mid-office of the COSMOPlat main platform, which is managed by IT.

The types of data being collected include:

- Financial data: number of paid users, conversion rate, etc.[1]
- Ecosystem data: number of registered business users, number of registered users, etc.
- Capability data: number of connected endpoints, number of industrial apps, etc.
- Third-party resource indicators

IT monitors, analyzes, and governs these data to understand business landscape across the company, and it is also further co-developing a framework with external business partners of a secure, confidential and compatible data inter-operation model. The below RV (Recreational Vehicle) case will help you have a general view on how data flows between COSMOPlat and our customers, and related interactions in between (Fig. 26.2).

26.3 Use Cases of MCM

26.3.1 Best Practices of Big Data Technology

There are three typical application directions of big data in manufacturing industry: One is intelligent production, which enables the production system to have the ability of agile perception, real-time analysis, independent decision-making, accurate execution, as well as learning and improvement, so as to comprehensively improve the production efficiency. The other is intelligent products, which encapsulates the trained artificial intelligence system into the hardware and endows the products with the ability of intelligent response to external changes and user needs.

[1] MQTT Protocol. MQTT is an OASIS standard messaging protocol for the Internet of Things (IoT) (MQTT.ORG).

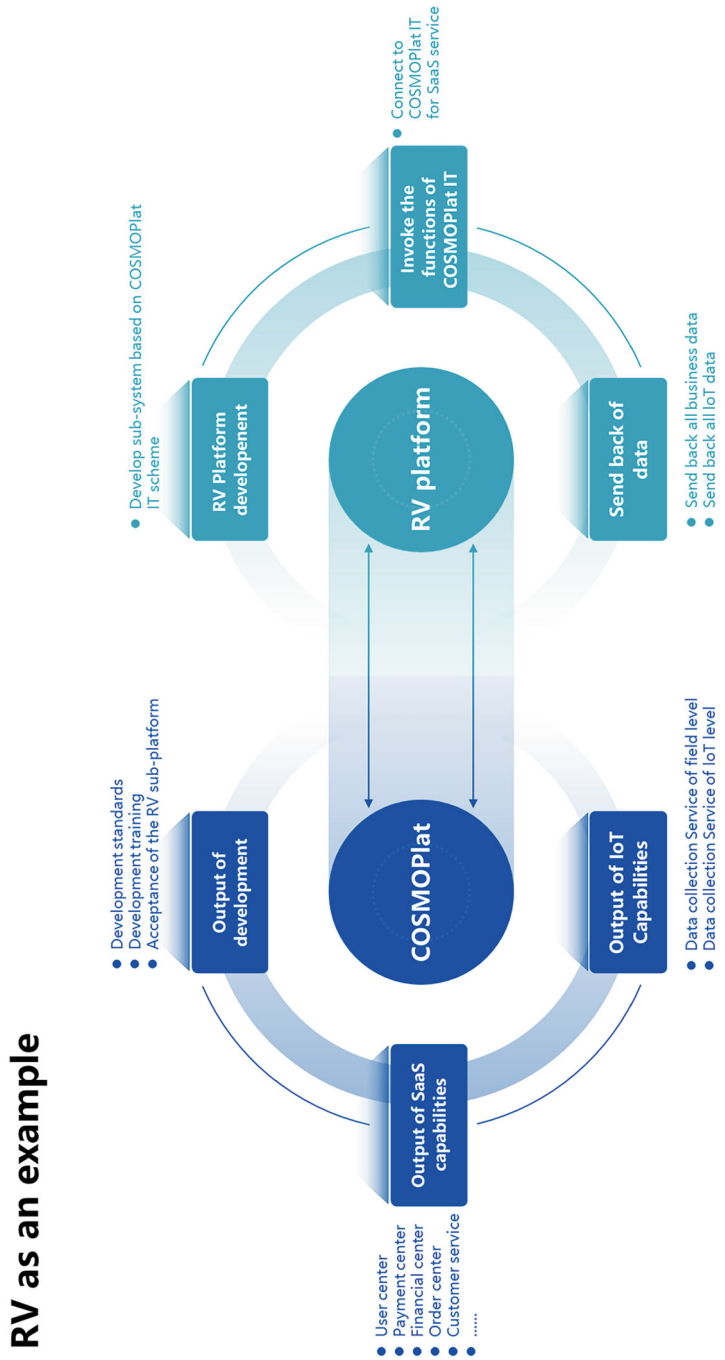


Fig. 26.2 How data flows in RV sub platform

The third is intelligent services, which can monitor the status of products in real time, quickly respond to the needs of users, and provide value-added services such as sale by rent, billing on time, remote diagnosis, fault prediction, remote maintenance, and integrated solutions, promoting the transformation of manufacturing enterprises from providing "products" to providing "products + services."

UC1: User Interaction

Big data combined with deep learning to make intelligent prediction of user demand.

In the past, if the research and analysis of user needs is inaccurate, it will most probably lead to a long marketing cycle, and market changes are not synchronized. In traditional ways, user needs are fragmented and personalized; user's ideas are collected one by one, making it nearly impossible to form a feasible scheme to achieve precision marketing and data transparency; and user behavior and user tag trajectory are not visualized. In this sense, user product demand prediction needs to be changed to user scenario prediction so as to create user experience value.

In the era of IoT (Internet of Things), the consumer market is undergoing a trend change from the traditional "enterprise-centered" mode to "user-centered." COSMOPlat has established a community interaction platform for direct communication with users, turning the uncertainty of users' needs into certainty, and the community interaction solutions can also meet the changing needs of users. Based on the social platform and scene interaction, a large number of users with "high aspirations" can gather 100,000 or even more users in a short time. Based on the Internet platform, one user's pain points and needs become a group of people's (10,000, 100,000, or more). Finally, it will realize the personalized customization and large-scale sales of household appliances.

Take Haier's infant and mom customization as an example: the mini dryer originated from the wisdom of mothers in the infant and mom community. Last year, maternal and infant groups put forward the demand for clothes dryers, which has raised heated discussion among the maternal and infant community. The interactive exposure was more than 3 million, user's participation in the interaction reached more than 100,000 times, the 180-day interactive product design took more than 10 iterations according to the user's interactive demand, and eight product functions were iterated. At present, the product has been put onto the market, and its sterilization rate reaches 99%+ and is capable of collecting threads and scraps. The inner barrel has the function of forward and reverse rotation and regular shaking, which can prevent clothes from winding or wrinkling. In addition, the drying process follows six professional procedures, making it possible to automatically sense the humidity of clothes so as to meet various needs of different users.

UC 2: Customized Experience

Big data combined with deep learning to create sleep curve of Haier Tianzun air conditioner

When heating season comes, many consumers will turn on their air conditioner overnight for heating. Due to the great temperature difference between day and night, the air conditioning also needs to adjust the temperature frequently, which is the headache of many air conditioning users. The birth of Haier Tianzun air

conditioner provides intelligent indoor temperature solutions for users, bringing an end to the complaint of frequent temperature adjustment during sleep. By studying and analyzing different needs of the elderly, children, and people from the north and the south, Haier launched an automatic sleep curve to bring users a peaceful and comfortable sleep environment.

Users can change the sleep curve according to their own preferences and name it with their favorite names. What's more humanized is that "sleep curve store" is also provided in the app, and users can choose more functional modules according to their own needs. The intelligent sleep curve mode completely solves the problem that the air conditioner does not match human body temperature during sleep and frees consumers from adjusting temperature with a remote control all the time.

After deeply exploring the sleep curve of Chinese people, Haier Tianzun air conditioner provides diversified sleep modes according to different seasons, user groups, and genders. The summer curve and winter curve classified by season also provide a variety of choices according to different time periods. Users can choose the 12:00–14:00 sleep curve in the nap period accordingly, and they can also choose the 23:00–08:00 sleep curve when night comes. For different users, such as the elderly and children with low immunity, Tianzun air conditioner will automatically select a soft sleep curve with low temperature difference to control the indoor temperature. For adults who are greedy for cool or warm environment, quickly reaching the consumer's favorite temperature would be a highlight function of Tianzun intelligent sleep curve to enhance user's comfort. In addition, the sleep curve can be remotely controlled by using smart phones, pads, and other terminal devices. Just download the app of "Haier intelligent air conditioner," and you can choose to turn on the sleep curve. Tianzun air conditioner will adjust its operating temperature and state according to your settings in the app.

UC3: Machine Visual Auto-inspection of Refrigerator Appearance

In the traditional production process of refrigerators, the appearance detection remains a bottleneck; high labor intensity of employees, complicated defect detection procedures, and various manual detection standards lead to the low detection rate of defective product to be only 92%, causing customer's complaints from the market.

By statistically analyzing on-site problems, Haier gathered first-class resources to develop a visual automatic inspection system (Fig. 26.3) for the detection of refrigerator appearance. The system conducts online quality inspection for refrigerator appearance by intelligent learning of qualified products. Inspection scope includes printed matter, door body and appearance, etc.

By using the appearance visual automatic detection technology instead of manual work, the automatic detection efficiency is improved by more than 50%, and detection rate reaches more than 99.5%. In addition, the detection information is visualized, and unqualified products will be kept on production line and not be sent to the market. The application of the system not only improves the production quality and efficiency but also increases the user satisfaction.

UC4: Voice Recognition Combined with In-depth Analysis to Provide Product Online Noise and Abnormal Sound Detection Solution.

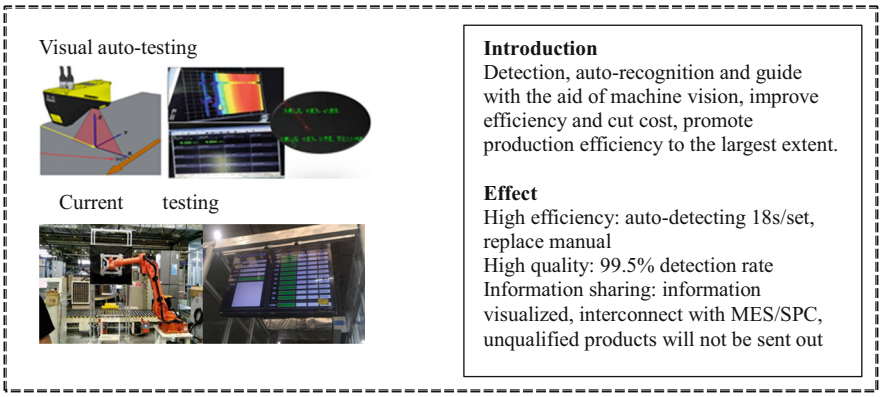


Fig. 26.3 Automatic visual inspection of appearance

In a typical detecting environment, the background noise is lower than 30 dB; in a traditional way of manual detection of the decibel value, the error is quite large. For each detection, the abnormal sound is tested through earphone, which contains the problem of voice leakage, as well as the difficulty to record the written statistics and poor controllability of the maintenance state, making it even more difficult to get an accurate result. Regarding this, a more advanced detecting method is becoming more and more people's prospect.

The COSMOPlat cloud platform collects the users' needs from Microblog, WeChat and search engine, etc., finds out various needs of users for all brands of air conditioning, and takes the air conditioning sound as the main concern through data analysis.

Air conditioning sound mainly includes noise and abnormal sound, and their level can be distinguished by decibels. There are tens of millions of different sounds. Relying on big data and artificial intelligence technology, a platform for online noise and abnormal sound detection (Fig. 26.4) is built. The platform can independently

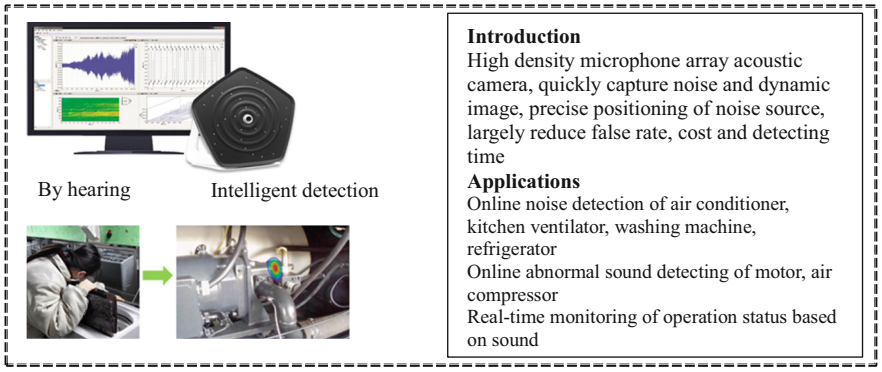


Fig. 26.4 Online noise and abnormal sound detection

learn to identify abnormal sound and conduct automatic control and reduce the miscalculation of artificial detection, so as to improve the accuracy of identification and reliability of the product.

While focusing on noise problems, the production process is traceable. By studying the big data of production process, the causes of abnormal sound (including poor installation of air conditioning fan and motor or the main frame) are analyzed, and key improvement measures for further prevention in advance are then summarized.

The independent learning to identify abnormal sound and control not only builds a modern detection way consisting of automatic isolation and transfer, real-time information interconnection, automatic storage, traceability, automatic process traceability, and problem closed loop, which largely improves the accuracy of identification, but also brings the benefit of labor cost saving of 5 million; more importantly, user experience is greatly improved during this process.

26.3.2 Value Chain of Use Cases

In the value chain of Haier's personalized customization solutions (Fig. 26.5), there are three parties: industrial Internet platform enterprises, home appliance manufacturers, and home appliance buyers. For home appliance buyers, the solution solves the user's demand for personalized customization and improves the user's experience of purchasing and using home appliances. For home appliance manufacturing enterprises, the pursuit of "improving quality, reducing cost, and increasing efficiency" has been realized. Last but not the least, industrial Internet platform enterprises can continue to generate revenue in the process of providing services for users and manufacturing enterprises.

At present, MCM has built a transformation service solution for whole industry users to get whole process participation globally, covering 15 industry ecologies including clothing, emergency supplies, agriculture, etc., established 7 centers nationwide, covered 12 regions, and has been replicated and promoted in 20 countries.

With the promotion of MCM, COSMOPlat has got comprehensive leadership in the following three dimensions:

- Standards leading. COSMOPlat is the only industrial Internet platform authorized by ISO, IEEE, and IEC to develop MCM standards.
- Model leading. At the end of 2019, COSMOPlat ranked No. 1 Industrial IoT Software Platform of Q4 2019 by Forrester, the world's leading market research and consulting organization.
- Brand leading. As an ecological brand, COSMOPlat is the only industrial Internet platform in the top 100 Chinese brand value list in 2019.

From the leader of China's "lighthouse factory" to the winner of German industry 4.0 award, MCM has fundamentally solved the problem of SMEs having customers

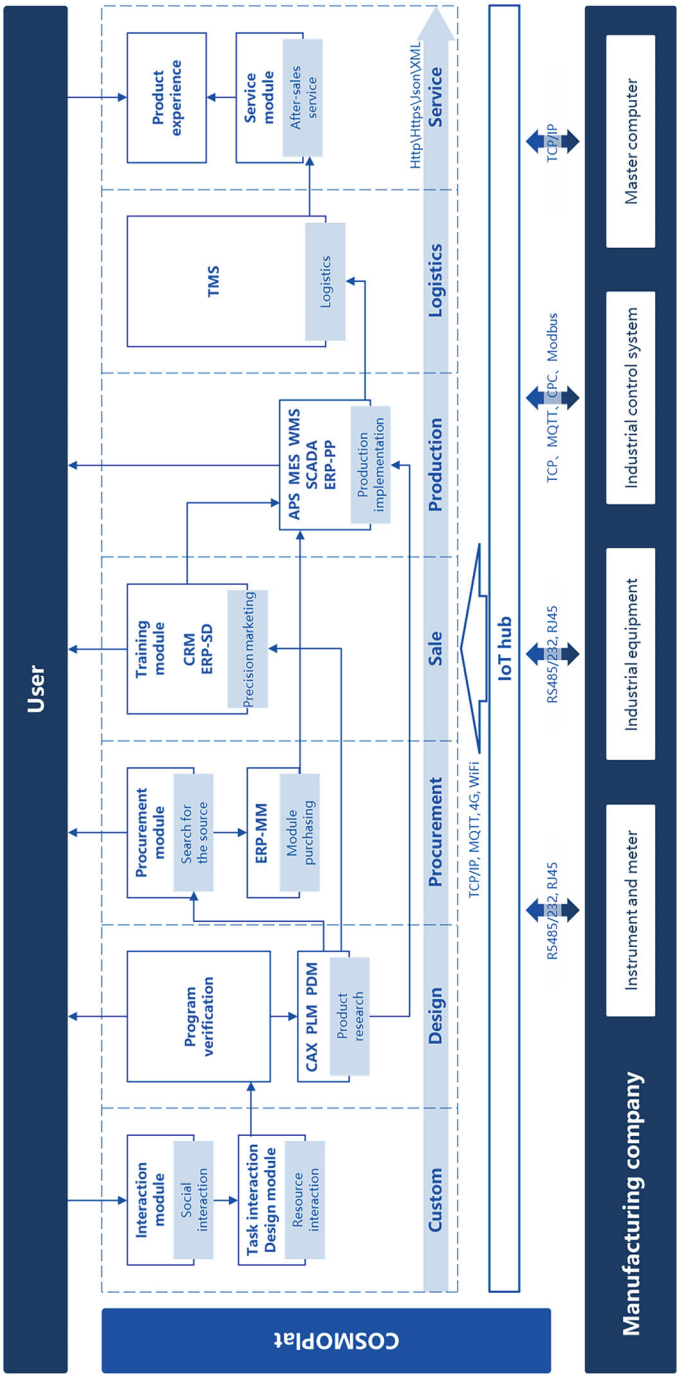


Fig. 26.5 Value chain of user cases

but no users, empowering SMEs for digital transformation and upgrade while promoting integration and innovation of traditional industries with industrial Internet. Based on the ecological system of multilateral interaction, co-creation, and sharing, it gradually completes the transformation from "lighthouse factory" to "lighthouse base", embracing an overall industry upgrade which leads to a high-quality development for enterprises as well as the entire economy.

References

1. Lucheng, C. (2021). Haier COSMOPlat: To identify a new path for the integration of industrialization and industrialization and open up a new ecological path for industrial internet [N]. *China Electronics News, 11*(3).
2. Lucheng, C. (2019). Haier COSMOPlat: Building and fostering a new and ever-growing ecosystem for the industrial internet [N]. *China Industry News, 25*(3).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 27
Huawei and International Data Spaces



Martin A. O'Brien, David Mohally, Götz P. Brasche, and Andrea G. Sanfilippo

Abstract In a digitalized and deeply interconnected industrial ecosystem, it is of paramount importance to create mechanisms that seamlessly guarantee standardization and verifiability, interoperability (including regional legislation), transparency, and trustworthiness, in particular in the intermediation of all businesses and stakeholders, including SMEs. In this respect, the International Data Spaces and GAIA-X initiatives pave the way to a framework for collaboration that use secure and trusted data services to safeguard digital sovereignty.

Huawei is a leading ICT provider, operating in more than 170 countries, and an active member of more than 600 standardization bodies and industry associations, among them IDS and GAIA-X. With their international footprint, IDS and GAIA-X are of the utmost importance for Huawei in Germany, Europe, and globally. Here, we provide a brief overview of our understanding of the data ecosystem's inherent issues, and how regulations address them, followed by more specific examples of how IDS and GAIA-X objectives can be supported by Huawei, with emphasis on validating concepts in the manufacturing domain as a prominent reference example for an important European market vertical. We illustrate the use case of an evolved implementation of Industry 4.0 that integrates machines and cloud services. 5G connectivity for trusted networked production is added to the example. We finally highlight the use of GAIA-X compliant federated AI for control, maintenance, and match-making of demand and supply actors.

M. A. O'Brien (✉) · D. Mohally
Service Provider Operations Lab, Ireland Research Center, Huawei, Dublin 2, Ireland
e-mail: martin.o.brien@huawei.com; davidmohally@huawei.com

G. P. Brasche
Intelligent Cloud Technologies Lab, Munich Research Center, Munich, Germany
e-mail: goetz.brasche@huawei.com

A. G. Sanfilippo
Technology Planning and Cooperation, Munich Research Center, Munich, Germany
e-mail: andrea.sanfilippo@huawei.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_27

27.1 Inherent Issues with Ecosystems

Digitalization, including data and digital infrastructures, is substantially transforming industry landscapes, in particular affecting the production, logistics, and health sectors, to name a few. In the current global framework, German and European firms at large are challenged to embrace digitalization to secure their competitiveness. In this respect, International Data Spaces (IDS) [1] provide an IT architecture with the aim of safeguarding data sovereignty. In order to understand how IDS helps, it is important to firstly understand, and classify, the nature and inherent issues with ecosystems, and secondly the concerted actions of regulators inside and outside Europe.

27.1.1 Functions and Natural Characteristics of Ecosystems

Firms need to interact with one another within ecosystems for mutual benefit to deliver their offerings to a market. The level at which this occurs can often depend on a host of factors. Two really important factors are the level of modularity and the level of coordination needed [2].

Level of modularity is the degree to which components of a system can be produced by different producers. For example, vehicle components tend to be produced by just a few suppliers, whereas construction materials tend to be produced by many firms and can be used to create any blueprint.

Level of coordination is the degree to which firms need to cooperate with one another to bring an offering to the market; again, vehicle component manufactures need to work closely with car manufacturers to create vehicles, while construction material firms and home builders largely don't.

These factors influence how industries are organized and their characteristics. Figure 27.1 shows a representation of industrial ecosystems with respect to varying levels (low to high) of modularity and coordination.

In the so-called vertically integrated ecosystem, the hub firm (buying firm) takes ownership of the supply chain because the level of coordination and control it can bring improves the value proposition, e.g., Tesla™ vs Nissan™ electric car charging or Apple™ vs Google™ phones.

In the hierarchical supply chain ecosystem, the hub maintains hierarchical control—not by owning its suppliers but by fully determining what is supplied and at what cost.

In business ecosystems, members retain control and claim over their assets. No one party can unilaterally set the terms, for example of, pricing and quantities. Standards and decision-making processes are mostly distributed, even though standards, rules, and interfaces are often set by hub firm(s).

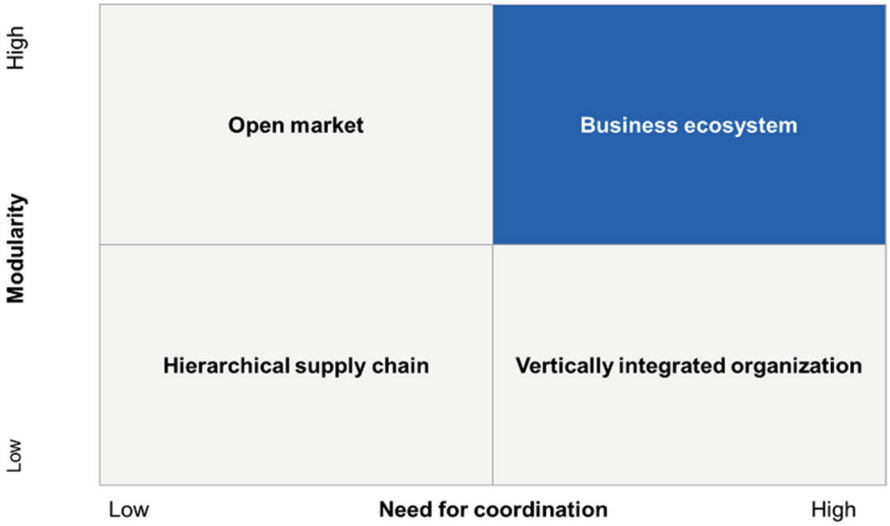


Fig. 27.1 Representation of the levels of modularity and coordination [2]

In open market ecosystems, firms build offerings from generic modular components and require little coordination. Often components have multi-purpose utility, e.g., nut and bolts and sheet metal.

Within the technology sector, business ecosystems are becoming the most dominant; firms use business ecosystems for a number of reasons [2]:

- They face an unpredictable but highly malleable business environment that requires them to collaborate with others in order to shape or reshape the industry for the common good.
- The scope of required capabilities are too broad to be kept in-house.
- The individual components of the end-user's solution are modular; however, requirements for coordination are high, for example, firms need to identify the required third-party partners, specify their roles, and align their activities.
- They can benefit from access to external capabilities and data that enable fast scaling of the ecosystem.

Technology business ecosystems tend to exist in three forms, often revolving around two-sided or multi-sided platforms [2, 3].

Fixed Core Ecosystem

Orchestrator offers the core thereby allowing the participants to tailor the rest of the offering based on a predefined list of functions or by allowing partners to define new use cases, e.g., Nespresso™ and Apple™.

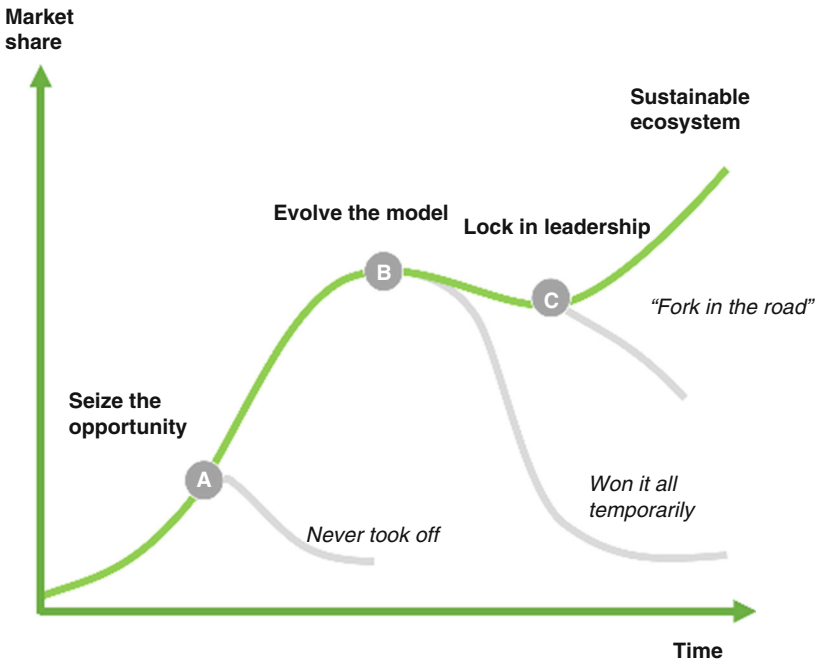


Fig. 27.2 Life cycle, market share as a function of time, cit. Martin Reeeves et al. (2019), MIT Sloan

Transactional Ecosystem
Orchestrator serves as an intermediary (offering a multi-sided platform) for standardized transactions between participants of an ecosystem, e.g., Uber™ and eBay™.

Innovation Toolkit Ecosystem
Communities of people or organizations that work toward the development of solutions that draw on common standards, e.g., Linux.

When functioning well, platforms can take fragmented and inefficient ecosystems and transform them through acceleration of industry standardization and orchestration of value creation, to the benefit of most stakeholders. This has been seen in many industries from cloud computing to short-term rentals. It is imperative for platform orchestrators to maintain control, to maximize returns, but this eventually comes at a high cost to complementors, as the business ecosystem life cycle matures.

Technology platform-based ecosystems tend to follow a life cycle shown in Fig. 27.2.

(A) *Seize the Opportunity*

- Be the first or radically disrupt, scale fast investing persistently and sufficiently.
- (Cumulative success rate 50%)

(B) *Evolve the Model*

- Broaden ecosystem scope and increase engagement.
- (Cumulative success rate 25%)

(C) *Lock in Leadership*

- Manage vested interests, maintain differentiation, and renew the platform.
- (Cumulative success rate 15%)

Creating a platform-based business ecosystem often depends on good timing and rapid investment (A) once winner-takes-all and network affects take hold and the market has "tipped" (B) those few that succeed can generate deeply entrenched advantages over all challengers.

As business ecosystems evolve, business ecosystem orchestrators tend to consolidate control of their ecosystems; (C) this is often where control starts to shift from a distributed control to centralized control; this pattern has replayed in many industries from mobility to operating systems [4].

At their core, business ecosystem platform orchestrators want to maintain control over their complementors by determining who can access the platform and under what terms. They become the global regulators controlling competition, price, licensing, and enforcement [5]. Platform orchestrators tend to seek to own the control points on the platforms that determine who can and can't access the ecosystem and under what terms. Common control points include ownership of key platform APIs, access to key data, and ownership of key platform services or IPR.

Platforms have to maintain uneasy paradoxical balances; the biggest danger to platforms can come from the complementors they orchestrate, which are well characterized by the platform paradoxes [6]:

- The Control Paradox—Too much control drives partners away, too little hurts standardization and orchestration of value creation.
- Weak Partner Paradox—Strong partners view you as a threat; your future depends on weak partners that can be controlled.

When platforms become very large, controls imposed in the lock-in phase can become overbearing, and this can ultimately stifle innovation.

Platform-based ecosystems are a massive part of the current and future economy, and regulators around the world are becoming aware of their negative externalities and are taking actions to address this.

In this context, the European Strategy for Data primarily aims to create a single market for data that will ensure Europe's global competitiveness and data sovereignty and to create a strong legal framework—in terms of data protection, fundamental rights, safety, and cybersecurity—in its internal market with competitive companies of all sizes and varied industrial base. The challenge is highlighted by a document published by the European Commission in February 2020 states [7] "A small number of Big Tech firms hold a large part of the world's data. This could reduce the incentives for data-driven businesses to emerge, grow and innovate in the EU."

International Data Spaces along with regulatory oversights are key to addressing governance challenges that platform economics bring to the data economy.

27.2 Concerns and Actions of Regulators

The market size of global platforms were measured [3] at $7 trillion this exceeds the combined GDP of Germany and France. A survey by the Centre for Global Enterprise [3] showed that North American and Asian platforms are currently dominating the digital economy. Europe's digital competitiveness has been a concern for some time. Surprisingly, while Europe has emerged as an important consumer of platform services, it has generated relatively few platform companies [3] (Fig. 27.3).

The failure of European companies to seize opportunities and "take-it-all" has been attributed to high initial scaling costs in the European digital market which is fragmented in terms of language, consumer preference, and rules and regulations compared to more uniform markets like the USA and China. The risk-averse nature of European investment culture and the lack of supports from governments is another. Europe's Strategy for Data is the plan to address these and other systemic problems. It has a number of aims that are outlined in Fig. 27.4.

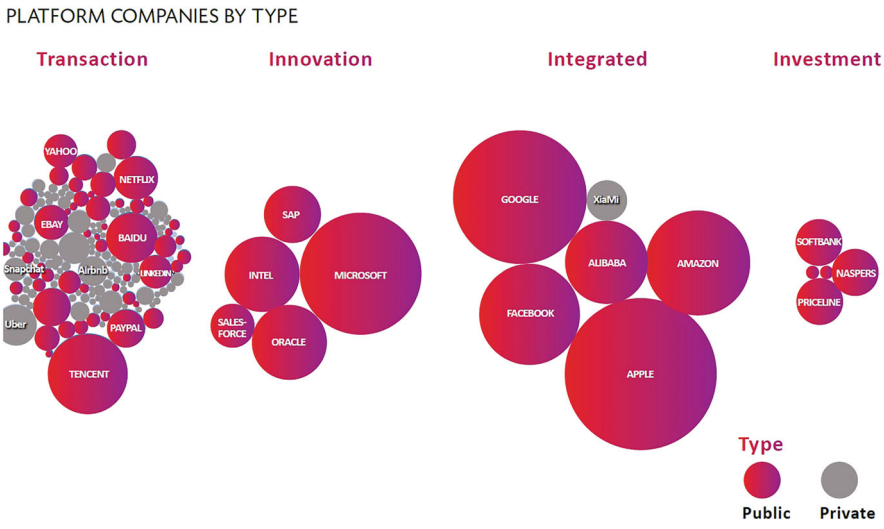


Fig. 27.3 Platform companies by size [3]. Source: Global Platform Survey, The Center for Global Enterprise, 2015

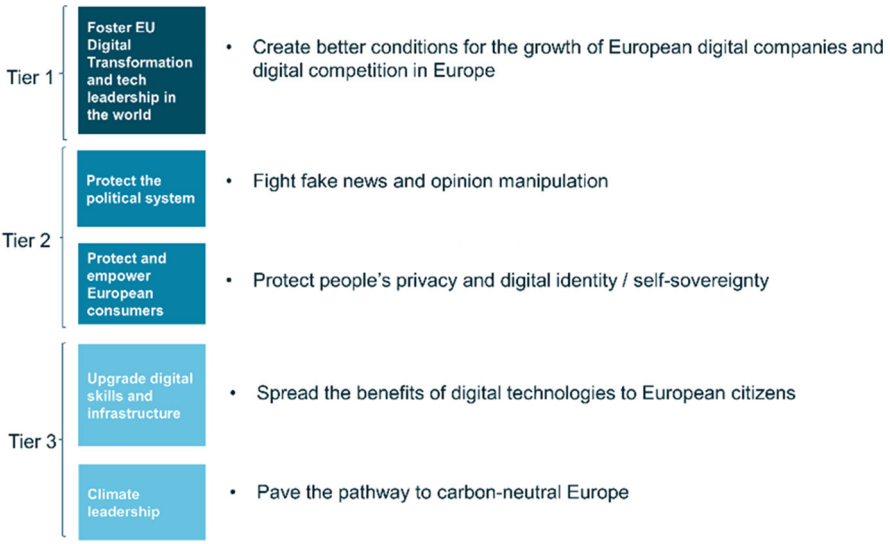


Fig. 27.4 Europe's Strategy for Data [7]

Backing up this policy, the EU has prepared legislative actions designed to regulate digital business such as the Digital Services Act (DSA) [8, 9], Digital Markets Act [10] (DMA) and the Data Governance Act [8] (Fig. 27.5).

27.3 Issues with the Data Ecosystems and How Europe Intends to Address It

The EU in its Strategy for Data [7] sees several issues holding Europe back from realizing its potential in the data economy.

Availability of Data
It aims to increase the use of public sector information by business (G2B—data sharing), and the use of privately held data by other companies (B2B, data sharing).

Imbalances in Market Power
Small numbers of players may accumulate large amounts of data, gathering important insights and competitive advantages from the richness and variety of the data they hold. The "data advantage" can enable large players to set the rules on the platform and unilaterally impose conditions for access and use of data.

Data Interoperability and Quality
Application of standard and shared compatible formats and protocols for gathering and processing data in an interoperable manner across sectors and vertical markets should be encouraged through the rolling plan for ICT standardization.



Fig. 27.5 Synoptic table of the EU legislation actions for DSA, DMA, and data governance [8]

Data Governance
Enforcement of data governance in society and the economy.

Skills and Data Literacy
General data literacy in the workforce and across the population is relatively low, and participation gaps exist.

Data Infrastructures and Technologies
The EU believes it needs to reduce its technological dependencies in these strategic infrastructures; it sees difficulties in the supply and demand side.

On the Supply Side

- EU-based cloud providers have only a small share of the cloud market. The Hyperscalers represent more than 70% of cloud provisioned in Europe; figures from Synergy [11] show that since 2017, the European cloud market has grown more than threefold to €5,9 billion in the third quarter of 2020, but the European market share has declined from 26% to 16%.
- Subjectivity to legislation of third countries and compliance of cloud service providers with important EU rules and standards.
- Sector-specific data spaces, lack of ontologies, and widely accessible application programming interfaces (APIs) limit the applicability of comprehensive and EU-wide solutions.

On the Demand Side

- There is a low cloud uptake in Europe (1 company in 4, only 1 in 5 SMEs).
- European businesses often experience problems with multi-cloud interoperability, in particular data portability.
- Currently digital service users cannot make self-determined decisions on data use, due to a lack of control and transparency over stored or processed data.

Empowering Individuals to Exercise Their Rights
Potential of Article 20 of GDPR to enable novel data flows; there are calls to give individuals the tools and means to decide at a granular level what is done with their data.

Europe hopes to accomplish these aims through:

A. A cross-sectoral governance framework for data access and use
B. Enablers: Investments in data and strengthening Europe's capabilities and infrastructures for hosting, processing, and using data.
C. Investment in a High Impact Project centered on European data spaces and federated cloud infrastructures
D. Competences: Empowering individuals, investing in skills and in SMEs
E. Creating common European data spaces in strategic sectors and domains of public interest
F. An open, but proactive international approach

In this context, the GAIA-X project [12, 13], initiated by various European countries' representatives (22 founding members from France and Germany,

comprehensive of industry and other organizations), was kicked off in 2019, and it was given a structure in 2020, with a non-profit association called GAIA-X AISBL. The aim of GAIA-X is to establish a framework for collaboration and enable, or accelerate, the use of secure and trusted data services, with emphasis on SMEs, leveraging existing open standards. Great attention has been given to data processing and storage (EU/EEA area), transparency (incl. applicable jurisdiction), cybersecurity and portability (incl. practices to facilitate the switching between providers).

27.4 Role of IDS in Helping to Address these Concerns

Clearly International Data Spaces capabilities together with GAIA-X will play a key role in helping Europe achieve its stated goals of establishing an attractive secure and dynamic digital economy by providing the standards and trust infrastructure required to enable the establishment of European data spaces.

However, IDS and GAIA-X are not enough on their own to address imbalances in market power as well as data interoperability and governance issues.

To address this, the EU is preparing ex ante regulatory proposals aimed at large cloud "gatekeeper" platforms which will address this issue.

Articles 5 and 6 of the proposed Digital Markets Act [10] will compel cloud gatekeepers to:

> 5(e) refrain from requiring business users to use, offer or interoperate with an identification service of the gatekeeper
> 6(f) allow business users and providers of ancillary services access to and interoperability with the same operating system, hardware or software features that are available or used in the provision by the gatekeeper of any ancillary services.
> 6(h) provide effective portability of data generated through the activity of a business user or end user and shall, in particular, provide tools for end users to facilitate the exercise of data portability, in line with Regulation E 2016/679, including by the provision of continuous and real-time access.
> 6(i) provide business users, or third parties authorized by a business user, free of charge, with effective, high-quality, continuous and real-time access and use of aggregated or non-aggregated data [. . .].

These measures and investments will allow data to flow more freely and facilitate the establishment of data spaces through IDS and GAIA-X which in turn will increase contestability and innovation.

27.5 Laws, Regulations, and National Standards in China on Data Protection

Around the world, at a governmental and societal level, there has been a push towards a more explicit set of rules to protect consumer data and privacy as technology services continue to fulfil an increasingly more important role in



Fig. 27.6 KPMG: Overview of Draft Personal Information Protection Law in China [14]

economics and everyday life. China is no exception in this respect; the cadence of laws, regulations, and national standards in China on data protection has been especially intense in recent years (Fig. 27.6).

On April 26, 2021, following public consultation, the second draft of the Personal Information Protection Law (PIPL) was submitted to the Standing Committee of the

National People's Congress (NPC) of China. Similarly to Europe's GDPR, it lays down a comprehensive set of rules around data collection and protection [14].

Liu Junchen, deputy director of the Legal Affairs Committee of the Standing Committee of the NPC, noted the importance of the protection of personal information for the development of the digital economy in China:

> . . . the formulation of a personal information protection law is an objective requirement to further strengthen the legal protection of personal information protection; is a practical requirement for maintaining a healthy cyberspace; and is an important step in promoting the healthy development of the digital economy [14].

China's PIPL applies to the country's citizens and to companies and individuals handling their data. It contains 70 articles and shares many of the same principles and values of GDPR including transparency, fairness, purpose limitation, data minimization, limited retention, data accuracy, and accountability [14–17].

In particular, the key parts of the law:

- Clarify the role and liabilities of the data processor.
- Set out the lawful basis for the processing of data.
- Specify that consent must be informed, specific, freely given, and indicative of the wishes of the data subject; it also details the additional consent required for the processing of sensitive personal information.
- Define the data subject's rights: right to information on data processing, right to access and request copy of personal data, right to correction, right to object to processing, right to withdrawal of consent, and right to deletion.
- Set out data localization requirements and clearer rules on cross-border transfer of personal data.
- Define the protective obligations for data processors that hold personal data, such as regular compliance audits, risk assessments, data breach reporting, and remedial measures in response to data breaches.

Serious violations of the PIPL, such as illegal processing of personal data or failure to adopt necessary measures to protect personal data, can result in fines of up to $7,4 million or up to 5% of the preceding year's revenue. There are also terms for personal liability in the context of a violation of the PIPL. Like GDPR, PIPL will apply extraterritorially to protect the interests of Chinese data subjects.

Global Data Security Initiative

In order to address data security risks and enhance dialogue and cooperation in cyberspace governance, China has issued the Global Data Security Initiative [18], which proposes three principles and eight initiatives for global data security governance. These principles include agreeing global data security rules through consultation and co-operation, balancing security and development, and adherence to the principles of fairness and justice. These principles provide a feasible path for the healthy development of data protection and cyberspace. As a leading global information and communications infrastructure service provider, Huawei will maintain consistent cooperation, transparency, and openness in the field of data security,

actively participate in global industry collaboration, and contribute to building a fair, secure, trustworthy, and stable cyberspace.

Huawei's Data Lake Governance Center (DGC) [19] can help build an enterprise-class data governance platform and is used by both public and private sector enterprises. With DGC, organizations can manage data ingestion, data governance, data development, data services, and data visualization from end to end. DGC is focused on the elimination of data silos, compliance with data standards such as IDS and GAIA-X, acceleration of data monetization, and facilitation of digital transformation [19, 20].

27.6 Why International Data Spaces Are Important

IDS will help unleash innovation in the data economy.

Platform orchestraters can enable considerable value creation that can be difficult for complementors to resist, but they pay a price; the platform orchestrator becomes their regulator, and it often controls aspects such as access, standards setting, competition, price, licensing, and enforcement [5].

There are also considerable business risks as the platform life cycle matures: the risk of the platform assimilating core product functions, platform lock-in and lock-out, security and sovereignty issues, loss of control of emerging industry standards, data sovereignty, and decision-making. Many industries are rightly unwilling to accept these risks or hand over control.

There are many reasons why firms want to maintain control and sovereignty over their data, and these were outlined by Prof. Achim Wambach at the GAIA-X Summit on November 18, 2020 [12]: To protect user's privacy (there are severe penalties for failure here), to protect sensitive data and trade secrets, and to gain a return on investment.

Many sectors of the economy handle sensitive data that depends on trust and security such as manufacturing, health, energy, telecoms, and government services, to name a few. Often, in these sectors, data is also fragmented and in heterogeneous formats.

Trust and interoperability are key, and the data economy in these sectors will not reach its full potential unless: a trust infrastructure is in place that ensures data can be shared accurately and securely, there is interoperability and data portability, and parties act in good faith and are verified. IDS and GAIA-X can provide this trust infrastructure and be the anchor for data ecosystems.

IDS and GAIA-X can be the trusted regulator in the data economy [12] orchestrating value creation, setting standards, offering federated platform services, and incentivizing and encouraging the exchange of data through data portability and interoperability. This will lower switching costs and also reduce co-dependencies.

The benefits are numerous [12]:

- Direct benefits from the data, for example, the ability to create sophisticated AI models used on a host of applications through data partnerships.
- Reduced ecosystem transaction costs between business ecosystem participants of all types leading to more efficient supply chains.
- Prevention of captured and siloed data markets.

27.7 Specific Examples on How IDS and GAIA-X Can Be Used from Huawei Perspective

The International Data Space Association (IDSA) concepts have been successfully prototyped by Huawei's internal IT department and applied to a real business scenario with one of its supply-chain ecosystem partners which requires daily bidirectional data exchange of sensitive production configuration and pricing data as well as confidential design and development documents.

Also for the early validation of the emerging GAIA-X concepts, Huawei started several activities to prototype pilot use cases together with partners in joint R&D projects.

1. Networked Production

   Under the umbrella of a GAIA-X project co-funded by the German government, a use case that had been tested as a German-Dutch demonstrator for a 3D production environment for individualized USB sticks is being extended to validate the feasibility of the concept of trusted and GAIA-X-compliant networked production. Conceptually, the use case is an evolved implementation of a state-of-the-art production process that builds upon Industry 4.0 standards. It demonstrates two physically distributed factories, third-party ISVs and associated cloud services cooperating in the production of a USB stick. Partners have the option to personalize the USB stick by pre-installing a software package, and they also have the option to change the individual form factor and color based on the request of a field-test customer. The integration of production is carried out across company boundaries, and the offering requirements of the value-added partners are instantiated according to the individual production order demands. Cloud services are also used to monitor the production and ensure quality control.

Furthermore, 5G connectivity is incorporated into the trusted networked production example to enable the integration of remote cloud-based services for real-time feedback, as illustrated in Fig. 27.7. Retrofitting of legacy machines, e.g., by deploying smart AI-services from the edge, will also be demonstrated as part of the project.As a secondary use case, remote servicing will also be implemented; this will enable the ad hoc provisioning of maintenance based on the concepts of a common application catalogue, identity services, and federation services.

2. Federated Trustworthy AI

   With a view to the increasing use of AI across all market verticals and distributed market players along multidimensional value chains, there is another

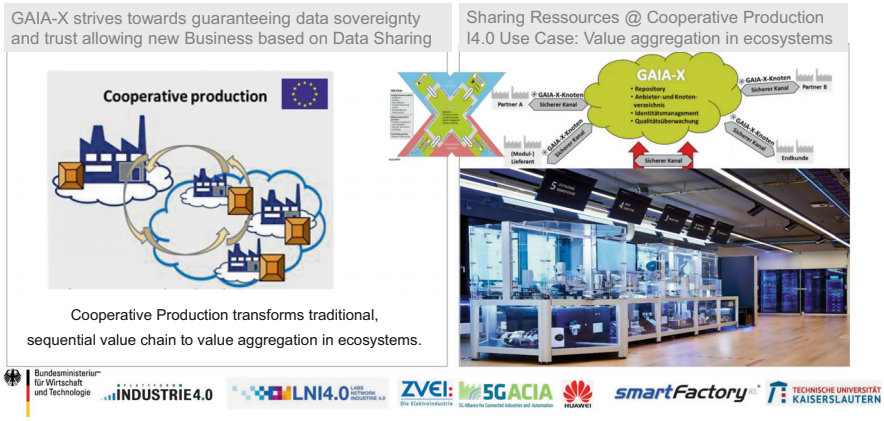


Fig. 27.7 5G connectivity is added to networked production-enabled cloud-based services

important aspect to embrace which goes well beyond sovereign data infrastructures and common data spaces, but is equally important for its adoption: the trustworthiness of AI itself.

In this context, Huawei Cloud is starting to explore the use of GAIA-X-compliant federated artificial intelligence for: control and (predictive) maintenance, match-making of demand and supply actors, as well as auction and contract provision of wind power plants operated by different utility providers. The idea is that the manufacturer of the wind turbines will offer an AI service that allows monitoring, control, and predictive maintenance of the turbines. A potential barrier is that while the power plant operator may be very much interested in using intelligent ready-made value-added services which are tailored to the turbines it bought, it may not be willing to share sensitive data about the actual operation and utilization of its wind farm, e.g., providing sensitive data may comprise future equipment tender negotiations or violate local cybersecurity regulations. To overcome this, the wind turbine manufacturer bundles the AI with the physical machinery. The master AI model(s) is (are) initially trained and hosted in a public cloud environment and operated "as is" at the wind power plant sites. Due to the implicit nature of AI as a learning system, the model continues to be trained on the operational data without that data leaving the wind farm. Instead of sharing the operations data with the manufacturer, the updated AI model trained onsite is fed back to the cloud and synthesized, resulting in an improved master AI model. This can then be redistributed to the customers, i.e., wind power plant operators. Without going into the details, there are variations in the approach to the federation with regard to central or decentral orchestration and homogeneous and heterogeneous data sources to be integrated. The selection will depend on the specific setting and can be implemented in combination, i.e., decentralized within the wind farms and centralized as seen from wind farms of different providers to the cloud backend of the turbine manufacturer. In any case,

privacy by design is ensured with regard to data sharing. However, there are still several research questions to be solved with regard to federated learning such as optimizing the required communication among participating nodes and in particular on the characteristics of the robustness and integrity.[1]

For both use cases, Huawei and partners are also addressing questions related to the federation of data and services as active assets within a networked production ecosystem related to the concept of digital twins and the required extension of the asset administration shell (AAS).[2]

While the past focus of the AAS definition could be placed more on hardware, firmware, and connectivity-related aspects within the shop floor or physical factories and their physical environment, now and in the future, virtual factories, digital twins, and the increased integration of intelligent services require us to put more emphasis on the software dimension and cloudification aspects. Huawei is working with the ZVEI working group "IT in Automation" on a corresponding white paper for the "software type plate: as part of the AAS.

Moreover, considering the cloud as an embedded part of the (eco-)system of production-as-a-service requires new approaches to the integration of enterprise data across the information technology (IT) and operation technology (OT) domains. The cloud will have to integrate all steps of the manufacturing life cycle and offer holistic trustworthy data management. The cloud enriches the picture with business-focused data logic and offers the required platform services for the shop floor. Maintenance or value-added services can make use of AI and analytics cloud features, and product development and product twins will benefit from dedicated high-performance compute clusters. The increasing abstraction at scale at the manufacturing level, e.g., for production design based on production twins can be supported through an integrated cloud-edge IoT platform with rich and powerful data ingress, data lake, data storage, and advanced analytics capabilities.

In the aforementioned projects, Huawei considers digital twins for the manufacturing sector that are truly enabled through the cloud-edge continuum are a crucial component of an open, trustworthy, and sovereign data infrastructure. The various activities in that space driven by the EU and national governments in close interaction with industry and academic institutions put Europe at the forefront of the future of production.

[1] In more general terms, trustworthiness of AI implies new policies and system assessment criteria related to fairness, explainability, accountability, robustness, reliability, and safety of AI. Various expert groups have been looking into the definition of a general assessment list for Trustworthy Artificial Intelligence (ALTAI); see, e.g., https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[2] In the world of I4.0, each asset is given an Asset Administration Shell (AAS). The AAS consists of a number of sub-models in which all the information and functionalities of a given asset are described. It allows for the use of different communication channels and applications and serves as the link between I4.0 objects and the connected, digital, and distributed world; s. "The Asset Administration Shell: Implementing Digital Twins for use in Industrie 4.0, ZVEI publication."

27.8 Conclusions

In the current global framework, German and European firms at large are challenged to embrace digitalization to secure their competitiveness. International Data Spaces (IDS) and GAIX-X provide an IT architecture and a collaboration framework with the aim of safeguarding data sovereignty. These capabilities are extremely relevant for Huawei as its goal is to foster a fertile business environment in Europe and globally, governed by trusted services and a high level of interoperability, fully complying with regional digital sovereignty.

In the first part, we highlighted the current inherent issues and limitations in the data spaces ecosystem. Platform-based ecosystems are a relevant part of the current and future digital economy. However, it has been observed that business ecosystem platform orchestrators tend to maintain control over their complementors by determining who can access the platform and under what terms, and this reflects into offers, pricing models, and access to key data and IPR, thus creating unbalanced data-driven businesses, affecting SMEs in particular.

Moreover, various sectors deal with sensitive data that depends on trust and security, e.g., in production, health, and telecommunications. However, there is a widespread lack of interoperability and high heterogeneity of data, often fragmented.

Regulators have become increasingly aware of the problem, and they are introducing key policies, in particular a legal framework encompassing data protection, fundamental rights, safety, and cybersecurity (European Strategy for Data). Measures have also been taken to facilitate competitiveness while achieving a high level of trust and interoperability, for all stakeholders of all sizes (SMEs to Big Tech). Here is also where the IDS and GAIA-X initiatives will come into play.

The "cloudification" of the main industrial processes with leverage of artificial intelligence (AI)-based solutions and approaches (e.g., basic AI, machine learning, neuronal networks, and deep learning) gradually leads to the dynamic and flexible establishment of multiple interwoven industrial ecosystems; it is paving the way for the creation of various business models and richer offerings which will ultimately benefit efficiency, increase throughput, and reduce CAPEX. They also allow different levels of business relationships among the stakeholders. This creates an unprecedented level of opportunity, growth, and relative complexity, but it also makes it more apparent that aspects like interoperability, legislation, digital sovereignty of the data flows verifiability, and trustworthiness of services need to be considered carefully.

In this context, Huawei has started to look at contributing suitable solution architecture designs and open software building blocks to the open-source community, to support the ability to govern dataspaces and enhance existing governance frameworks with additional recommendations and insights. For Huawei, it is, imperative that the bidirectional data exchange of structured and non-structured data (like product configuration, pricing, design, feedbacks, and deliverables) with other ecosystem partners can occur frequently and efficiently in a trusted, safe, and controllable environment. Implementing the IDS concepts in data flows can lead

to more scalable services with our partners and ultimately allow augmentation of the overall ecosystem.

We then considered a use case in manufacturing. The use case is an evolved implementation of Industry 4.0 in a factory. The integration is carried out across company boundaries, with value-added partners that are constantly being integrated in order to fulfil a particular production order. 5G connectivity is integrated into this example of trusted networked production to further enable the deterministic integration of remote cloud-based services. Retrofitting of legacy machines, e.g., by deploying smart AI-services from the edge is also demonstrated in the use case.

Finally, we mentioned how Huawei Cloud is also starting to explore the use of GAIA-X-compliant federated artificial intelligence for control and (predictive) maintenance as well as match-making of demand and supply actors.

With its engagement, Huawei hopes to contribute to the emerging European data economy in strict compliance with the related regulations.

References

1. International Data Spaces Association. Retrieved from https://internationaldataspaces.org/
2. Jacobides, M. G., et al. (2018). Towards a theory of ecosystems. *Strategic Management Journal, 39*, 2255–2276.
3. Evans, P. C., & Gawer, A. (2016). *The rise of the platform enterprise, the Center for Global Enterprise*. Accessed from https://www.thecge.net/app/uploads/2016/01/PDF-WEB-Platform-Survey_01_12.pdf
4. Reeves, R., et al. (2019). *How business ecosystems rise (and often fall)*. Accessed from https://sloanreview.mit.edu/article/how-business-ecosystems-rise-and-often-fall/
5. Tirole, J. (2015, July 9). *A few remark on the role of intermediaries in the digital economy*. Lecture, Boston.
6. Three paradoxes of Platform – CACM. (2015, February). Accessed from https://dl.acm.org/doi/10.1145/2700343
7. A European strategy for data. (2020). Accessed from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066
8. Digital Services Act. (2020). Accessed from https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN
9. Data Governance Act. (2020). Accessed from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767
10. Proposal for a Digital Markets Act. (2020). Accessed from https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608116887159&uri=COM%3A2020%3A842%3AFIN
11. Synergy Research Group. (2021). *European cloud providers struggle to reverse market share losses*. Accessed from https://www.srgresearch.com/articles/european-cloud-providers-struggle-reverse-market-share-losses
12. GAIA-X Homepage. Accessed from https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html
13. GAIA-X Summit 18th and 19th of November 2020. Accessed from https://events.talque.com/gaia-x-summit/en/6iq6yI5LPSxaIRA6cmnq
14. Shek, H., et al. *KPMG cybersecurity, overview of Draft Personal Information Protection Law in China (2019)*. Accessed from https://assets.kpmg/content/dam/kpmg/cn/pdf/en/2020/11/overview-of-draft-personal-information-protection-law-in-china.pdf

15. Zhang, G. & Yin, K. (2020). *A look at China's draft of personal information protection law*. Accessed from https://iapp.org/news/a/a-look-at-chinas-draft-of-personal-data-protection-law/
16. Leung, H. T. (2020). *The Mainland unveils new draft data privacy law*. Accessed from http://csj.hkics.org.hk/site/2020/12/28/the-mainland-unveils-new-draft-data-privacy-law/
17. Rogier Creemers, R. (2020). *China's draft 'Personal information protection Law' (full translation)*. Accessed from https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinas-draft-personal-information-protection-law-full-translation/
18. Ministry of Foreign Affairs of People's Republic of China. (2020). *Global initiative on data security*. Accessed from https://www.fmprc.gov.cn/mfa_eng/zxxx_662805/t1812951.shtml
19. Data Lake Governance Center (DGC). Accessed from https://www.huaweicloud.com/en-us/product/dlg.html
20. Huawei Fusion Insight Drives the Efficiency of Qinghai Green Energy. Accessed from https://e.huawei.com/de/videolist/cloudandbigdata/35db3d2b2c1f4babb558c95fa09df439

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 28
International Collaboration Between Data Spaces and Carrier Networks



Akira Sakaino

Abstract NTT, headquartered in Japan, is a global ICT company with data centers, network facilities, R&D centers, and business locations all over the world. It is developing new data infrastructures that utilize next-generation information communications technology and digital twin computing. This section introduces NTT's R&D activities, services and solutions, and future initiatives to ensure security, compliance, fairness, transparency, and interoperability in the global data space. The details are as follows:

1. Concerns and issues regarding the cross-border sharing and use of industrial IoT data between multiple companies
2. Use cases considered by RRI (Robot Revolution and Industrial IoT Initiative) and requirements for global data spaces
3. Next-generation optical and wireless network "IOWN" and highly reliable data infrastructure "SDPF with Trust" proposed by NTT to realize data spaces that meet requirements for global data spaces
4. Demonstration system and international interconnection experiments using IDS Connectors to connect data spaces between Japan and Europe
5. Ideal international data platform architecture that connects GAIA-X compliant data spaces to networks of telecommunications carriers in various countries
6. Concept of a digital immune system "Global autonomic nerve" that combines data spaces with IoT and AI to achieve sustainable governance of economic activities and the global ecosystem

A. Sakaino (✉)
NTT Communications Corporation, Tokyo, Japan
e-mail: akira.sakaino@ntt.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_28

28.1 NTT's Business Domains, Mission, and Services

NTT has 300,000 employees and offices in 88 countries and provides regional, mobile, long-distance, international voice and data communication services, as well as smart energy and urban development solutions in 190 countries. In order to provide secure ICT solutions to customers, it has become a new social mission for NTT to develop an international data management platform that guarantees interoperability, trust, reliability, transparency, and sovereignty. The following NTT services could help the mission.

- IP network: quick and safe tier-1 global network
- Virtual private network: enterprise broadband networking
- Data center: scalable and reliable data centers in 20 countries
- Enterprise cloud: private, multi-tenant, and third-party clouds
- SD exchange: connect enterprise cloud and various clouds
- Cloud management: optimize multi/hybrid-cloud environment
- Managed security: global, integrated security services

28.2 Key Requirements for Data Spaces

28.2.1 Why We Need Common, Secure, and Fair Data Spaces

Today, new business models such as the sharing economy and the circular economy have emerged, and the need for data sharing to improve productivity is growing. As the value chain between companies expands internationally, it is also necessary to share the data globally, seamlessly, and promptly in cyberspace for quality control and risk hedging, etc. If we share all the data on the global value chain (design, manufacture, lease, billing, maintenance, recycling, etc.), we can respond quickly to the diverse needs of users while respecting SDGs adopted by the United Nations. However, if data is leaked or misused, companies would suffer a large loss. Data monopoly, censorship, and regulation can impede international data flows. Inconsistent data flow regulations, rules, and standards could also make data sharing difficult. Based on this recognition, I have established "Global Data Management Platform Study WG" within RRI (Robot Revolution and Industrial IoT Initiative) and discussed the requirements for an ideal international data platform with the WG members.

28.2.2 Use Cases for International Industrial IoT Data Utilization

In considering the requirements of the WG, we assumed a use case where a company not only sells products but also controls the quality of products in use by many users and performs predictive maintenance, and manufacturers should share product data with overseas customers and component manufacturers.

28.2.2.1 Use Case 1: Mobility as a Service (MaaS)

Use Case (Assumed Situation)
A car-sharing company globally provides MaaS using vehicles from various manufacturers. When a user depresses a brake pedal of a vehicle to slow down while driving, a vibration sensor, which is attached to brake mechanism for detecting failure of brake system, exceeds a threshold value, and an alert is issued requesting confirmation of safety to a car-sharing company.

Scenario
The car-sharing company asks the vehicle manufacturer through the maintenance company to analyze sensor data. The manufacturer verifies the manufacturing history with the parts manufacturer to determine the cause. The maintenance company inspects and replaces vehicle parts. An insurance company contracted with a car-sharing company pays benefits to the maintenance company (Fig. 28.1).

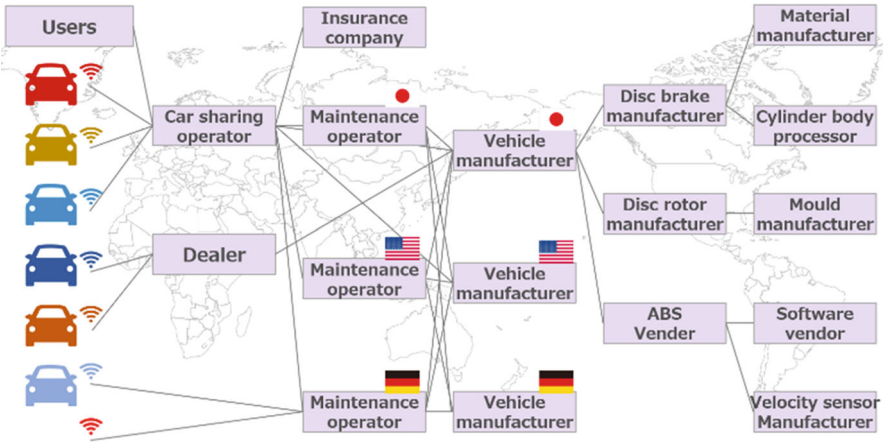


Fig. 28.1 Collaboration between companies for MaaS (©2021, NTT)

28.2.2.2 Use Case 2: Factory as a Service (FaaS)

Use Case (Assumed Situation)
A manufacturer that exports agricultural machinery to foreign countries entrusts the predictive maintenance for its products to a FaaS operator*. The vibration sensor of the machinery exceeded the threshold value, and an alert requiring replacement of worn parts was issued.

*FaaS operator: A company that manufactures, delivers, and replaces genuine parts on demand in cooperation with overseas parts factories and maintenance companies.

Scenario
The machine manufacturer asks a FaaS operator to replace parts. The FaaS operator selects the nearest and best partner factory in terms of quality and delivery time. The machine manufacturer discloses drawings and recipes for the parts to the selected partner factory, which is confidential detailed data that contains the company's proprietary manufacturing know-how and is not normally given outside the company because it should not be disclosed to competitors. The partner factory uses the confidential data and makes the parts, and the maintenance company receives and replaces the parts (Fig. 28.2).

The study of these use cases reveals that MaaS/FaaS operators need data spaces that can securely share sensitive data with manufacturers, partner factories, and maintenance company and control the scope of disclosure and terms of use of the data (Fig. 28.3).

28.2.2.3 Key Requirements for Future Global Data Platforms

Based on the above use cases, the WG considered the concerns and challenges of data sharing for each company, setting up the business scene, business flow, players,

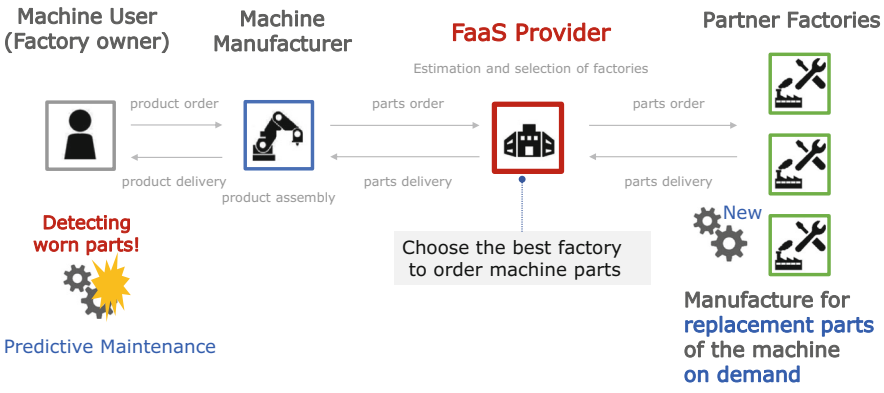


Fig. 28.2 Business model and process for FaaS (©2021, NTT)

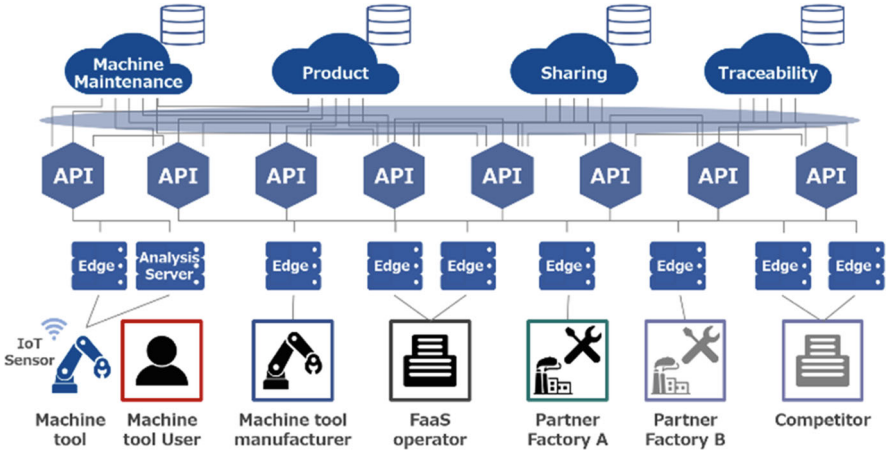


Fig. 28.3 Data spaces architecture required for FaaS (©2021, NTT)

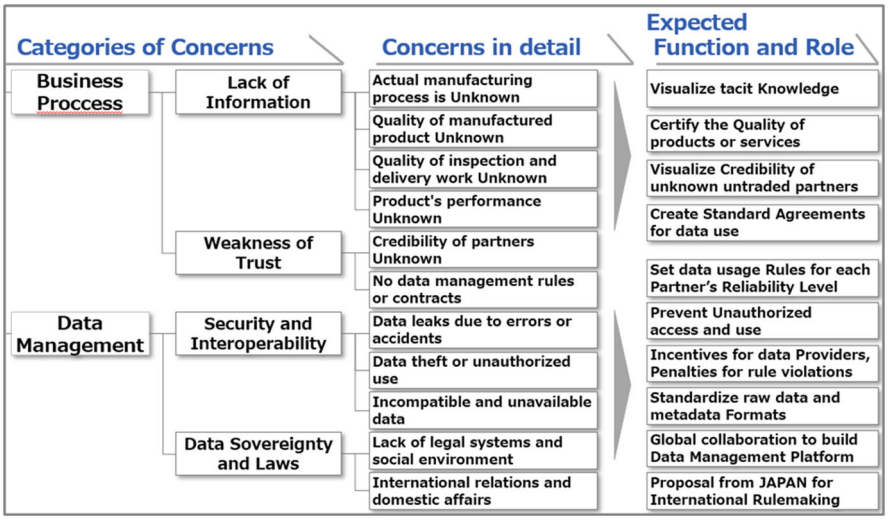


Fig. 28.4 Major concerns and challenges of data sharing (©2021, NTT)

a value network between each player, each player's business role, value provided, and activity (Fig. 28.4).

We listed 134 concerns and identified the challenges to address them. Then we detailed the expected functions and roles of data platform and identified major system requirements that will be important in the future for international data sharing as follows (Fig. 28.5).

In order to create and operate the data spaces that meet these requirements, I hope to combine Fraunhofer's achievements with NTT's technologies and services below.

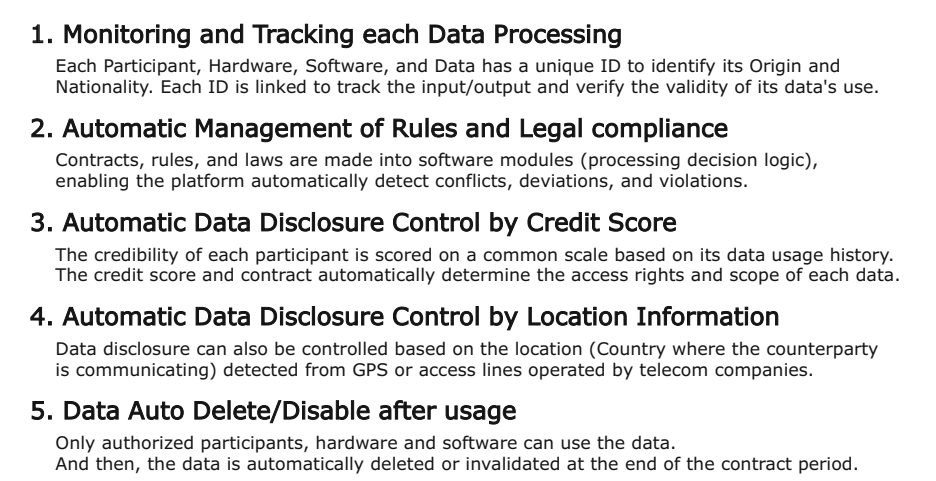


Fig. 28.5 Major system requirements for international data sharing (©2021, NTT)

28.3 NTT R&D Related to Data Spaces

In 2020, NTT has published documents, "Global infrastructure for data sharing between businesses" and "A Global Data Infrastructure for Data Sharing Between Businesses" [1]. To promote value creation using data, NTT is working on building an international and open ICT infrastructure that enables data to be provided, shared, and utilized securely, at low-cost, and quickly among companies in all types of industries around the world while maintaining the data owner's data sovereignty. We aim to collaborate and grow with companies, communities, and governments around the world who are working to overcome similar challenges.

28.3.1 IOWN (Innovative Optical and Wireless Network)

The IOWN [2] is an initiative by NTT for networks and information processing infrastructure including terminals that can provide high-speed, high-capacity communication utilizing innovative technology focused on optics, as well as tremendous computational resources, which create useful data spaces combining networks, edges, and clouds. This is done in order to overcome the limitations of existing infrastructure with innovative technologies, optimize the individual with the whole based on all available information, and create a rich society that is tolerant of diversity by effectively utilizing data.

NTT has established the IOWN Global Forum to engage in discussions with people from various industries. It aims to realize a smarter world where data in

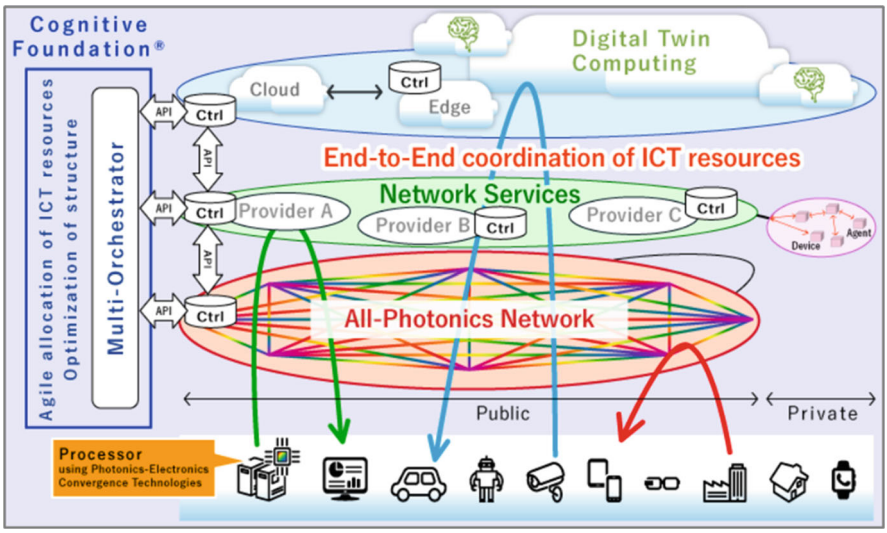


Fig. 28.6 Concept of IOWN by NTT (©2021, NTT)

different industries will be brought together in order to create a fully connected and intelligent society (Fig. 28.6).

NTT plans to develop the Cognitive Foundation Data Hub, which is an infrastructure that consists of multiple servers distributed over a wide area. With its data brokerage functions and shared data storage, it enables nodes, e.g., sensor nodes and AI analysis nodes, to instantly exchange or share large data objects.

28.3.2 4D Digital Platform

4D digital platform [3] integrates sensing data in real time into the Advanced Geospatial Information Database with its highly precise and abundant semantic information and performs a variety of high-speed analysis and future prediction. By combining with various IoT data, it can offer various values such as increasing smoothness of road traffic flow, improving ease of use of urban assets, and enabling cooperative maintenance of social infrastructures. As a cross-industry platform supporting people's lives, and as the one of the key elements of Digital Twin Computing, a part of NTT's IOWN initiative, we intend to leverage NTT R&D and NTT Group technologies and assets toward sequential commercialization beginning in FY2021 (Fig. 28.7).

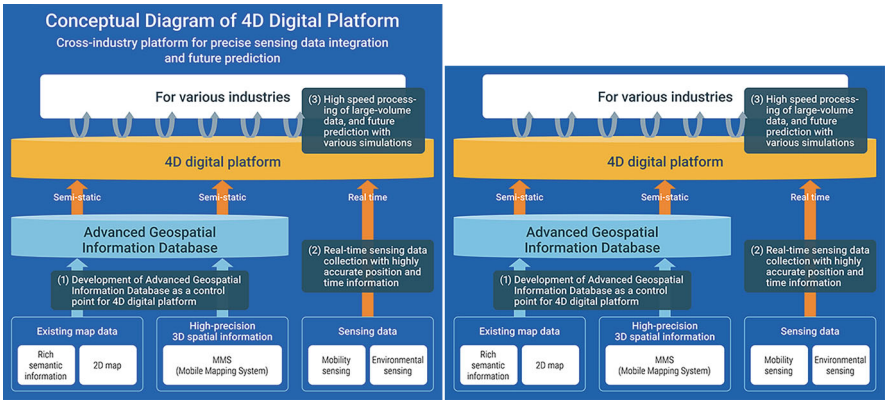


Fig. 28.7 Conceptual diagram of 4D digital platform (©2021, NTT)

28.3.3 Digital Twin Computing (DTC)

NTT proposes digital twin computing [4] as an extension of the conventional concept of digital twins. By freely combining and performing calculations on digital twins of objects and humans in diverse industries, we are able to accurately reproduce combinations that could not be comprehensively handled up to now, such as humans and automobiles in cities, and thereby make predictions about the future. With DTC, NTT will create a diverse virtual society in which objects and people can interact with each other beyond the constraints of the real world and will enable the real world to expand and ascend through integration with virtual society. We aim to create innovative services that have never been possible before, which includes expanding human potential by expanding the scope of human activities to virtual societies or social design and decision support for solving complex social issues through large-scale simulations and predictions (Fig. 28.8).

In the development and operation of DTC, it is necessary to consider issues such as human understanding, privacy, digital ethics, and values of the digital society. However, such problem-solving and social implementation cannot be achieved by players involved with ICT technologies alone. Collaboration among experts across a wide range of academic and specialist fields and players in various industries will be required, such as social sciences, humanities, natural sciences, applied sciences, and interdisciplinary fields. And above all, secure data spaces are essential to achieving our goals.

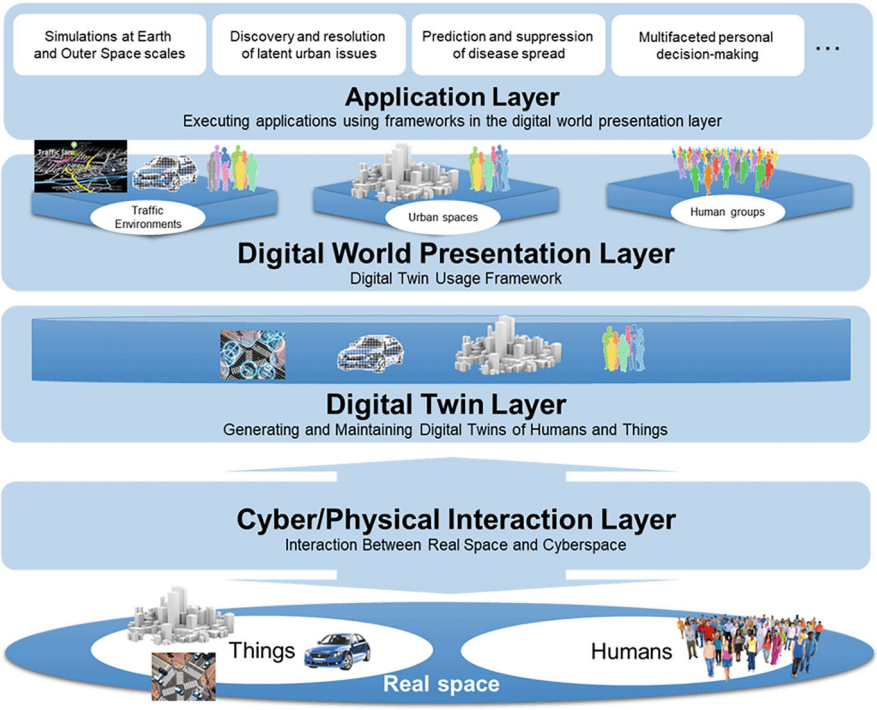


Fig. 28.8 Digital twin computing four-layered architecture (©2021, NTT)

28.3.4 "DATA Trust" and "Smart Data Platform with Trust"

DATA Trust is a concept proposed by NTT for safe and secure data sharing between enterprises and embodied on Smart Data Platform (SDPF) with Trust by NTT Communications. The platform provides features such as linking data owner with each data, access control, data usage tracking, and data isolation for each user, consensual data sharing, partial sharing of data, and anonymizing data. It can separate and control various types of data, such as open data and contract-protected and confidential data.

NTT is applying this technology to provide Digital Utility Cloud [5]. It manages machine tool operation status data, personnel data like work logs collected by mobile devices, and static data such as manuals and specifications safely and securely. Companies that use this platform can take advantage of this data to save costs and resources required for development and improve customer service.

28.4 Initiatives to Interconnect GAIA-X and NTT Platform

28.4.1 Collaboration Between IDS Connector and SDPF with Trust

Since October 2020, NTT has been collaborating with IDSA in a demonstration test as the first phase of contributing to the development of a secure, global data-management platform that assures interoperability between data platforms which will be built and managed in countries worldwide [6]. The test environment for sharing highly confidential data securely will include IDS Connectors, the core technology of GAIA-X, a federated data infrastructure for Europe, and NTT Com's IoT platform and Smart Data Platform with Trust. The demonstration, in addition to assessing the practicality and operability of a new structure for appropriately controlling the access rights of each data based on related laws and contracts, will shed new light on the requirements, etc. of platforms designed for international data management. The results are expected to lead to the establishment of global data management platforms that smoothly link local data platforms in countries across the globe (Fig. 28.9).

In the demonstration, a test environment has been built to test various cases of international data sharing, such as remote monitoring of machines overseas, etc., to verify the practicality and operability of data sharing. IDS Connectors and Smart Data Platform (SDPF) with Trust will be deployed in a test environment in Japan to test system interoperability and the management of specific data-usage rights. NTT has already created a simple demo system using IDS Connectors that simulates the use case 2 discussed in the RRI WG mentioned above.

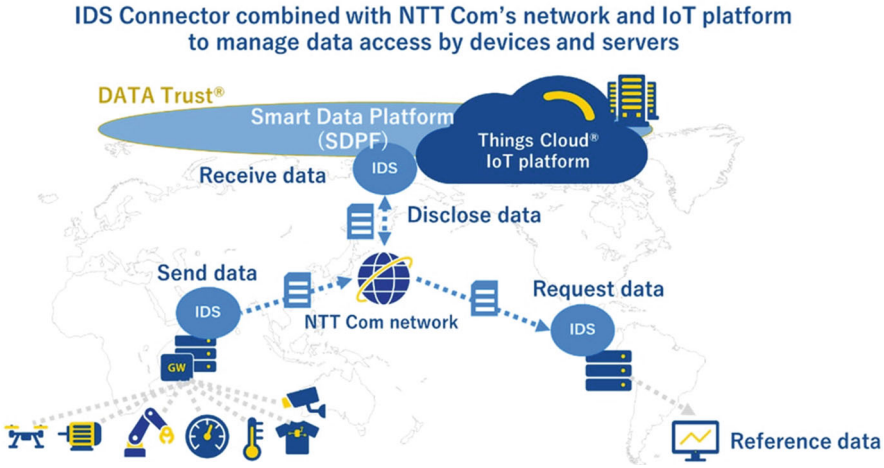


Fig. 28.9 IDS demonstration system on NTT's IoT platform (©2021, NTT)

NTT will proceed with further verification and testing by applying the current test's findings in a test operation environment (test bed), aiming to verify the interoperability of various hardware and software using the IDS Connectors. The new global data management platform will be jointly developed by Japanese and overseas companies and organizations. At the same time, we will determine concrete requirements for the platform together with various organizations and companies active in Japan and overseas, including the RRI. Going forward, NTT hopes to support the formulation of basic specifications through public-private-academic collaboration.

28.4.2 Interconnect Factories in Europe and Japan with IDS

In order to achieve the global target "carbon-neutral society," we must optimize the industrial structure by globally sharing and monitoring data on economic activities that emit CO2. In December 2020, NTT started the data-management Proof of Concept (POC) to visualize CO2 emissions across manufacturing processes with the aim of transforming into a sustainable and law-carbon economy collaborating with SIEMENS at the Switzerland Innovation Park Biel/Bienne [7]. NTT is testing the innovative approach toward global data sharing between businesses using Smart Data Platform with Trust. The POC is performed in the test and experimentation platform of a drone manufacturing line. It highlights the role of data sharing for a sustainable development and circular economy. SIEMENS is supporting the POC by integrating its IoT platform Mindsphere as a central data hub on NTT's virtual private cloud and on-premises servers.

NTT plans to interconnect the test system with GAIA-X prototype using IDS Connectors in 2021. We would like to install IDS Connectors to link with GAIA-X in NTT's data centers and interconnect the Swiss factory and Japanese cloud with GAIA-X prototype system in EU via NTT's SDPF with Trust. We would like to expand this experimental system to build and operate an international joint testbed for interoperability experiments with IDS/GAIA-X in cooperation with companies and organizations around the world and to contribute to the establishment of data spaces that can share various data globally in a fair and secure manner.

28.5 Future Prospects and Expectations for Data Spaces

Today, we face many problems that threaten the environment, human rights, health, and peace. In order to solve these problems and achieve the SDGs, we need to have a mechanism to constantly monitor the activities of the ecosystem of companies, individuals, and organisms and to detect and correct abnormalities and injustices. I hope to create a global cyber-physical system with a mechanism that uses ICT to check the state of the economy and the natural environment and implement defense

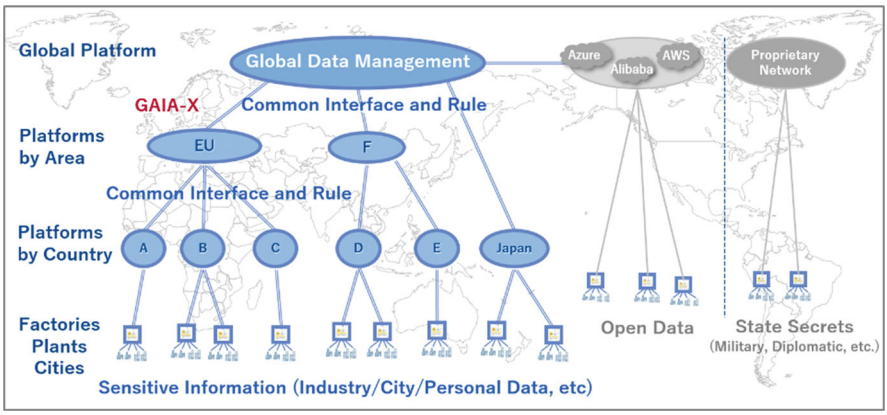


Fig. 28.10 Data spaces hierarchical interconnect architecture (©2021, NTT)

measures and repair them when they deviate from their ideal state, which is just like the immune system and autonomic nervous system of the human body. To complete such a "Global Autonomic Nerve," I would like to work with Fraunhofer and other companies, organizations, and governments around the world to create peaceful data spaces for the SDGs. To this end, it is necessary to build a common, global, open, and fair data management platform that can connect and share data not only within the EU but with all countries and regions around the world. However, rules for using data spaces, such as the definition of critical data, the concept of data sovereignty, and laws governing the export of data, will vary from country to country. Therefore, it might be a good idea to construct global data spaces by establishing a national data space within each country or region that operates in accordance with each law, setting up a data flow management gateway at the each border, and interconnecting them by global common rules and interfaces.

I think GAIA-X should also have secure gateways to interconnect with data spaces in countries outside the EU to enable global data sharing. In addition, to make data spaces more operational and convenient, we need to prepare orchestrator systems that can easily and securely integrate and manage data, devices, edges, servers, clouds, software, gateways, and networks (Fig. 28.10).

As a global carrier, NTT will be able to contribute to designing, building, and operating such an international intercontinental data space architecture with IDSA and Fraunhofer and all other stakeholders, by combining existing NTT services and new technologies including IOWN, 4D digital platform, Digital Twin Computing, and SDPF with Trust. I look forward to promoting these activities together in various fields such as industry, mobility, energy, logistics, health, and cities toward a hopeful future for the ecosystem of mankind and the earth.

References

1. NTT's documents on data spaces. Accessed from https://www.rd.ntt/e/sic/news/research/2020/326.html, https://www.rd.ntt/e/sic/news/research/2020/363.html
2. IOWN. Accessed from https://www.rd.ntt/e/iown/0003.html, https://group.ntt/en/newsrelease/2020/04/16/200416a.html, https://iowngf.org/
3. 4D digital platform. Accessed from https://www.rd.ntt/e/4ddpf/0001.html
4. Digital Twin Computing. Accessed from https://www.rd.ntt/e/dtc/, https://www.rd.ntt/_assets/pdf/iown/reference-model_en.pdf
5. Digital Utility Cloud. Accessed from https://www.ntt.com/en/about-us/press-releases/news/article/2019/0912.html
6. Testbed for IDS and SDPF. Accessed from https://www.ntt.com/en/about-us/press-releases/news/article/2020/0928.html, https://www.ntt.com/en/about-us/press-releases/news/article/2021/1014.html
7. Switzerland Innovation Park Biel/Bienne. Accessed from https://japan-innovation-park.com/en/ntt-poc-2020-12-14/

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 29
From Linear Supply Chains to Open Supply Ecosystems



Fabian Biegel and Nemrude Verzano

Abstract The 2020s will be the beginning of an "Age of Data." Data will help us unleash huge potentials in private but particularly also in the industry. At the same time, partly driven by the availability of this real-time information, the customer-supplier relationships are transforming from linear supply chains into networks and data-driven ecosystems. Efficient standards and technologies are key to this transformation.

However, the biggest hurdle to sharing data continues to be the lack of trust in secure, transparent, and trustworthy information exchange. Concepts and instruments like the International Data Spaces are essential to guarantee data sovereignty.

The SAP Asset Intelligence Network is therefore complemented by the core elements of IDS to maintain trust and transparency throughout the entire industry value network.

29.1 Introduction

Forbes has proclaimed the 2020s the "Age of Data." The EU Commission forecasts a tripling of the value of the European data economy in the next 5 years. McKinsey and PwC assess the market potential enabled by data at around US$13-16 trillion by 2030. Or, to take it beyond financials: Open data could reduce malaria mortality rates by up to 92%, saving up to 400,000 people a year from death. So, we know about the value of data and the exchange and sharing of data in particular.

However, the biggest hurdle to sharing data—in addition to technical problems, capital expenditure, and lack of skills—continues to be the lack of trust in secure, transparent, and trustworthy data ecosystems.

Mechanisms and tools are needed to address this deficiency in trust, in particular for business-to-business scenarios where we share sensitive or even confidential information. We find such a tool in the International Data Spaces.

F. Biegel (✉) · N. Verzano
SAP SE, Walldorf, Germany
e-mail: fabian.biegel@sap.com; nemrude.verzano@sap.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_29

29.2 Data at the Center of Industrie 4.0

The digital transformation of the manufacturing sector took hold early on with the proclamation of the fourth industrial revolution in 2013. Industrie 4.0 even places the availability of data at the center of its definition[1]:

Industrie 4.0 has its foundation in the 'availability of all relevant information in real time through the networking of all entities involved in the value chain and the ability to use the data to optimize the value added at any given time.'

Looking at the business impact of data for Industrie 4.0, the PAiCE study by the German Federal Ministry for Economic Affairs and Energy[2] estimates, for instance, that artificial intelligence has the potential to contribute one-third of the total growth of the manufacturing industry and thus a gross value added of over 30 billion euros within 5 years. In particular, applications for predictive analytics, intelligent assistance systems, robotics, intelligent automation, and intelligent sensor technology offer the greatest opportunity.

Not yet at the center of attention, but already being recognized as potential disruption of the established collaboration models and gaining in importance in recent years, is the breaking up of traditional supply chains in favor of digital platforms[3] and agile networks[4] of business partners in continuously changing roles.[5]

Not least, these will form the core of Industrie 4.0: dynamic ecosystems that make information their central asset not only for optimizing traditional production processes but rather for embarking on new, expanded business models based on data and artificial intelligence.

However, this requires a comprehensive approach to building these industrial data spaces based on efficient tools and standards for manufacturing data ecosystems to thrive.

29.2.1 Data Ecosystems and Data Spaces

Traditional supply chains have proven efficient and robust over decades of industrial production. Disruptions to very linear supply chains however, such as the Eyjafjallajökull eruption in March 2010, low water levels (2019) in the river Rhine, or more recently the COVID-19 pandemic, have shown how fragile they

[1] Umsetzungsstrategie Industrie 4.0, Plattform Industrie 4.0, 2015.

[2] Potenziale der Künstlichen Intelligenz im produzierenden Gewerbe in Deutschland, iit – Institut für Innovation und Technik, 2018.

[3] Wertschöpfung durch digitale B2B-Plattformen, Plattform Industrie 4.0, 2020.

[4] Collaborative data-driven business model, Plattform Industrie 4.0, 2020.

[5] Value Networks as the Foundation for Digital Business Models, Platform Industrie 4.0, giz and CCID, 2020.

also are. German carmakers were still producing at the same volume at the start of the COVID-19 pandemic in spring 2020, while northern Italy—and thus key suppliers to the industry—had already been hit hard and were no longer able to meet their supply commitments. This led to significant interference in the production schedule. Automotive manufacturers experienced similar disruptions in 2021 when major chipmakers were unable to supply according to plan due to extraordinarily high demand from the better-paying consumer electronics industry.

Now, suppliers can't be easily exchanged when it comes to complex parts like in the automotive industry. Digitization and frequent information exchange however can help adopt resource and production planning with very low latency and react at an early stage making the supply chain more resilient to external disruptions. This is especially true if comprehensive information networks with direct information flow between parties at all levels of the supply chain can eliminate gatekeepers that previously impeded the flow of information.

While these data networks enable dynamic data exchange between peers, forming actual data spaces holds even greater opportunities for the partners. The use of Big Data in the industry for optimization, for instance, or AI-based algorithms for dynamic (or even predictive) asset management requires building application-specific, virtual data pools for analytics and machine learning beyond the one-to-one data exchange for transactional purposes.

Even more important for data networks and data spaces than in linear customer-supplier relationships, however, are trust in the partners involved and the appropriate handling of business-critical data. Thus, we heavily rely on the right tools to guarantee data sovereignty.

29.2.2 Standards

Business networks and ecosystems that build their core around information and data depend on finding interoperability concepts, common languages, ontologies, rules, and standards to exchange this information efficiently and use it effectively.

The benefits of such data ecosystems follow the mechanisms of classical network economics: the potential value of an existing network with n established network members increases—to put it simply—exponentially with each new partner.[6] But this also means that size, growth, and efficiency in the exchange of information are the keys to success. Easy onboarding of new members and their eventual offboarding with low hurdles and reliable, established standards are core to this effort.

Technical communication standards and architectures like MQTT messaging or OPC UA have been core to industrial data exchange for a decade. Same holds true

[6] For example, by 2n + 1 applying Metcalfe's law as first described in Metcalfe's Law and Legacy, George Gilder, 1993.

for semantics and ontologies like ECLASS or the NAMUR data model. In recent years, the demand for complex industrial data spaces has been added, which realize a comprehensive view of each asset in the form of a digital twin that covers the entire life cycle up to the recycling aspect in a circular economy. The Asset Administration Shell is one result of these efforts.

29.2.3 Trust and Security

As cited at the beginning, the biggest hurdle to widespread transformation to industrial business networks is trust in network partners and, perhaps even more so, in reliable, secure technology tools that technically underpin this trust.

Especially in data-intensive partner consortia, guaranteeing one's own data sovereignty is inviolable: the ability to not only monitor but actually control at all times where and how "my" data is stored and processed—but even more so, which third parties can access this data and whether it is handled according to the previously agreed terms.

While decentralized concepts such as *Linked Data* and *SOLID* from Tim Berners-Lee and the Massachusetts Institute of Technology are increasingly gaining recognition in the private sector, we see the International Data Spaces at the forefront of these efforts in the area of industrial applications and the "gold standard" for the trustworthy exchange of business information.

However, trust needs to be established throughout the entire technology stack which on the one hand side includes technical security of any data exchange infrastructure and Cloud service involved but also operational aspects and legal framework conditions.

Europe's GAIA-X initiative for a trustworthy federated data and cloud infrastructure is aimed to design and implement an infrastructure ecosystem that meets these requirements. Along with IDS's data sovereignty concepts and toolset, GAIA-X provides a comprehensive reference design and policy framework to underpin these ambitions for innovative yet trusted data ecosystems.

29.3 SAP Asset Intelligence Network

As being described above, the challenges of decentral infrastructures and business applications are not limited to technical challenges, but especially valid to compliance and regulatory aspects. One of the emerging areas is asset management to address the maintenance, operation, and further services that are relevant to an asset or rather device or machine. While real-time transparency is important for specific applications, new business models can also be realized with an efficient asset management. Therefore, the exchange of consistent and interoperable master,



Fig. 29.1 Traditional exchange of master data (©2021, SAP)

transactional, and unstructured IoT data is getting increasingly important to many companies, regardless of the tools from different vendors that are being used.

The traditional way of data exchange is being depicted in Fig. 29.1 and has turned out to be impractical in many ways to enforce collaboration within a company and externally with suppliers or service providers. In many cases, time-consuming processes are in place to on- and offboard partners, service providers, and even employees of the same company because of the heterogeneous applications and tools in the IT landscape of a company. Real-time and sensitive data of assets are difficult to share and need a lot of effort and approvals. Often enough, data is being shared by email.

SAP addressed these challenges since several years already and provides the solution SAP Asset Intelligence Network (AIN), which simplifies cross-company collaboration and consolidates asset-oriented activities from operation over maintenance up to machine learning-driven services like predictive maintenance. Different stakeholders are able to access asset information at a central place to get a transparent overview about the current status and the history of all activities that are stored in different backend systems. Spare part ordering can also be organized, or firmware updates made available, and operators notified. Based on assigned user profiles, the access to data and information can be managed and limited—see Fig. 29.2.

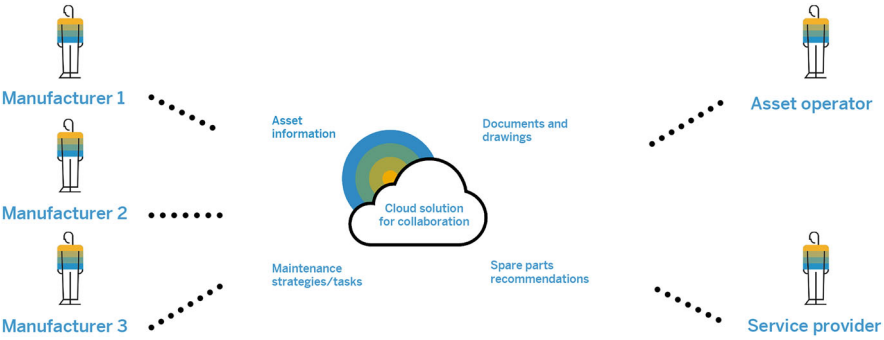


Fig. 29.2 SAP Asset Intelligence Network allows cross-company collaboration and provides transparency of the asset (©2021, SAP)

AIN is being used to store the digital twin of an asset. As soon as the unique asset has been produced and serialized, the digital twin can be onboarded in AIN.

29.4 Data Governance with IDS

The asset management organized in a centralized manner like AIN helps many companies create transparency of an asset and enforce collaboration. From documents like CAD files and manuals, over current IoT data, up to maintenance strategies, AIN enables every stakeholder to access relevant data centrally. But why adding a de-centralized approach like IDS to a well-functioning environment? This article is meant to emphasize that established mechanisms can be extended with new approaches without disrupting the running and existing processes. Usually, existing IT applications are not being turned off to introduce new features or capabilities but transition phases being preferred instead.

As described, AIN already provides data access capabilities to enforce collaboration across companies, but this traditional kind of user management is not sufficient to cope with data being stored at different places on different cloud or edge platforms. Especially IoT data that need to be processed by service providers cannot be governed in a manner that the usage and the processing of the provided data are being controlled.

Current data sharing approaches are limited and cannot sufficiently cover all requirements that have been partly described above to support a scalable enterprise usage. IDS allows data governance if needed and can enrich existing sharing features of AIN to improve the collaboration. Especially a controlled data space, where unstructured IoT data of the assets need to be shared with different stakeholders, will help many companies overcome machine learning relevant challenges to manage data access and to monitor how a trusted consumer is processing the data. For example, lengthy procedures to get IoT data to a machine learning service provider and monitor the exchange between the stakeholders can be dramatically improved by using IDS components. While the compliance and commercial aspects are being managed by the Clearing House, access to the data is being managed by an IDS Connector. How data is being processed depends on the scenario and respective data apps that will be running in a connector. Once the processing of the data is done, the results can be transferred by the IDS Connector as well. The predictive maintenance feature of AIN can be enriched with IDS capabilities to organize and monitor data exchange in a much more compliant manner.

The example above shows how IDS can enhance an existing application and its sharing features with data governance capabilities (Fig. 29.3). There are more potentials to apply IDS to existing and established business scenarios. Pay per use of assets, collective intelligence, and governance of documents are just some to mention that would benefit from the IDS approach.



Fig. 29.3 Enhancement of an existing application with IDS capabilities (©2021, SAP)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 30
Data Spaces: First Applications in Mobility and Industry



Christoph Schlueter Langdon and Karsten Schweichhart

Abstract Traditional approaches to handling of data are becoming outdated. On the one hand, data storage is already costly today, and tomorrow's growth will create a serious cost problem. On the other hand, certain advanced analytics applications require more data than many companies can hope to collet themselves. Therefore, a more federated system with just-in-time data and data sharing seems rational. Unfortunately, many companies worry about data sharing. They are concerned about a loss of data sovereignty, the right to have control over one's data. Once a file is sent, anything could happen to it. This is where the International Data Spaces (IDS) standard comes in to enable a setting that can ensure data sovereignty in theory. To find out in practice, Deutsche Telekom has incorporated IDS technology into solutions and applied them in the automotive business and in industry. We report about first findings from real-world applications.

30.1 Introduction

Much has been said about the value of data, such as "data is the new oil," originally coined in 2006 by Clive Humby, the co-creator of British multinational retailing giant Tesco's Clubcard loyalty program and founder of a successful data analytics agency (see [1]). But few companies beyond Amazon, Facebook, and Google seem to have mastered refining and distilling crude data into value. Their market capitalization overshadows all others (see [2]). If data savviness correlates with market capitalization, then the stock market is proof that somehow being truly data-driven and monetizing data is so much easier said than done. One solution to "democratize" value creation with data is the concept of a data space. Yes, yet another concept born in an academic research lab, in this case in the labs of the Fraunhofer Society, Europe's largest application-oriented research organization, which began creating International Data Spaces (IDS) technology in 2015. On the one hand, we love the

C. Schlueter Langdon (✉) · K. Schweichhart
T-Systems, Frankfurt, Germany
e-mail: christoph.schlueter-langdon@t-systems.com; Karsten.Schweichhart@t-systems.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_30

rigor that comes with scientific methods, such as validating the efficacy of a vaccine (are you sure it will help . . . rather than kill me?). On the other hand, however, the business community has learned that not every idea from science translates into economic value and profit. How fitting, then, that experts and engineers at Deutsche Telekom, a for-profit entity, have put data spaces and IDS (International Data Spaces) technology to the test. "A data space is a shared, trusted space for transactions with data [. . .] based on common standards (or values, technologies, interfaces) that allow and promote transactions with data" [3, p. 109]. Deutsche Telekom has been counting on IDS technology and became a pioneer participant in the International Data Space Association (IDSA) created to promote IDS. Deutsche Telekom is deeply involved at multiple levels, at the Board level as well as in the trenches, such as the launch coalition. Contributions include real-world IDS prototyping in the National Platform Future of Mobility's (NPM) RealLab Hamburg ("Mobility with IDS: Adding the 'N' in NPM to RealLab HH," link), "Rulebook" writing, and formulizing certification procedures. Based on our IDS work and empirical findings, we conclude rather provocatively that data space is a law. And we explain why, using examples from two verticals—mobility and industry—to substantiate and illustrate our findings. Lastly, we will abstract from these examples to share with you the commonalities across use cases and verticals that have been observed with our IDS-based data space applications to present a conclusion as a type of blueprint or recipe for business success with a data space.

30.1.1 Data Is Broken . . . and Data Space Is a Law

Date spaces are shaping up to be a kind of law like Moore's law (the doubling of computer power every 2 years), which is not a law of nature but rather a trend that follows or is enabled by fundamental developments. In the case of Moore's law, it is miniaturization and scaling in manufacturing; with a data space, it is economics of specialization and modularization. The latter is not new at all but has been spreading throughout the physical world (think container shipping, [4]) as well as in software. In software, modularization has spread across the stack from the transport layer (example of the Internet's packetized messaging, the Transmission Control Protocol/Internet Protocol or TCP/IP) to the application layer with web services in the early 2000s (such as Microsoft's .Net; [5]) and Dockerization yesterday ("Dockerize an application," link). Today, attention and innovation have shifted to the data layer. The reflex has been to hoard and pool data on hard drives (it's mine, I want to keep it), in databases and data warehouses since Michael Porter's seminal Harvard Business Review article on IoT, and in data lakes [6]. Today, it is clear that this approach is broken.

- First, it is costly to store data; it is money spent without an immediate benefit.
- Second, additional money is required to keep track of it. Without keeping track, the data may as well not exist in the first place. Tossed on a pile, pumped into a

lake . . . good luck finding anything in this fishing expedition: A structured search space is required for the search to be successful, such as using keywords and tagging.

- Third, these keywords must be carefully chosen and hierarchically sorted for anybody to find anything without searching forever. For experts, a library classification system such as the Dewey Decimal Classification (DDC) comes to mind. For the rest of us, the food analogy is helpful here: We all buy food. We do not slaughter any animals in our bedrooms or grow vegetables in our living rooms; instead, we buy food portioned, packaged, and labeled. We do this going to a "food space" or supermarket, where we consult the store directory, walk to the appropriate aisle, find the right shelf, and compare items on it based on labels. This is what happens pretty much anywhere in the world. With data, there is no supermarket, no directory, no aisles, no shelves, no labels (the Telekom Data Intelligence hub provides a hosted data store with a directory, complete with a self-service process to put data products on the shelf; link). Even basics such as portioning and packaging are a mystery: We know we buy eggs by the dozen, butter by the pound, and milk by the liter. But how do we consistently measure data quantity? By terabyte or length of a time series, or frequency of time series values, etc. (see "Data: Quantity or quality"; link)? And here it is: we dream of monetizing data but have not even agreed on how to measure any orders.
- Fourth, beyond the measurement problem, with data, it seems that most of us do slaughter animals in our bedrooms and grow vegetables in our living rooms. Data tends to be home-made and made to order, just like cars before Henry Ford turned them into a product [7] and economized on production using factory automation [1]. This explains why data scientists spend more than 80% of the time of a data analytics project on data: hunting it down, cleaning it, slicing and dicing it to understand the information contained in it, and finally preparing it for "dinner" (see our research on "Data is broken"; link).
- Fifth, data is going to explode; look no further than the introduction of 5G and IoT, the Internet of Things, which turns every other thing into a website where everything is tracked. Any solution that has been developed to mitigate some of the aforementioned issues is a Band-Aid at best; it will not stop the bleeding. Estimates suggest that the amount of data will increase from 33 zettabytes (1 zettabyte $= 10^{21}$ bytes) in 2018 to 2142 zettabyte by 2035 [8]. If the cost of data storage is already painful today, it will be a nightmare tomorrow. If your data scientists are already struggling to cook dinner for the family, wait for disaster to strike when you open the restaurant.

The natural solution is to modularize data. Applications pull data, and applications have gone from monolithic to service orchestration (see microservices, dockerization), so data could follow suit and become more federated and distributed, too. Just as elsewhere, economics could trigger a shift to just-in-time data and an *n:n* data infrastructure. Why has this not happened yet? A key problem so far has been data sovereignty, or, more precisely, the lack of it. Data sovereignty is the right to have control over one's data. Once a file is sent, anything could happen to it. This

where the IDS standard comes in to enable a setting that can ensure data sovereignty. It's emergence as a data-sharing standard coincides with new regulation. As a key pillar of its data strategy, the European Commission is proposing the European Data Governance Act (DGA) to foster the availability of data for use by increasing trust in data intermediaries and by strengthening data-sharing mechanisms across the European Union.

30.1.2 Data Exchange and Trading for Better and New Business

No doubt about it: Data is becoming a fundamental raw material for the success of national economies. On the one hand, it is in the public interest to enable and even boost the availability, exchange, and use of data. On the other hand, it is in the individual interest of companies to benefit from the added value that greater use of more data promises. But there are some obstacles that are now well-known and widely discussed. There is a strong lack of:

- Interoperability in technology and semantics
- Data quality
- Legal frameworks according to data
- Security
- Digital maturity of many companies and—most importantly—a fundamental mistrust between participants, often competitors in the markets

To overcome these obstacles, new principles, new technologies, and new business roles are introduced and will be further expanded and innovated in the future. Data sovereignty principles and services are at the heart of this. They ensure that control of data, especially sensitive data, is maintained, and are ready to bring more trust into the data business. Additional strong security technology is applicable, and legal frameworks and laws define the trusted environment that every thriving business needs. Last but not least, new roles to enable and maintain data sharing are emerging: data intermediates, data brokers, and data traders, often certified and neutral to economic sectors, because trust in competition means security and neutrality. You could compare them to banks, which care about money; only, they care about data. Both elements are fundamental. No economy will work without them in the future. The following use cases from the areas of mobility and industrial production show how the benefits of data exchange and data sharing already work today, applying existing technologies und generating advantages for stakeholders. In future business history books, they will be known as "early adopters."

30.2 Transition to Data Space-Enabled Mobility

30.2.1 From Auto (Hardware) to Mobility (Service)

The automotive and transport sector is facing a structural reset and will change as profoundly as it did when Henry Ford industrialized the business in the 1910s with the introduction of the moving assembly line, which made individual motorized mobility affordable for the public. Today, radical new vehicle technology such as electrification and autonomous driving is emerging just as digitization is spreading from other industries like publishing and music [9, 10] to automotive with symptoms such as "connected car" services and direct online sales to consumers (DTC). And all of this coincides with climate change and the saturation of urban space with cars to create a perfect storm, which in turn cumulates into a sociocultural and political shift away from car-centric priorities ("Verkehrswende"; see, e.g., [11], link). As a consequence, and in terms of business model change, industry experts foresee a revenue shift from selling hardware to mobility services (see "Metamorphosis of auto into mobility," link).

30.2.2 Traffic Is Broken: New Connected Mobility to the Rescue?

How do we get to our destination faster, in a more environmentally friendly way, but also more cheaply and hopefully more conveniently, especially in congested urban areas and cities? One solution is intermodal travel, linking different modes of transport for a seamless journey from point A to point B: start the first leg by car, switch to rail, and finish the last leg with a micromobility offer, such as an e-scooter or on-demand shuttle bus (see Fig. 30.1 for cascading intermodal scenarios). On the one hand, many major cities are experiencing traffic gridlock, more and longer traffic jams [12], and citizens are complaining about air and noise pollution (we published an analysis on "Stuck in Traffic," link). On the other hand, new technologies are enabling mobility innovation. For one, the widespread use of smartphones and apps (mobile maps, turn-by-turn navigation, payment by smartphone) has enabled new connected mobility service offerings, such as ride-hailing, carpooling, and car-sharing services from new companies like Uber, Lyft, and Free Now. For another, better battery technology has driven innovation in two-wheelers such as e-bikes and e-scooters—in Germany, laws even had to be changed [13]. As a result, smartphone-based mobility platforms with multimodal or Mobility-as-a-Service (MaaS) offerings have emerged in major cities, such as Jelbi from Berlin's main public transport company Berliner Verkehrsbetriebe (BVG) and Switch from Hamburger Hochbahn (HHA), Hamburg's public transport operator. Streets have also become visibly more colorful, dotted with electric scooters or e-scooters—for example, from Bird (black), Lime (white), Tier (green), and Voi (red)—and shared

Simple **intermodal** model with **smart mobility** (see Schlueter Langdon 2020)

	A	to			B
	First leg			Near B	Last leg
	Segment 1			Segment 2	Segement 3

Experiments with **Berlin** digital twin (see Tuescher 2019)

S 0	Personal car			Self parking	Walking
S 2	Same			Smart parking	Same
S 3	Same			Same	Smart scooter

Intermodal model with smart mobility ... plus **public transport**, ÖPNV

	A	to				B
	First leg	Hub to Hub				Last leg
	A -> Near Hub A	Near HA	Hub A transfer	Inter Hub	Hub B transfer	Near Hub B -> B
	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6

Experiments with Berlin digitial twin

S PT1	Personal car	Parking	Walking	ÖPNV	Walking	Walking
S PT2	Same	Smart parking	0	ÖPNV	Same	Same
S PT3	Same	Parking	Walking	ÖPNV	0	Smart shared
S PT4	Same	Smart parking	0	ÖPNV	0	Smart shared

Fig. 30.1 A simple intermodal model and "smart" scenarios

bicycles from providers such as Donkey Republic (orange) and Next (silver). But despite many new transport options, especially for short distances, no intermodal service has yet emerged. Where is the problem? Wouldn't micromobility make public transport, which is tied to rigid routes and time tables, more flexible and thus more attractive? What are we still waiting for? The mobility analytics experts at Deutsche Telekom are trying it out now, both with data analytics in simulation tests in our Berlin digital twin and in everyday life with funding from the Federal Ministry of Transport in the RealLab Hamburg.

30.2.3 Synthesis of Simulation and Real-Life Data Space Prototyping

Deutsche Telekom has a long, rich, and successful history supporting mobility and transportation. Its T-Systems business has been the leading information and communication technology (ICT) provider to the German automotive industry for the past decade (Automotive IT 2020 [14], pp. 18–20, link), from operating connected car back-end systems ("Globally connected vehicles," link) to providing smart charging infrastructure ("Smart electric vehicle charging," link). Deutsche Telekom is already actively participating in the automobility transformation: from contributing to the German government's National Platform Future of Mobility (NPM)

initiative to providing connectivity solutions for micromobility. Furthermore, Tier, a leading micromobility company, which received EUR 250 million in funding in 2020 on top of the more than EUR 130 million it had already raised, was launched at Hubraum in 2018 ("How Tier mobility are set to change transport forever," link), Deutsche Telekom's tech incubator with campuses in Berlin, Krakow, and Tel Aviv ("Our mission," link). Now, mobility analytics experts at the Telekom Data Intelligence Hub ("Smart mobility," link) at Deutsche Telekom IoT are working with T-Systems to combine simulation experiments (theory) with real-life prototyping of a mobility data space to enable new mobility offerings for their business clients (practice).

30.2.4 Simulation and Berlin Digital Twin: Benefits in Theory

End-user benefits are a necessary condition or *conditio* sine qua non for the success of intermodal travel. Without clear end-user benefits, everything else is a waste. But how can we find out? How can benefits be estimated if intermodal travel does not even exist yet? How can we test the impossible to prove the probable? Simulation can provide answers. It is a valued scientific tool (e.g., [15–17]) that has evolved in the field of economics from the foundations laid by Nobel Prize winners Simon [18] and Smith [19]. For mobility, there are already various tools available in Germany, for example, SUMO (Simulation of Urban Mobility), an open-source traffic simulation by the German Aerospace Center [20, 21], and the virtual measurement campaign (VMC) software by the Fraunhofer Institute for Industrial Mathematics ITWM [22]. For our purpose of a rough estimation of end-user benefits, a significantly simplified model and procedure was designed, coupled with a scientifically rigorous implementation and experimental strategy (for a detailed description, see "Simulating intermodal mobility," link; also "Calculator powered by machine learning: Mobility-as-a-Service," link). For our simulation in a digital representation of Berlin, our Berlin digital twin, the results speak a clear language: intermodal travel is faster (for a detailed description, see "Berlin digital twin and intermodal travel," link). Experiments show that time savings can quickly exceed 10% and double with "smart" linking of transport options (see Fig. 30.2 with results for scenarios S_2 and S_3), for example, by recommending not just any parking space but the one with an e-scooter connection nearby to speed up the last leg (see scenario S_3 of smart parking with smart e-scooter in Fig. 30.1). All in all, simulations help develop insights beyond the obvious. Particularly "smart" options reveal very specific data needs, such as parking occupancy prediction data, which turn into requirements for our real-life prototyping effort.

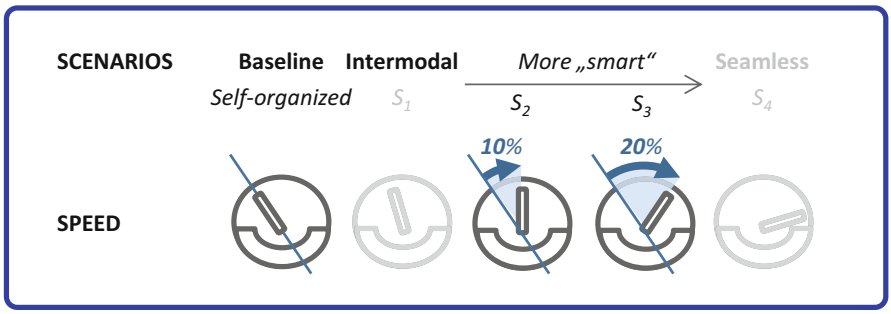


Fig. 30.2 Speed advantage of intermodal travel

30.2.5 Data Spaces: From "Mine, Mine, Mine" to "Win-Win-Win"

What are we waiting for? Okay, consumer benefit is only the first requirement for success. And in this case, the time advantage was the only factor taken into account. It may be the most important, but it is only one decision variable for the customer. Other factors include cost and convenience, or more precisely, the right combination of all these parameters. And because of climate change, environmental protection must also be considered. Coming from the supply side, all companies involved must understand to what extent and in what way consumer benefits can be monetized, i.e., turned into money. This second prerequisite for success requires data sharing. "Smart" only works with data. Take Uber, for example: matching a passenger with a driver and vehicle depends on data, namely, the locations of passenger and driver, availability, and traffic conditions, each with the same timestamp. Without this data, Uber would not be able to orchestrate any transport service. "Uberization" requires that all stakeholders—passengers and drivers—share this data in near real time. The same applies to intermodal mobility: data sharing is key. The problem is obvious: some transport options are in direct competition with each other (public transport, cabs, ride-hailing, e-scooters, etc.). As a result, some service providers are competitors and do not trust each other. Everyone wants to "own" the customer or the customer connection and, therefore, keep the customer data to themselves. In addition, compliance with data protection regulations, namely, GDPR (General Data Protection Regulation of the European Commission, DSGVO in German), can serve as an excuse not to share important data. Many will remember the famous scene in the Pixar blockbuster movie "Finding Nemo," where a flock of seagulls are foolishly chasing food without any coordination and with every seagull screaming: "mine, mine, mine." It is not unlike what is happing with urban mobility in 2020. So far, everything is multimodal, and there is no coordination yet to orchestrate different modes and providers to create a seamless end-to-end-user journey. If intermodal worked, it would be a triple win outcome – "everyone's a winner": consumers (better travel), providers (more business), and cities (cleaner, safer, less noise).

30.2.6 From Simulation to Reality: IDS in NPM and RealLabHH

This dilemma could be remedied by technology that is designed to facilitate data sharing between parties that do not necessarily trust each other, such as competitors. This is where our work on the IDS standard comes into play (IDSA RAM 3.0, link). IDS is a DIN SPEC standard for data exchange while maintaining data sovereignty ([23, 24] and "IDS is officially a standard," link). Or, to put it in simple terms: IDS technology enables parties that do not trust each other to trust a particular data transaction. And that is exactly what our involvement with the real laboratory Hamburg or RealLabHH is all about, determining whether such IDS-based solution is possible. The simulation has helped establish that there are clear benefits for the consumer, for the demand side of the business calculation. What about the other side, the supply side? With RealLabHH, it is now possible to look at the supply side, how IDS can support the orchestration of a service offering involving multiple transport companies. To this end, Telekom is working with the Telekom Data Intelligence Hub (DIH, link) team and the Urban Software Institute (link) to build a fully functional demonstrator that will be presented with RealLabHH at the 2021 ITS World Congress (link). Since IDS is a core technology of the European GAIA-X distributed data infrastructure (link), this will effectively be a first mini-GAIA-X mobility data space demonstrator with real companies involved, such as Hamburger Hochbahn ("Startschuss für das Reallabor Digitale Mobilität Hamburg—mit Bundesminister Scheuer," link). More details on this project in our DIH story "DT and NPM" (link) and "Mobility with IDS" (link). The goal is to show how a federated data structure with sovereignty controls based on IDS can create benefits for all stakeholders (see Fig. 30.3): new, better travel options for citizens, such as intermodal travel, and new business opportunities for both established public transport companies (e.g., convenience offers combining rail and on-demand shuttles) and new micromobility providers (e.g., connecting e-scooters to public transit). Specifically, a planning app for a door-to-door service between Hamburg and Berlin is to be implemented as a demonstrator. With the initial IDS components, such as connectors, a broker, and identity management up and running, the project is already in the midst of fine-tuning and experimenting with data usage policies and enforcement options to facilitate automated machine-to-machine interaction that can ensure data sovereignty for all stakeholders. The timing could not have been better. Because at the same time, the German government is building the first infrastructure components for a German Mobility Data Space with its Acatech Datenraum Mobilität project (DRM; [25]). If these DRM components are installed in time, they could be used for the RealLabHH demonstrator to ensure compatibly with DRM, making it a first, truly operational GAIA-X use case.

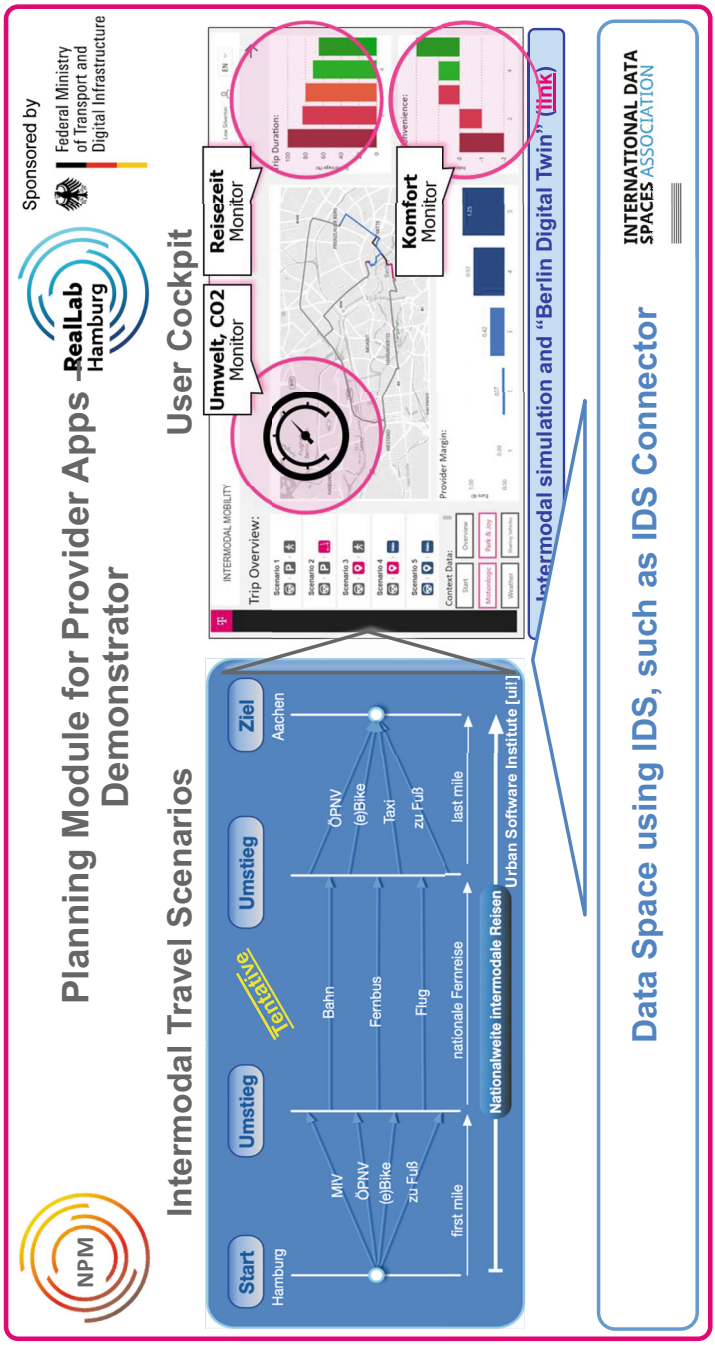


Fig. 30.3 An intermodal travel demonstrator with IDS-based data interoperability

30.2.7 Catena-X Automotive Network and Deutsche Telekom

Many industry observers were caught off guard when German Chancellor Angela Merkel proclaimed data spaces as a top priority for the German automotive business at the "Autogipfel" auto summit in the fall of 2020 [25]. What is a data space? Today, the term is used almost as a matter of course. Only a few months later in the spring of 2021, the German federal minister for economic affairs and energy, Mr. Altmaier, hosted the president of the German Association of the Automotive Industry (VDA), Ms. Müller, together with the CEOs of BMW and Daimler, Messrs. Zipse, and Källenius respectively, for the christening ceremony of Catena-X, a network for secure and cross-company data exchange in the automotive industry. And data spaces were right on center stage. One key enabler of the Catena-X network is the pan-European GAIA-X data infrastructure initiative, which in turn features data spaces built on data spaces technology, specifically the International Data Spaces (IDS) standard, DIN Spec 27070. Catena-X has evolved from first CEO-level talks. In December 2020, Bloomberg News announced that some of the biggest German companies had joined forces to build a German auto alliance [26]. The founders of the partner network include BMW, Deutsche Telekom, Robert Bosch, SAP, Siemens, and ZF Friedrichshafen. In the meantime, additional companies have joined the initiative including Mercedes-Benz AG, BASF SE, Henkel AG & Co. KGaA, Schaeffler AG, German Edge Cloud GmbH & Co. KG, ISTOS GmbH, SupplyOn AG, the German Aerospace Center (DLR), Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V, and ARENA2036 as well as several small- and medium-sized companies (SMEs; see [27]). Now, all participants have pledged to cooperate and collaborate jointly in an open network to advance and accelerate cross-company data exchange throughout the entire automotive value chain. Specific examples include continuously connected data chains to create digital twins of automobiles. Deutsche Telekom is an active participant with its experts on cloud computing, digital twins, GAIA-X and IDS, and, of course, mobility analytics from its Telekom Data Intelligence Hub team.

30.3 Industrial Data Spaces

For the manufacturing industry, there is tremendous value in data spaces; in shared data spaces, the value will become excessive. But to get there, there are some serious obstacles. Data is used and needed throughout the complete industrial production environment, and all processes converge on the shop floor: engineering, production, logistics, order management, service, supply chain, product life cycle, and life cycle of production machines themselves.

Very often, this data is sensitive, because it contains intellectual property (IP), unique selling propositions (USP), competition-critical information on products, production processes, or business models. For this reason, the use of data sharing

or shared data spaces is not yet very well-known in this area of industrial production, mainly for competitive reasons.

On the other hand, there are great benefits in using and sharing more data. With respect to the mentioned sensitivity, data sovereignty can be seen as a critical enabler for more data availability.

This chapter focuses on examples and use cases that are close to the shop floor in terms of the production/manufacturing process and its machines. There is no doubt that there are other valuable use cases in all other mentioned processes.

30.3.1 Most Wanted Use Cases with Own Data: And with Shared Data

In production, there are currently seven most wanted use cases that are built on data and data sharing (Table 30.1). The challenge is that you cannot start at level 7; you have to start at level 1 and work your way up step by step. The good news is that you gain more business value at each step.

1. Dashboards: An information and assistance service for those responsible. The prerequisite is that machine data is available. Visibility through dashboards can be provided on various devices such as screens, computers, cell phones, or smartwatches.
2. Transparency: In addition to the dashboard, not only individual machines or processes are displayed, but all relevant ones in the context of interest. This step provides added value, and it provides an overview and enables dependencies to be identified. Even without another analytic algorithm, transparency is of very high value, e.g., getting an overview of a complete process or production line. The challenge is to make data from different sources available. However, since we are dealing with our own sources, this is quite easy. The added benefit: reaching the maturity of transparency for a defined context is the best starting point to go the next steps.
3. Condition monitoring: In addition to transparency, some limits and values are defined, which are monitored automatically. The degree of transparency and assistance is thus higher.
4. Collaborative condition monitoring: This is the first step that is not possible without data exchange. The basic idea is that limits and values in the operation of machines and productions lines will become much more precise, if as many real experience values of operational data can be used to define them. This can be a data exchange between machine operators, but it is most fruitful to involve machine builders, integrators, or component producers.
5. Anomaly detection: This is the first step where AI comes into play. Translation: First, large amounts of data are needed to train the anomaly algorithms, much more data than a company controls itself. Second, deep data is where high business value can be achieved, because problems can be identified first before

Table 30.1 Shared data needed for high-level use cases

Use case levels	Component builder	Machine builder	Machine user/operator
1. Dashboard	Own data		
2. Transparency			
3. Cond. monitoring			
4. Coll. cond. monitoring	Shared data		
5. Anomaly detection			
6. Predictive maintenance			
7. New business models			

humans are able to notice them. Therefore, good predictive maintenance is a prerequisite for the next step.

6. Predictive maintenance: The idea is simple and well-known; maintain early before something is broken, but not too early so as not to waste time and money. The value of predictive maintenance lies in the precision of the maintenance point. This can be derived first from precise limits and values for step 4, collaborative condition monitoring; second, from high-value anomaly detection; and third, from a lot of operational data exchanged with the company. The rule: better predictive maintenance with better data exchange.
7. New business models, the holy grail: Selling lifelong service by subscription instead of selling a machine once: air instead of pumps, holes instead of drills, mobility instead of cars, etc. The business case of such business models only holds if you can ensure that the service works when it needs to work, for example, 24/7. Translation: You have to manage maintenance before something breaks, i.e., you have to master level 6, predictive maintenance, enriched with excellent spare parts and service processes.

Key to climbing the seven value stages toward new business values is the right use and handling of data.

New core competencies around data are required: access, transport, and storage of data, quality management of data, exchange of data, analysis of data, and, last but not least, the creation of data-driven business models such as ecosystems from it. Security and legal issues are important in all steps and need to be managed carefully. All in all, it boils down to managing data as an asset, just as all other assets are managed in the company.

But the final and most mission-critical competence to open the door to high-value use cases is sharing and exchanging data to enrich one's data base to a valuable extent (Table 30.1).

30.3.2 The Umati Story

One example that enables new data-driven business models in the manufacturing industry and points the way is umati, the universal machine technology interface.

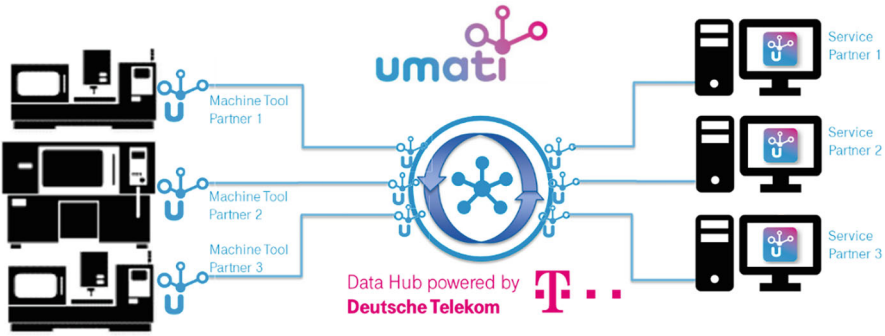


Fig. 30.4 Cross machine data sharing enabled by umati

This is a community of component and machine manufacturers and operators organized at the Verband Deutscher Maschinen- und Anlagenbau (VDMA), the German Engineering Federation. They provide an open platform communications (OPC) companion specification for cross-industry machine builders in all segments, which is to become a new world standard language for machines, a common semantics.

With reference to steps 1–7 of the most wanted use cases mentioned above, this is important from step 2, transparency, onward. For valuable information within a production line, e.g., about energy consumption, the information of all machines in the context must be available in the same semantics. It becomes mission-critical from step 4 onward, when it comes to data exchange. Collaborative condition monitoring requires collaborative semantics.

But it worked very well. The umati demonstrator from VDMA powered by T-Systems runs as a 365 × 24 × 7 testbed for all interested machine builders (see Fig. 30.4). Via self-onboarding, they can connect their machines and forward the data to the dedicated dashboards—worldwide. The demonstrator shows that data transparency and exchange across machine builders are possible.

The implementation of the umati concept in the new business models uses the principles of the demonstrator, expanded by business model-driven roles, governance, security, and services. Selling a machine service instead of the machine, as mentioned above, requires the machine builder to connect as many of its customers' machines to its data-driven predictive maintenance and service platform. With the umati concept and the demonstrator principle, this is possible, worldwide. But does the customer want to do it? And why? There are two drivers for customers. First, own high-value benefits and, second, sufficient trust and control in the process. Point 1 must be served by the entire business model, so that no data provider and partner are forgotten on the benefit side. Point 2 can be ensured through the use of data sovereignty concepts.

30.3.3 Data Sovereignty for More Industrial Data Exchange

The last and most difficult obstacle to collaborative data use and data exchange is trust between partners. Moreover, a concept of usage control is helpful to maintain control over the data: Who does what and when how often with my data—this is the level of ambition. Only those who properly master the obstacle of trust and control can climb the steps to the high-value use cases with the most attractive business cases and benefits.

The data sovereignty concept of IDSA and GAIAX is a strong concept for sustainable data exchange. It takes care of all points of trust and control between the partners exchanging data.

The concept of digital sovereignty opens a world of new business opportunities and even completely new markets. And—moreover—it enables and accelerates existing ones with new business capabilities by adding data sovereignty services. This "add on" is valuable wherever a lack of data or a lack of trust, or both, present themselves as barriers for further growth.

The enablers for business models based on data sovereignty originate from the core services of digital sovereignty technologies and architectures such as IDS or GAIA-X.

What is needed for adding data sovereignty to accelerate data exchange in your business or even to organize and build data spaces to bring together data providers and data customers is the integration and use of the following services:

- For bringing parties together, data intermediary and service intermediary, if necessary, neutral partners are proposed for competitive reasons
- For semantic interoperability: vocabulary publisher and provider
- For governance key services: identity authority, clearinghouse, and certification services
- For support and integration: software developers with specialized data sovereignty know-how.

These roles and their related assets are specifically defined, e.g., in RAM, IDSA's reference architecture. To summarize briefly, beyond the key source of data suppliers, seven key roles are relevant to running data sovereign data spaces (see "New Business Models for Dataspaces," [28]).

Data suppliers produce and/or own data that can be made available in the dedicated data space. Depending on the business model and the operational model in place, the basic roles typically assumed by a data supplier are data creator, data owner, and data provider.

Data intermediary: The data intermediary acts as a trusted data broker and manages data exchange in data spaces. The data intermediary knows the stakeholders, can take care of common roles, routes data, stores data on demand, and manages metadata and the ecosystem of available data sources. An organization acting as a data intermediary can also assume other basic intermediary roles at the same time. Consequently, assuming additional basic roles means assuming

additional tasks and duties for the data intermediary to execute. But one key rule is required: a data intermediary takes care of the data, but as a trusted intermediary, it never uses or analyzes the data itself.

Service provider: Services offer various functions, such as data analysis, data integration, data cleansing, or semantic enrichment of data. The service provider is a platform operator that provides services (i.e., an app, app store, including computing time as a trustee), metadata about services, or both.

Service intermediary: Entity that ensures the up-to-datedness and allocation of services on the data space via the management of metadata (yellow pages for services) about new and existing services. Provider of an interface for the data providers to provide metadata on their available services.

Vocabulary provider: Vocabularies can be used to annotate and describe data assets. A vocabulary intermediary technically manages and provides vocabularies (i.e., ontologies, reference data models, metadata elements). Vocabularies are owned and governed by the respective standardization organizations. A vocabulary intermediary typically assumes the basic roles of a vocabulary publisher and/or a vocabulary provider.

Clearinghouse: Functional instance that ensures the verification of financial/data-based transactions (both in terms of data exchange and monetary transactions) in the data space. Neutral role for the management of transactional metadata.

An identity authority provides a service to create, maintain, manage, monitor, and validate identity information from and for IDS participants. This is imperative for secure operation of the data space and to prevent unauthorized access to data. Each participant in a data space therefore inevitably has an identity (describing the respective participant) and uses an identity (for authentication). Without an identity authority, sovereign data exchange is not possible in the data space. It is an essential component of the data space that can be provided by a company to all participants of the ecosystem.

Certification establishes trust for all participants within the data space by ensuring a standardized level of security for all participants. Certification must be performed in two areas: certification of the operational environment and certification of components. This procedure is performed by two roles, the certification body and an evaluation facility. The evaluation facility performs the certification, while the certification body monitors the process, manages quality assurance, and provides guidance throughout the process. These roles ensure that only IDS-compliant organizations are granted access to the trusted business ecosystem. In this process, the certification body is responsible for setting up the certification scheme, which includes specifications of supported components and profiles, criteria catalogs, test specifications, and certifications processes. In addition, the certification body approves evaluation facilities and monitors their actions and decisions. At the time of publication, the IDSA headquarters acts as the initial certification body.

30.3.4 Make Your Choice

Of course, there are new business opportunities out there to offer some or even all of the above services for dedicated and upcoming industrial data spaces. On the other hand, it might be a good idea to buy these kinds of basic data sovereignty services from a trusted supplier and build a data space or collaborate with data spaces or with partners in data spaces producing the most wanted use cases.

Make your choice of which role you want to play and manage it properly. If you choose more than one role, pay attention to the different management goals.

1. Be a data provider: Look for the data spaces most relevant to you, where you can find the most customers for your data, and the greatest benefit for you
2. Be a service provider: Where can you get the best data you need for your service?
3. Be a data space platform provider: What makes this platform attractive to many data providers and service providers?
4. Be a data or service consumer: What data space platform gives me the best offers?
5. Be a basic service provider: Attract as many platforms as possible using your service as a standard

Make your choice—and start.

30.4 Conclusion

Deutsche Telekom has been actively involved in IDSA right from the start and is heavily involved at many levels from Board membership and "Launch Coalition" to "Rulebook" writing, because IDS technology is trying to do the right thing—data sharing with sovereignty—and looks promising. But Deutsche Telekom is a for-profit entity, and while all of us in business like the rigor of science, we have also learned that the proof is in the practice. That is why Deutsche Telekom also took the risk of using IDS technology. And we did not just start yesterday. Instead, we began early, in 2018, and applied it in more than one vertical to increase the sample size and broaden the empirical base, so to speak. Our pioneering applications of IDS technology in mobility and industry now give us insights into what it takes to turn IDS into a business success. We have abstracted and condensed our cross-industry lessons learned to share it with you.

- First, joining the opportunity is easy. It is also affordable, because it is cloud-based and comes with platform-as-a-service (PaaS) offerings that let you "think big, start small, and fail and scale fast." Find out for free: Test-drive the Telekom Data Intelligence Hub (link).
- Second, the benefits of IDS are real, and the technology is ready for prime time. We have certainly started to put it to use. For example, GAIA-X, the pan-European hyper-cloud initiative of the German and French governments,

relies on IDS as the core technology for its federation services layer (GXFS), and Deutsche Telekom is among the partners helping to implement it.
- Third, the German government is building a first GAIA-X-compatible data space for mobility under the auspices of Acatech, the German Academy of Science and Engineering [25].
- Fourth, both GAIA-X and the German government have chosen to rely on Deutsche Telekom, among others (link). And surely, Deutsche Telekom has proven that it is a trusted, neutral provider of critical infrastructure. It is also Europe's largest telecommunications company, one of the most trusted brands, and #GoodMagenta and #GreenMagenta.

References

1. Schlueter Langdon, C. A., & Sikora, R. (2020). Creating a data factory for data products. In K. R. Lang, J. J. Xu, et al. (Eds.), *Smart business: Technology and data enabled innovative business models and practices*. Springer Nature.
2. Schlueter Langdon, C. (2019). *The auto power shift to data – From asteroid to data sandwiches and exchanges. Working paper (WP_DCL-Drucker-CGU_2019-01)*. Drucker Customer Lab, Drucker School of Management, Claremont Graduate University (link).
3. Federal Government of the Federal Republic of Germany. (2021). *Data strategy of the federal government – An innovation strategy for social progress and sustainable growth*. Cabinet version 2021-01-27, Federal Chancellery, Berlin. Accessed from www.bundesregierung.de/publikationen
4. Levinson, M. (2006). *The box: How the shipping container made the world smaller and the world economy bigger*. Princeton University Press.
5. Schlueter Langdon, C. (2003). The state of web services. *IEEE Computer, 36*(7), 93–95.
6. Porter, M. E. & Heppelmann, J. E. (2015, October). How smart, connected products are transforming companies. *Harvard Business Review*. Accessed from https://hbr.org/2015/10/how-smart-connected-products-are-transforming-companies
7. Crosby, L. & Schlueter Langdon, C. (2019, April). *Data is a product*. American Marketing Association Marketing News, link.
8. Handelsblatt. (2019). *Grenzen des Speichers. Grafik des Tages (2019-05-14): 24-25*.
9. Schlueter Langdon, C., & Shaw, M. J. (1997). A strategic framework for developing electronic commerce. *IEEE Internet Computing, 1*(6), 20–28.
10. Schlueter Langdon, C., & Shaw, M. J. (2002). Emergent patterns of integration in electronic channel systems. *Communications of the ACM, 45*(12), 50–55.
11. Agora Verkehrswende. n.d. *Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), World Economic Forum (WEF). 2020. Transport for under two degrees – The way forward. 10 key insights for the decarbonisation of the transport sector*. Commissioned by the German Federal Foreign Office (link).
12. TomTom Traffic Index Ranking. (2019). *As our world changes, traffic tells the story*. Link.
13. Federal Government of the Federal Republic of Germany. (2019). Cabinet passes regulation for small electric vehicles: Green light for e-scooters. Cabinet version 2021-05-22, Federal Chancellery, Berlin. Accessed from link.
14. Alexander Stroh, C. (2021). *automotiveIT-Ranking: The top 25 IT service providers in Germany*. Link.
15. Schlueter Langdon, C. (2005). Agent-based modeling for simulation of complex business systems: Research design and validation strategies. *International Journal of Intelligent Information Technologies, 1*(3), 1–11.

16. Schlueter Langdon, C. (2014). 3-step analytics success with parsimonious models. In J. Wang (Ed.), *Encyclopedia of business analytics and optimization* (pp. 1–13). Idea Group Publishing.
17. Schlueter Langdon, C. (2020). *Simulators in business: Testing the impossible, discovering the probable. Working paper (WP_DCL-Drucker-CGU_2020-05)*. Drucker Customer Lab, Drucker School of Management, Claremont Graduate University (link).
18. Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA.
19. Smith, V. L. (1962). An experimental study of competitive market behavior. *Journal of Political Economy, 70*, 111–137.
20. DLR. (2020). *Eclipse SUMO – Simulation of urban mobility*. Link.
21. Krajzewicz, D., Erdmann, J., Behrisch, M., & Bieker, L. (2012). Recent development and applications of SUMO – Simulation of urban mobility. *International Journal on Advances in Systems and Measurements, 5*(3&4), 128–138.
22. Fraunhofer ITWM. (2020). *Taffic simulation*. Link.
23. Otto, B., Lis, D., Juerjens, J., Cirullies, J., Opriel, S., Howar, F., Meister, S., Spiekermann, M., Pettenpohl, H., & Möller, F. (2019a). *Data ecosystems: Conceptual foundations, constituents and recommendations for action*. Fraunhofer ISST report.
24. Otto, B., Steinbuß, S., Teuscher, A., Lohmann, S., et al. (2019b). *Reference architecture model version 3.0*. International Data Spaces Association (link).
25. Delhaes, D. (2020). *Merkel drängt Autokonzerne: BMW, Daimler und VW sollen Datenschatz teilen. Handelsblatt (2020-10-28)*, link.
26. Wilkes, W. (2020). *BMW and SAP join forces to build German auto data Alliance - Siemens, Deutsche Telekom also join the cloud data group*. Bloomberg (2020-12-01).
27. Bergmann, C. (2021). *The alliance for secure and cross-company data exchange in the automotive industry is picking up speed*. Deutsche Telekom press release (2021-03-02).
28. Schlueter Langdon, C. (2020). *Stuck in traffic; How bad is it . . . do we age faster . . . how can we fix it*?. Link.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 31
Competition, Security, and Transparency: Data in Connected Vehicles



Karsten Schulze

Abstract State-of-the-art cars are increasingly becoming computers on wheels, constantly collecting, storing, and transmitting data. This goes hand in hand with better market opportunities for certain service providers with access to the data.

There is still no clear legal regulation on who the customer can share their data with—and in what way—nor how transparency and security can be guaranteed for such data in the process. Without legal regulation of data access, there will be no way to ensure a level playing field among providers and freedom of choice for consumers in the future.

Three basic principles apply to the access to vehicle data:

- Third-party service providers should be able to develop new services without depending on car manufacturers.
- Independent service providers, such as independent workshops, insurers, and automobile clubs, should be able to reach customers through the same channels as the vehicle manufacturers.
- Vehicle manufacturers should not be allowed to monitor vehicle users or the service providers selected by vehicle owners.

31.1 Data in Connected Vehicles

Digital transformation is having a far-reaching impact on the automotive sector. Technological development in connected driving and new online business models will thoroughly reshape mobility. Sensors, digital data processing, and data communication enable new functionalities which make driving safer, more comfortable, and more efficient (see Fig. 31.1).

This generates ever-increasing data volumes, which are gaining more and more economic importance. After all, the evaluation of the data will allow service

K. Schulze (✉)
Allgemeiner Deutscher Automobil-Club e.V. (ADAC), Munich, Germany
e-mail: Karsten.schulze@adac.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_31

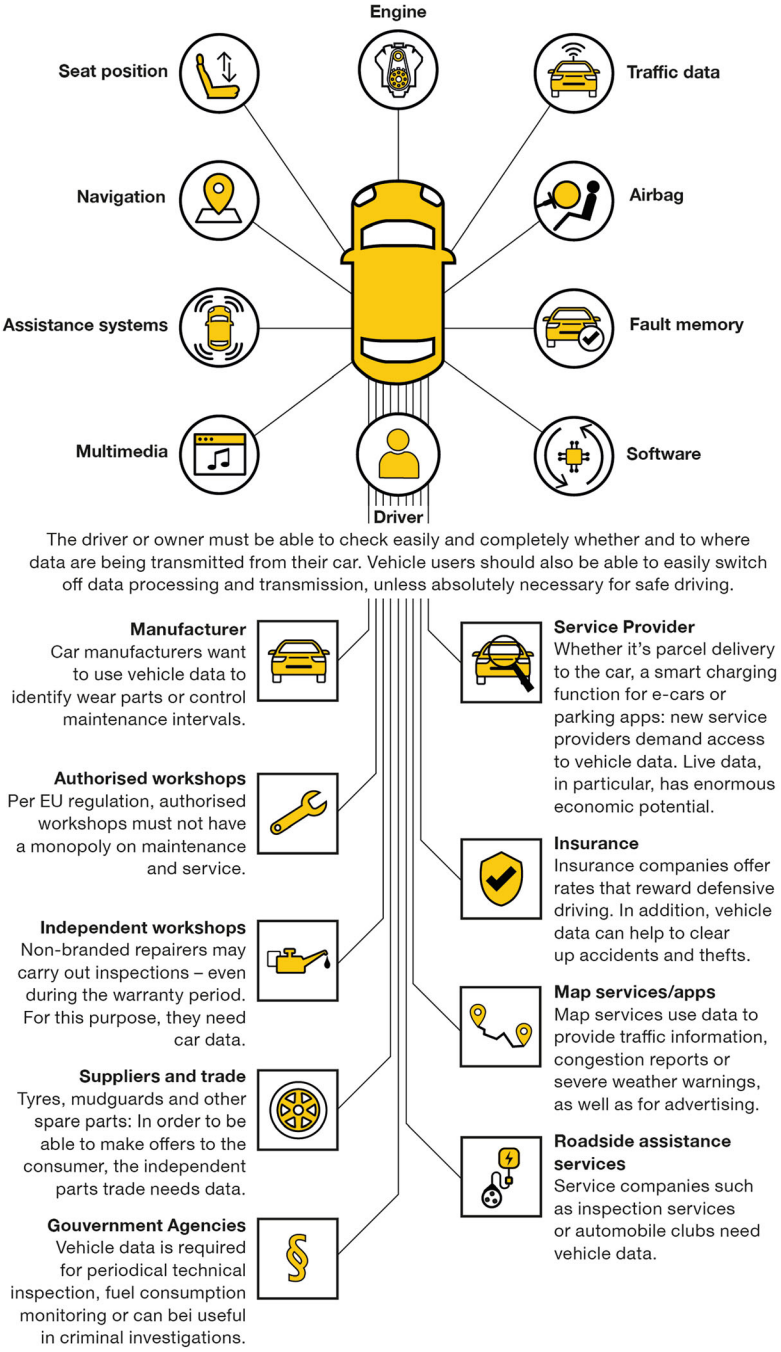


Fig. 31.1 Vehicle-generated datasets and potential uses

providers to optimize their car-related offers in the future. This goes hand in hand with better market opportunities for certain service providers with access to the data.

Data from connected vehicles are specially protected by law as personal data [1]. It is neither relevant whether the data comprise technical information nor whether the data are vehicle-generated or provided by the customer—in all of these cases, the data contain information relating to an identified or identifiable person as specified in the General Data Protection Regulation. However, there is still no clear legal regulation on who the customer can share their data with—and in what way—nor how transparency and security can be guaranteed for such data. Without such regulation, those who have initial access to the data will prevail in the market. In our case, the vehicle manufacturers have that privilege. As a result, they have a gatekeeper position in the market and can control the competition. In the medium to long term, this will have a negative impact on the market as far as the diversity of offers and suppliers and consequently also prices and freedom of choice for consumers are concerned. I take the view that the proposals of the automotive industry for regulating data access are not suitable to eliminate this gatekeeper position.

31.2 Plans of the Automotive Industry Thwart Competition

According to a Europe-wide survey conducted by the FIA (Fédération Internationale de l'Automobile), 78 percent of vehicle users want to choose their own service providers, e.g., workshop, insurance company, or roadside assistance provider [2]. To ensure this, vehicle manufacturers suggest that access to vehicle data be granted via the Extended Vehicle (ExVe)/Neutral Server ("NEVADA Share & Secure"). According to this concept, the vehicle data are made available via the manufacturers' own servers, and access by third parties is monitored and subject to a fee. In this way, the manufacturer takes on the function of both a rights administrator (who gets access?) and a resource provider (which data and functions are made available on what terms?). Currently, the vehicle manufacturer decides on data access solely on the basis of business relations between the companies concerned, which gives the manufacturer a considerable competitive and negotiating edge. For example, consumers who want to share their data with a third party via Nevada will hardly be able to do so if the third party does not conclude B2B contracts with the manufacturer. This concept enables manufacturers to find out which third-party providers access vehicle data, as well as when and how often. As a result, the manufacturers find out a lot about the business activities of other market players. In addition, direct access to the customer in the vehicle via the on-board system is reserved for manufacturers; the ExVe/Nevada concept does not provide such integration for independent market participants. In this concept, therefore, the consumers' right to transmit their data to service providers freely selected by them without prior filtering of offers is hardly enforceable.

Against this background, a broad spectrum of third parties involved in the market, including, e.g., car repairers, technical service providers, as well as ADAC, reject the

concept, because access to data must not depend on the goodwill of a manufacturer toward its competitors.

Therefore, I advocate a swift adaptation of the European legal framework to take account of the digital transformation in the passenger car sector: without legal regulation of data access, there will be no way to ensure a level playing field among providers and freedom of choice for consumers in the future.

31.3 Quick Regulation of Data Access Needed

In 2016, according to industry figures, sales of parts in the automotive aftermarket amounted to €20 billion in Germany alone. Labor costs were another almost €11 billion [3]. Access to vehicle data and users is thus a key factor for the competitiveness of many companies operating in this market. As connectivity increases, data access bears more and more economic relevance, and a swift political decision on the regulatory framework is therefore urgently needed. A FIA-commissioned economic study shows without pro-competitive regulation, independent service providers across Europe might face a loss in revenue of €15 billion by 2025. By 2030, this sum could rise to €33 billion due to further increasing degrees of connectivity [4]. In order to benefit from the full potential of connected vehicles and digital transformation in general, access to vehicle-generated data is essential. This is the only way for third parties to offer car-related services, thus ensuring the consumers' freedom of choice and competition in the long term. Political decision-makers must therefore promptly create a level playing field allowing all stakeholders to act on an equal footing. In fact, the following disadvantages are conceivable, to the detriment of competitors and consumers:

Unnecessary costs	Restrictions	Delays	Surveillance
Vehicle manufacturers charge independent third-party providers for access to vehicle data In doing so, the manufacturers make money from data which the consumer has generated and whose added value should therefore also benefit the consumer—e.g., in the form of attractive and inexpensive offers	Independent service providers could be denied access to relevant data This would hamper or altogether preclude the development of innovative services	Data would be available to third-party providers only with some delay This would make the services they offer less attractive compared to those offered directly by the manufacturers or by companies preferred by them	The server-based data access concepts proposed by the manufacturers allow them to obtain a comprehensive market overview They provide manufacturers, who are also market participants, with a comprehensive overview of their competitors' business models and relationships. This is detrimental to competition

Without regulating access to vehicle data to give all market players equal opportunities and consumers authority over their data, there is a risk of market imbalances in very specific cases. Below are some examples.

Maintenance and repair: Many consumers today exercise their option to have inspections and repairs carried out by independent workshops which are more affordable than service providers authorized by the manufacturers. In the future, manufacturers could make offers to the user directly in the vehicle – more price-sensitive customers would receive a discount, and appointments would be synchronized directly with their smartphone diary. Granted: that sounds convenient. Who would then still bother inquiring about other offers? However, if not every supplier has the same opportunities to approach the customer, competition will be eliminated in the long term. This will lead to rising prices.

Insurance: More and more consumers are opting for telematics tariffs to obtain discounts on their car insurance. For this purpose, the insurer assesses the driver's style on the basis of various parameters and calculates the insurance premium accordingly. However, the decision on who receives the data for such insurance offers should not be up to the manufacturer, but only to the users who should have control over their data.

Roadside assistance: While many drivers nowadays opt to use the services of a roadside assistance organization in the event of a breakdown, in the future, manufacturers could directly employ their own network when a breakdown occurs—alerted via the on-board system—and take the car to one of the manufacturer's authorized workshops. This would give car manufacturers a significant competitive advantage in the entire aftermarket, which would indirectly lead to rising prices for the consumer. In this case too, I insist that car drivers must retain full freedom of choice of their service providers.

E-mobility: The vehicle-to-grid function allows electricity from cars to be fed into the grid, thus helping stabilize the power networks. In ADAC's view, the question of how to handle car data is a decisive factor for successful implementation—ultimately, suitable indicative data could help distribution network operators take the potential energy demand into account in their operational planning [5]. Access to vehicle data is relevant for planning when which car can charge and when excess energy needs to be fed into the grid. In this case, too, a regulation for fair and secure data access is essential.

31.4 Ensuring Competition: Clear Rules for Access to Data

I believe that access to vehicle data must comply with the following basic principles:

- Third-party service providers should be able to develop new services without depending on car manufacturers.

- Independent service providers, e.g., independent workshops, insurers, and automobile clubs, should be able to reach customers through the same channels as the vehicle manufacturers.
- Manufacturers should not be able to monitor the vehicle owner/driver or the service providers selected by the vehicle owner.

In the medium term, ADAC considers the open telematics platform (OTP) to be the best solution for fair competition in motor vehicle maintenance and repair services. Only open, standardized, non-discriminatory, and secure access to the data in the vehicle will enable other players to compete with a manufacturer's products and services and to develop new services. The European eCall Type Approval Regulation (EU) 2015/758 already stipulates that in-vehicle eCall systems should be based on an interoperable, standardized, secure, and open-access platform. ADAC supports such a platform granting open access to all market participants, car manufacturers, independent repairers, insurers, automobile clubs, and other authorized third parties, thus guaranteeing customers' freedom of choice and fair competitive conditions. One major challenge concerning the OTP is to ensure IT security for the access over the vehicle's entire lifetime. ADAC is in favor of a central automotive gateway in the vehicle, which controls communication from/to the vehicle and only allows access to third parties authorized by the vehicle owner. The concept was developed by TÜV-IT and represents a viable IT security concept that enables secure and non-discriminatory access to the OTP [6].

In order to avoid competitive disadvantages on the part of independent market participants in the short term, ADAC is advocating the temporary use of a so-called shared server. This concept is technologically comparable to the Extended Vehicle, but the shared server is operated and controlled by a neutral administrator as an independent third party ("data trustee"). It must be ensured, however, that customer and business data from independent third-party providers are neither accessible to nor usable by a market participant. This is the only way to prevent car manufacturers from acquiring a dominant market position, as would be the case under the Extended Vehicle (ExVe)/Neutral Server concept.

In order to create the conditions for technical regulation in the form of an OTP or, on an interim basis, a shared server, data access needs to be codified rapidly at the European level within the type-approval context.

31.5 Ensuring Data Security Even with Competitive Data Access

System deficiencies making vehicles susceptible to, e.g., keyless go thefts or odometer tampering have shown the need for car manufacturers to enhance IT security. A connected vehicle requires a reliable security architecture for optimal protection against safety-relevant interventions by criminal hackers. From my point of view,

data processing in the car must be protected against manipulation and illegal access over the entire life of the vehicle; this should be checked during type-approval.

However, data security is not a viable argument for allowing vehicle manufacturers first access to vehicle data on manufacturer-owned servers. In my view, the question of the security of vehicle data should not be tied to their storage location, but rather to a security architecture that takes into account the risks of connected cars and provides the best possible protection for consumers. Such protection should be based on standards long since common practice in other industries such as the IT sector. This protection standard should be confirmed by an impartial body, e.g., the Federal Office for Information Security (BSI), on the basis of internationally certified processes such as Common Criteria (ISO/IEC 15408).

Against this background, data access must be ensured via an open and standardized platform, since no one has insight into the manufacturers' systems. Communication with the IT systems in the vehicle must also be encrypted via a secure communication interface. Throughout the vehicle's life, car manufacturers must provide safety-relevant software updates and, where required from a technological point of view, implement required hardware adjustments. Owners must be able to rely on safely using their vehicles in the long term.

31.6 Data Transparency and Data Ownership Create Confidence in New Technology

There is also a need for action on the issue of data transparency. Currently, only vehicle manufacturers know in detail what data are generated, processed, stored, and transmitted in cars. I therefore urge that consumers be informed in detail and transparently about the data exchange between their cars and the manufacturers. The vehicle user should be informed and asked for consent when transmitting/receiving data—as is also common practice in the IT sector, e.g., for software updates or fault code transmissions—unless it is a matter of safety-relevant updates. The data privacy agreements which car manufacturers regularly use to obtain new car buyers' consent to unlimited data access will not be sufficient.

ADAC recommends making it mandatory to publish a list specifying all vehicle data collected, processed, and transmitted for each model ("car data list"). An impartial body should be authorized to check this list to ensure that data protection rules are respected. If the manufacturers refuse a voluntary commitment, the legislator should create a legal basis. Except for certain data whose use is required by law (e.g., eCall), the vehicle user should have the option to easily deactivate the processing and transmission of data that is not absolutely necessary for safe vehicle operation.

31.7 Need for Political Action

- Swift creation of a legal framework for access to vehicle-generated data at EU level
- Definition of clear technical specifications for an open telematics platform to safeguard competition in the automotive aftermarket
- Binding regulations obliging manufacturers to guarantee IT security over the entire lifetime of the vehicle
- Creation of a consumer-friendly regulatory environment for data in terms of transparency (car data lists) and the ability to easily switch off data processing and transmission

References

1. Osborne Clark. (2017). *What EU legislation says about car data. Legal Memorandum on Connected Vehicles and Data*.
2. *FIA Region I: What Europeans think about connected cars* (2016).
3. Gesamtverband Autoteile-Handel e.V. (German Association of Car Part Distributors). *Martkvolumina*. Accessed from https://www.gva.de/branche/marktvolumina.php (German only).
4. Schöneberger advisory services: The automotive digital transformation and the economic impacts of existing data access models (2019).
5. Bundesverband der Energie- und Wasserwirtschaft e.V. (2018). *(German Association of Energy and Water Industries): Digitalisierung, Normung und Standardisierung in der Elektromobilität. Erfolgsfaktoren für die Energiebranche* (German only).
6. Verband der TÜV e.V. (2019). *(German Association of Technical Inspection Agencies): Verkehrssicherheit und Umweltschutz durch Fernzugriff auf Fahrzeugdaten* (German only).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 32
Data Space Functionality



Douwe Lycklama

Abstract Data spaces, simply put, are a functionality for users. They allow users to control their data while sharing it with others. This functionality is what makes data spaces so powerful. They allow data to be shared multilaterally. Those who join a data space can reach many others who have also signed up. At the same time, however, data spaces participants must also consider some aspects. On the one hand, companies have to make business considerations, as participation should, for example, lead to higher revenues, lower costs or better services. On the other hand, decisions must be made regarding measures and technologies for how data sovereignty is implemented in the data space, as this is a central functionality for data sharing. This article provides an overview of possible ways to implement data spaces, the business considerations and a practical approach using the iSHARE trust system.

32.1 Introduction

By reading previous chapters, you may have reached the conclusion that data spaces[1] can mean many things to many people. It is a vast idea, with many angles and incentives for various stakeholders, varying from business leaders, operational manager, policymakers, ethicist, citizens, and businesses to government, IT specialist, researchers, and scientists. This chapter will bring some practical approaches to data sharing.

In its most concrete appearance, a data space is functional to users. Users can control their data while sharing it with others. Control means that they can share the data for a certain purpose (e.g., "only for maintenance purpose"), time ("1 month"), and other conditions such as the nature of the receiver ("lawyer," "doctor,"

[1] More on Data Spaces in page 7–8 from https://design-principles-for-data-spaces.org

D. Lycklama (✉)
INNOPAY, Amsterdam, The Netherlands
e-mail: douwe@innopay.com

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_32

"government"). Data sovereignty is the term often used for this functionality of control, all in an easy-to-use fashion.

Functionality is what makes data spaces so powerful. The word "space" refers to the optionality for users, who want to engage in data sharing. Participants to the space have a lower threshold in engaging with each other (than people outside of the data space), leading to lower transaction cost. Sharing is not bilateral anymore but becomes multilateral. Sign up once, and reach many other who also did sign up. Very similar to GSM, email, and payments cards who all share a decentralized design, leading to vibrant ecosystems with low thresholds for participation.

The larger the number of participants, the larger the data space. There can be unlimited amount of data spaces. In fact, today, many data spaces exist already. Think of the many data-sharing activities happening in, e.g., pharming, banking, insurance, healthcare, energy, or social media for that matter. These can be seen as data space, as they have defined groups of participants who all have sharing and interaction optionality vis-à-vis each other. Most of them are also open to participants, as long as participants fit the requirement set up by the community.

This chapter is about possible ways to implement data spaces, the business considerations, and a practical approach through the trust scheme iSHARE.

32.2 Business Considerations for Data Sharing

Consider yourself as a decision-maker or influential in business, e.g., agriculture, finance, manufacturing, energy, or health, or government for that matter, because governments also engage in digital data sharing.

In today's world of digital business, data sharing is often performed by commercial intermediaries, often referred to as platforms. They put their own capital at risk by offering a data exchange service to a defined set of actors while investing heavily in growth, because the more users they have, the more value they create for their customers and themselves. They invest in breaking the chicken and egg problem, by enlarging the optionality for all users. Examples of such platforms are freight and transport providers, b2b e-commerce platforms for supplies, and agriculture data for buying and selling crop, or social media platforms in the personal data spaces.

Data sharing via data spaces can be seen as 2.0 data sharing. The biggest difference with the platform business model is that the governance of the data spaces is not geared around profit maximization of a few owners, but the whole community of users can benefit. Multiple actors participating in the data spaces make profits. This is not such a new concept as in today's world, already many of such cooperative structures exist to realize public goals. Current examples include data exchange in sea and airports, agriculture, energy, and healthcare, providing the ecosystem environment for actors to do their business.

Governance and profit distribution are the largest differentiators of data spaces, vis-à-vis commercial platforms. Data spaces allow actors to realize their own value add, based on the data space, while offering interoperability for the end customers.

As such, data spaces should be seen as infrastructure, similar to electricity grids, roads, and telecommunications.

Having a clear business goal is paramount for any form of data sharing. It should lead to either more revenue, lower costs, or better services. In most situations, it is a combination of these factors as they are not fully independent from each other. However, the time horizon is also crucial.

We can apply the three horizons model here (by Baghai, Coley et al.[2]). The short-term horizon will lead to continuation of data sharing, as we know it, including bilateral arrangements, data silo's and platform providers. Horizon 2 is where we are with data spaces. Conceptually and technically, all is there, after many years of R & D. The biggest challenge now is the adoption and implementation. At its core, this is the change process within organizations. It will require a longer time horizon before the business results can be monetized. Horizon 3 is not applicable anymore when it comes to the conceptual development of decentralized data sharing. All is there to make it happen.

Leadership with a strategic mindset is needed to overcome this longer time span of horizon 2. The impact and competitive advantage of engaging in data spaces transcend the shorter horizon 1 of operational managers, charged with caring about the day-to-day problems. Amara's law is applicable here as well: we tend to overestimate the effect of a technology in the short run and underestimate the effect in the end.

32.3 Data Sovereignty Spectrum: From the Technical and Legal Perspective

How to start in practice with data sharing? As mentioned, a business reason is paramount for engaging in data sharing. Business reasons vary from achieving higher revenues, lower costs, and higher quality of goods or services. In practice, this means combinations of these three elements. Figure 32.1 shows an overview of various data-sharing situations.

Whatever the reason for data sharing, data sovereignty is the central functionality and design principle.

Control on data can be viewed from both a technical and a legal perspective. Technically, control is given through identity access management techniques, which are built into the APIs, which are actually sharing the data. Data right holders can authorize third parties (data consumers) to access the data source controlled by the data rights holders, by giving and recording the proper consent. Figure 32.2 shows the role-play between data consumer and data provider, from both a technical and legal perspective.

[2] White, D., Baghai, M. & Coley, S. (2000). *The alchemy of growth*. Perseus Books.

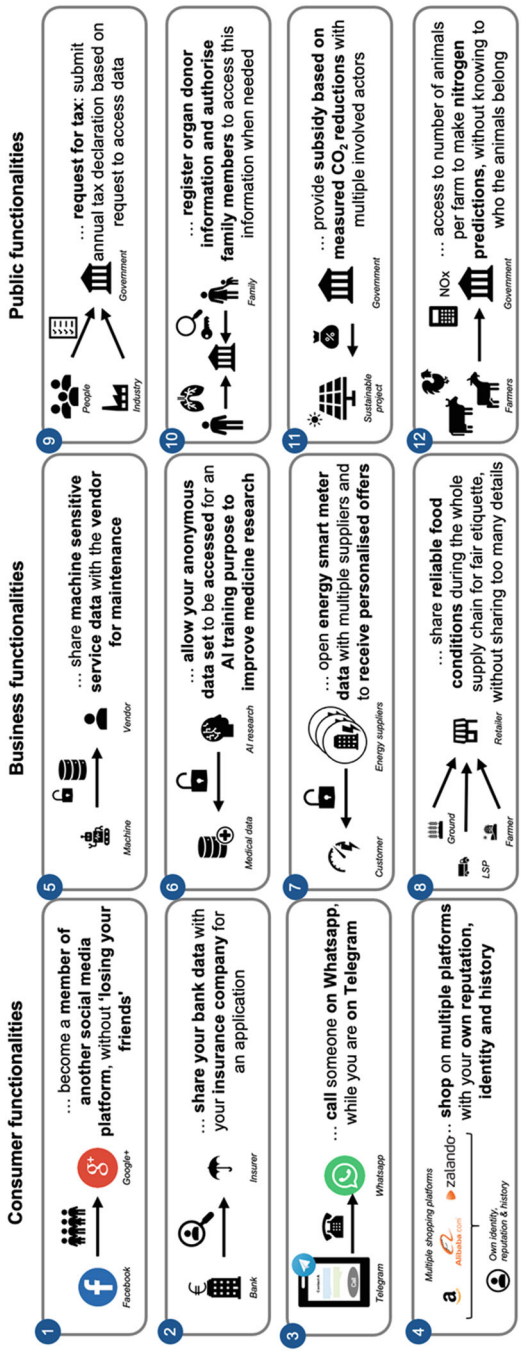


Fig. 32.1 Data sovereignty in functionalities (Source: INNOPAY)

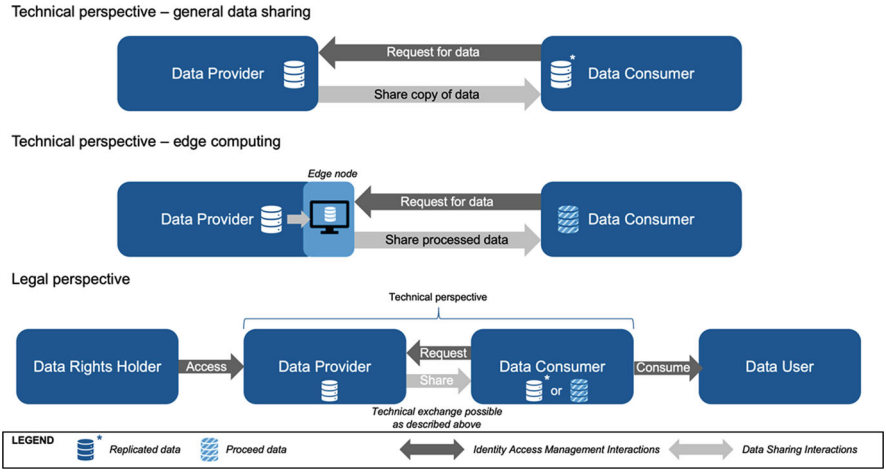


Fig. 32.2 Basic roles in data sharing (Source: INNOPAY)

The next technical question that arises is: how to control the data after the data has been shared. This is a hard problem, because data can easily be replicated and the data sharing party does not control the systems and users of the data-consuming party. One way to go about this is to apply edge computing, by doing the computations on the data at the source and only share the result to the data consumer. Then data control can be enforced in a technical fashion. However, this comes at a price in terms of complexity, cost, and manageability. This is one end of the data sovereignty spectrum.

In practice, decision-makers and engineers often think in terms of risk. For data sharing, this is not different. Edge computing solutions may be a technically sound solution; it may be often overkill (cost, effort) when it comes to many data-sharing practices. On the other end of the sovereignty spectrum is the procedural and legal approach.

Companies and people under verifiable consent then share data and the consumer has the legally binding obligation to only use the data for that purpose. The data provider has no technical means to control the data after it has been shared. He only has a technical and legal means to record a proper consent record, in which the terms of sharing are recorded in a digitally sound fashion and where the authenticity and the identity of the data consumer are secured. In this scenario, the data provider has strong legal recourse possibilities in case he detects that the data consumer is misusing the provided data.

Table 32.1 summarizes the main dimensions of the two ends of the spectrum.

The data provider has the trade-off between functional data control and legal data control. The latter comes with a lighter implementation.

The heart of this decision is a risk, cost, and speed trade-off, and this depends on the specific use case. The consequence of a possible data breach or misuse of data determines which scenario will often be used as a determinate. E.g. health data pose

Table 32.1 Trade-offs between functional data control and legal data control

	Edge computing	Data sharing
Data flow	The result of the computation flows to consumer, less data needs to flow	The data flows to consumer
Consent, authentication, identity	Present	Present, multilateral
Technical implementation	More complex	Less complex
Enforceability of misuse	Technical and legal	Legal

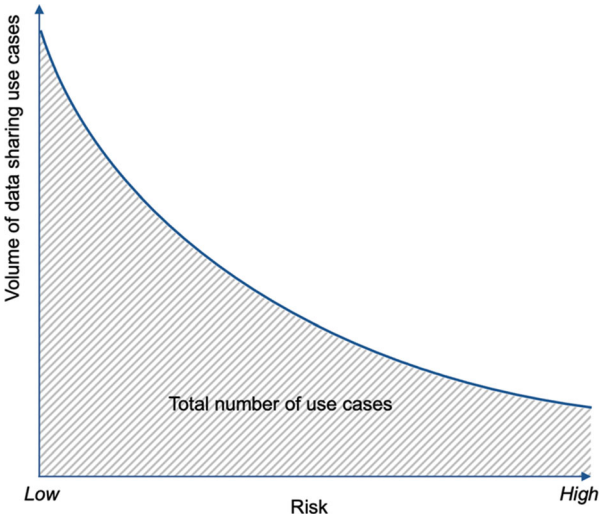


Fig. 32.3 Incidence of use cases as a function of risk (Source: INNOPAY)

a greater risk and therefore require a more stringent treatment because of privacy regulations than e.g. weather data or shipment information of a specific container. Confidentiality in both cases is paramount; and the way to enforce this is different. In the case of medical data, the edge-computing scenario may be used; on the other cases, the data is shared, and possible misuse is legally enforceable because the digital trail of consent, authentication, and authorization has a legal status.

Intuitively, the number of use cases for this technically lighter scenario is higher than the number of use cases for the edge computing solution. Both provide data sovereignty, but in a different fashion and against different cost and effort. The application area is large as there are many applications where actors prefer a lighter implementation as they can oversee the risk of misuse themselves, supported by the legal trail provided by the consent, authentication, and identification mechanism. Figure 32.3 summarizes the relation between risk and the number of use cases.

The remainder of this contribution focusses on iSHARE as a solution for this "entry level" data sovereignty. In the following sections, we explain more and show

iSHARE as a cornerstone of data sovereignty, because of its focus on consent, authentication, and identification.

32.4 The iSHARE Trust Mechanism for Data Spaces

First and foremost: iSHARE is not a platform nor a provider.

iSHARE is a trust standard, also referred to as "trust scheme." The set of agreements describe how APIs for data sharing can implement the identity access management part, in a functional, operational, technical, and legal way. iSHARE participants implement the set of agreement in their data-sharing solutions and participate in the governance of iSHARE.

The things iSHARE provides are the governance, the set of agreements, and the vetting of the participants. It is run by a foundation[3] and governed by its participants. Participants can be everyone who uses data, provides data, and offers technical services to these groups. Often a data user also is a data provider.

32.4.1 Standards and Contract

The set of agreements come with an optional contract, which binds the user to the terms set by the foundation and its governance. This ensures that sharing among participants occurs on the same legal terms. A minimum is defined such that participants have less to negotiate and agree upon when engaging in data sharing. Therefore, the more participants there are, the more options to share data a participant will experience. Thresholds for data sharing are reduced, and costs for data providers and data consumers for engaging into data sharing are reduced.

Data consumers and data users are referred to as "adhering parties," as they bind themselves to adhering to the iSHARE rules and regulations. The parties offering the technical services to actually run the network are referred to as "certified parties." The iSHARE Foundation and its future satellites certify them.

iSHARE only becomes a solution when it is integrated in an IT system of a data consumer and data provider. That IT system can be self-built by data providers and data consumers, but often, this IT system will come from a specialized (cloud) provider. In addition, iSHARE can serve in complement with existing identity and access management solution, offering additional security and legal trust.

[3] https://ishare.foundation

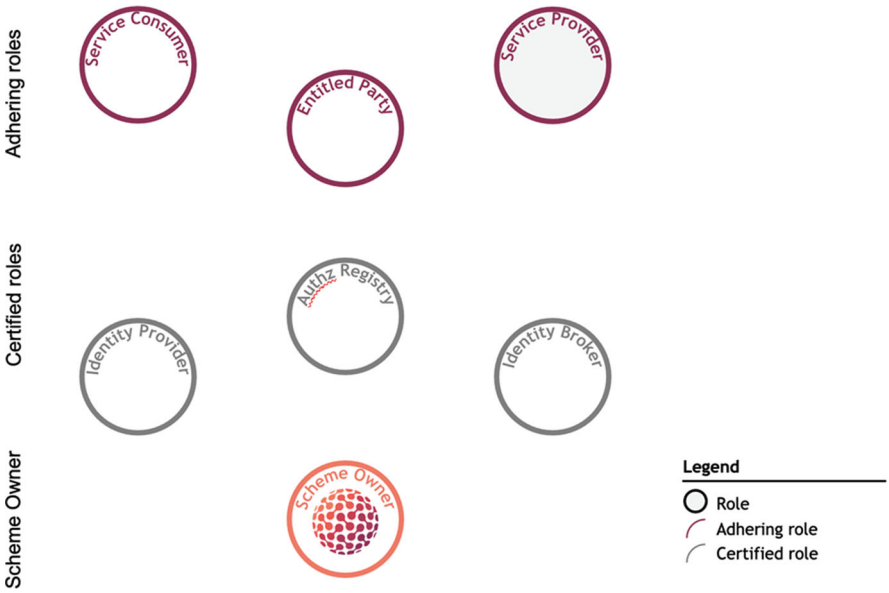


Fig. 32.4 iSHARE role model (Source: iSHARE Foundation)

32.4.2 Connecting Legal Entities to Data

Next to providing an IAM protocol for data sharing APIs, iSHARE also binds the data point to a specific legal identity through its identifier. The identities used in iSHARE are grounded in eIDAS and linked to local, certified, and legal business registries, such as tax identifiers, legal entity identifiers, or Chamber of Commerce. This connection makes data sharing a legal transaction and therefore enforceable through legal means. The iSHARE digital trail in combination with the underlying contracts can serve as proof of data sharing in case of an incident or breach.

32.4.3 iSHARE Roles and the IDS Reference Architecture

iSHARE fits in the IDS RAM architecture as trust solution. The roles within iSHARE are similar to the ones in the IDS RAM as both revolve around data consumers, data providers, and their technical representation thereof. Figure 32.4 shows the role model of ISHARE.

The roles are non-exclusive, except for the directory of participants. The distributed and federated directory, which is executed by the satellites of data spaces, holds the metadata of trust actors within the network. On a transaction basis, this list can be consulted via an API.

The non-exclusivity exclusivity of roles means that multiple parties can perform roles in competition with each other. This ensures the level playing field.

Next to the role of "data consumer" (represented by service consumer) and "data provider" (represented by service provider), we see the "identity provider" and the "authorization registry." The legal identity of the actor comes through the identity provider, which is in practice existing provider of login credentials and certificates. The authorization registry stores the consent of every data transaction. Data providers have the option to provide for their own authorization registry or to outsource this to a specialized third-party provider conforming to the iSHARE rulebook.

The exclusive role of "directory" is to keep track of all actors in the network by storing their digital certificate credentials. Every data transaction can be checked against the validity of the actor in the registry. Noncompliant actors are not in the directory or can be taken out of the directory. In technical terms, the directory is a set of APIs, which actors use in their operations of data sharing. This role is comparable with the DAPS within the IDS RAM, although DAPS stores more attributes of actors than the iSHARE registry.

Figure 32.5 shows where iSHARE fits in the IDS RAM role model.

iSHARE is drawn in as an option for trust provisioning and IAM (Identity Access Management). As such, it can be seen as "basic level data sovereignty," where data sovereignty is mainly enforced via legal means, based on the digital trace of data sharing and consents. The decision of using a lighter data sovereignty solution or a more ruggedized one via edge computing in connectors depends fully in the risk analysis of the particular use case.

32.5 Practical Example: Data Sharing Along the Chain

In practice, iSHARE is a crucial component for data sovereignty, as it is now possible to grant access to information on a very granular level. Access to one or more data points can be given by the data rights holder (also referred to as "entitled party"), under very specific conditions such as time, event status, and identifiers, all linked to legal entities.

Data hubs are very common along all sorts of industrial, logistics, financial, and healthcare processes. These are collections of data about processes, which are under the control of the hub owner. The data rights holder has limited control over his data, e.g., for reuse further down the physical chain.

iSHARE enables data hubs to offer data sovereignty functionality to their customers, which typically are the data rights holders. With this additional service, data rights holders are able to grant access to third parties to this data, by generating and distributing API authorization keys to specific data with specific purpose and conditions attached to this. In this section, we describe three cases of data hubs.

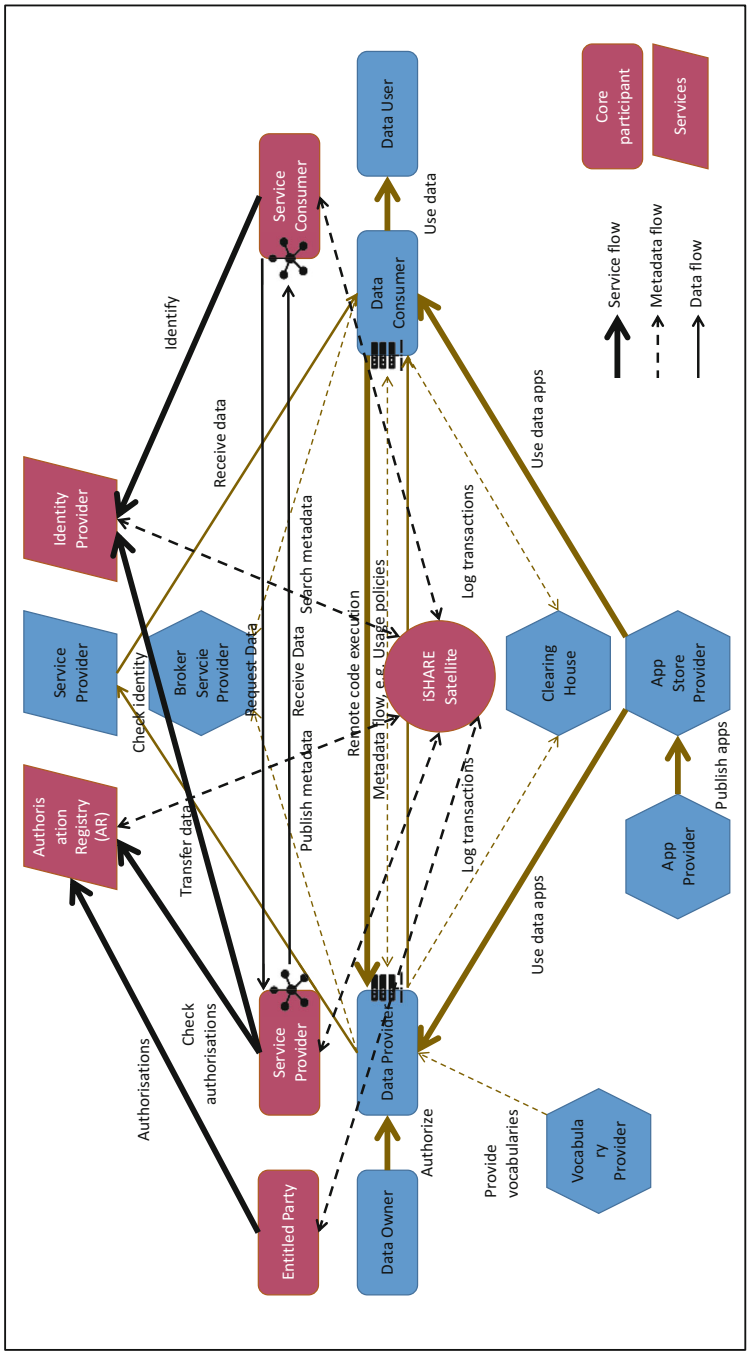


Fig. 32.5 iSHARE within the IDS architecture (Source: iSHARE Foundation, International Data Spaces Association)

32.5.1 Faster Inland Dispatch of Sea Freight

This data sovereignty can very well be illustrated with the case of European Container Terminal (ECT), part of the worldwide Hutchinson Ports company. In this use case, ECT was able to improve data security by using iSHARE for authorizing the pick-up of containers. This is an alternative for the standing practice using a PIN code, which is shared on a need-to-know basis. This is not very secure, as the ownership of such PIN code is very easy to change via messaging, calling, and post-it notes on computer screens.

The handling of inbound containers involves many actors. Think of customers, shippers, ship owners, forwarders, importers, inspections, and transporters. One large container vessel contains more than 15,000 40-foot containers. Unloading one ship leads to a lot of physical handling of containers by road freight, rail, and barge for transporting the container to the next handler in the physical chain, somewhere in Europe (hinterland).

By using iSHARE, the shipper (importing the goods) can receive data about the shipment earlier: instead of being notified once the container is unloaded, the shipper can receive the expected unloading moment in advance, because the sophisticated authorization mechanism of iSHARE facilitates legally sound consent. The carrier can then provide the information to the right shipper. Once this importing shipper has this information, his transporter is sent to pick up the container, also through a proper delegation of rights. So authorizations can also be delegated in iSHARE.

All the authorizations are temporarily, specific for the goal, limited to the need-know part of the information, and legally bound to a certain legal entity and its employees. Data is sovereign in the sense that the data rights holder is in control of granting the authorization. In logistics, the data rights holder (entitled party) may vary during the physical transport. Therefore, delegation of rights is a crucial functionality.

32.5.2 Faster Information for the Road Authority in Case of Truck Accidents

Another example is a data hub, which provides information about trucks and their loads at a certain moment in time. This information is important for the public road authority (in The Netherlands "Rijkswaterstaat"), who needs this information in case of a road accident where trucks containing hazardous information are involved.[4] In case of a road accident, cameras of road authority read out the license plates of the involved truck.

[4] https://www.ishareworks.org/en/news/how-ishare-and-deflog-help-rijkswaterstaat-save-time-and-reduce-safety-risks

This identifier collects all relevant information from various sources, including the on-board trip computer and the e-CMR. The road authority has received the proper and standing authorization from the transporters and shippers, such that this information on the cargo is instantly available. With this information, the road authority can better request and coordinate help from police, fire department, and (road and air) medical support. All the authorizations and delegations thereof are achieved using iSHARE.

32.5.3 Inland Container Terminal Sharing Data Down Stream

The last example (depicted in Fig. 32.6) is about an inland container terminal. Often a container is composed of freight of several importers. The freight forwarder, in collaboration with the inland terminal, would be able to share data about specific freight directly from the inland terminal to the end receiver of a particular freight, even if there are other transporting parties still be involved later to move the goods to the end receiver. The advantage is a better supply chain visibility for the end receiver, allowing them to better anticipate their production or service process. The inland terminal can provide new data services to downstream actors who would benefit from this information.

32.6 Conclusions

In the above, we have proposed practical approaches to data spaces. This starts with thinking in terms of functionality for users, in this case the data consumer and the data provider. In addition, the ease of doing business should be optimal, which means that the "data space" of data consumer and data provider needs to be covering most existing and potential data exchange situations for its users.

Contextual risk analysis for users is another essential approach when realizing data spaces. Each end of the risk spectrum requires different implementations. Data sharing use cases should be selected on this aspect. The data spaces architecture accommodates both ends of the spectrum. The use cases by iSHARE illustrate the possibilities in real-life situation.

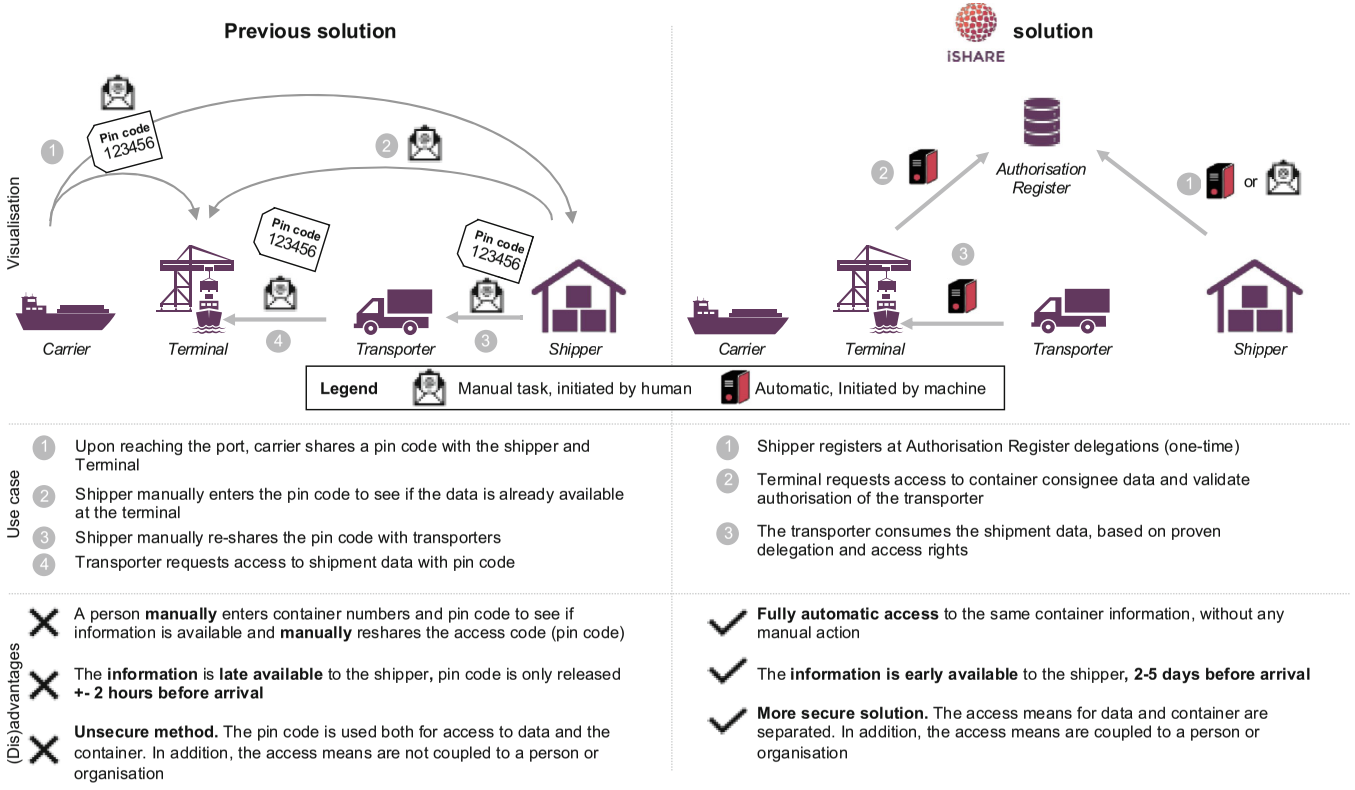


Fig. 32.6 iSHARE for inbound sea freight (Source: iSHARE Foundation)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 33
The Energy Data Space: The Path to a European Approach for Energy



Martine Gouriet, Hervé Barancourt, Marianne Boust, Philippe Calvez, Michael Laskowski, Anne-Sophie Taillandier, Loïc Tilman, Mathias Uslar, and Oliver Warweg

Abstract Trusted data spaces supporting energy services and fostering collaboration between all stakeholders are a cornerstone of the decarbonization of the sector. Today, a broad representation of European energy companies and academic and technological partners has joined GAIA-X to build the European energy data space.

M. Gouriet (✉)
EDF, Paris, France
e-mail: martine.gouriet@edf.fr

H. Barancourt
Atos Worldgrid, Grenoble, France
e-mail: herve.barancourt@atos.net

M. Boust
Capgemini Invent, Issy-les-Moulineaux, France
e-mail: marianne.boust@capgemini.com

P. Calvez
ENGIE, Courbevoie, France
e-mail: philippe.calvez1@engie.com

M. Laskowski
E.ON, Research & Development, Essen, Germany
e-mail: michael.laskowski@eon.com

A.-S. Taillandier
IMT, Paris, France
e-mail: taillandier@imt.fr

L. Tilman
Elia Group, Brussels, Belgium
e-mail: loic.tilman@eliagroup.eu

M. Uslar
OFFIS - Institute for Information Technology, Oldenburg, Germany
e-mail: uslar@offis.de

O. Warweg
Fraunhofer IOSB-AST, Ilmenau, Germany
e-mail: oliver.warweg@iosb-ast.fraunhofer.de

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5_33

The group represents all segments of the energy value chain and is from all around Europe.

Through this data space, we aim to address the following challenges: accelerate the deployment of low carbon energy solutions, foster energy efficiency and sector coupling (power, gas, and heating, integration of mobility and building/heating systems, etc.), enable more flexibility and renewable energy integration to the European electric system, accelerate the sector digitalization, and ultimately support Europe competitiveness, thanks to low energy costs.

To achieve these goals, strong collaboration between the actors is needed to identify and launch valuable use cases on key topics: renewables, hydrogen, nuclear, energy efficiency, electric vehicles, local energy communities, networks, or compliance and traceability.

The article is a collaborative effort, initiated in February 2021, from the French, German, and Belgium energy communities within GAIA-X national hubs. The intention is to provide insight on the work of the GAIA-X energy Domain, to share widely our ecosystem's expectations, and to provide an overview of use cases identified.

33.1 Mission and Goals of the Energy Data Space

The goals of the energy data space are:

- To support and accelerate the energy transition in Europe
- To develop businesses at European and worldwide scales
- To provide new services to European citizens, taking advantage of GAIA-X infrastructure, data protection services, and associated value allowed by these services

The GAIA-X Contribution to the Energy Domain

Energy European stakeholders expect strong business benefits and joint tangible outcomes from the creation of the GAIA-X ecosystem. We actively support the emergence of federated services and data platforms that create value and opportunities for all businesses—and ultimately for European citizens.

Through the GAIA-X initiative, we—at last—foresee a tangible and close way forward to easily build, assemble, and use trusted and value-creating cloud services, as well as create new products/services and foster new business models that are compliant "by design" with European regulations and values. For that reason, it is not only about the data but also the opportunity to build a strong European digital system layer on top of the existing energy system layer which should be seamlessly coupled.

Added Value to the Whole Energy Value Chain

Thanks to data-driven solutions, the energy data space aims to help manage the transition toward decarbonized energy and carbon neutrality. Other objectives are to

enable energy efficiency and sector coupling (green energy fluids, integration of mobility and building/heating systems, etc.) and also enable more flexibility and renewable energy integration into the European electric system. Within GAIA-X, energy stakeholders will foster European energy system optimization and competitiveness. The European ecosystem is being identified and organized to design relevant use cases on national and cross-border scales: the governance within this ecosystem needs to be clear to ease and accelerate the energy sector digitalization and its adoption. Thanks to data-sharing possibilities between stakeholders, valuable use cases will be addressed so that new services for European citizens and companies can be deployed.

33.2 Challenges Addressed for the Energy Data Space in Context of GAIA-X

Addressing these use cases requires cross-border common data knowledge representations, semantic data models, data collection, and sharing data capabilities. Accessing data in trusted and collaborative cloud infrastructures is compulsory to provide comprehensive offerings under high security standards. The GAIA-X strategy of enabling access to aggregated, federated, and interoperable trusted cloud services and AI services through the setting of common policy rules will ensure data, data knowledge representation, and services protection, but also interoperability and portability. These are crucial to reinforce trust and transparency in order to scale, digitalize, and provide new services at a European level. Moreover, there is a need to have real-time data exchange in order to improve the system flexibility and leverage the electrification of the system.

Since highly secured data will be processed in the energy data space, dataspace stakeholders will be very sensitive to providers with highly secured standards: labelled GAIA-X.

The challenge is also on governance and organization. Big energy companies—some of them being great competitors—are joining their efforts to provide to their national and cross-border customers new services that will help business growth, social improvement, and carbon neutrality reach.

Large amounts of heterogeneous data can be gathered to address the most valuable use cases; therefore, the whole value chain must participate to this joint effort:

- *Energy providers*: for the production and storage of energy for all the different kinds of decarbonized energies, energy efficiency, and new services
- *Contractors, partners, and engineering service providers:* all companies and partners to improve energy production and engineering
- *Gas, power, and heat network operators (DSO/TSO/Stadtwerke (engl. Municipal utilities in Germany):* with energy transportation and distribution data
- *Aggregators*: collecting and packaging data to provide it to consumers

- *Energy consumers,* as households, local collectivities, industrials (smart metering data, etc.)
- *EV station managers*: with vehicle and charging stations data.
- *Safety and certification bodies*: with safety and certification data.
- *Open services*: with public maps, meteorological, city, or transport data.

33.3 Solution: Data Space Description in a Holistic View—Detailed View on the Endeavor

33.3.1 Partners of the Ecosystem

There is already a broad representation of energy companies in the data space, from all around Europe, especially from Germany, Belgium, and France. These companies represent all segments of the energy value chain. They are already able to develop smart services (AI, IOT, machine learning, blockchain, etc.) and to integrate suppliers to develop these services. And they are supported by academic and technological partners.

33.3.2 Description Use Cases: Renewables

33.3.2.1 Renewables: Wind and Solar Asset Description Model

Solution
The wind and solar description model is the digital backbone to federate all the businesses and its ecosystem around one single source of truth from "DESIGN, MODIFICATION to others REFRESHMENT". This structured, agnostic asset management description supports business process orchestration in a context of new business model refocused on growth of decentralized and distributed energy and carbon neutrality "as a service."

Asset data management and valorization lay on a change in paradigm which needs to consider specificities of the data and the complexities of the life-cycle management:

- Different technologies offered by equipment manufacturers: Vestas, Siemens, etc.
- Diversity of contracts with equipment manufacturers and data providers
- Strong commitment to third parties on the performance rate of turbines
- Large volume of data and inconsistency making it impossible to share the results and use them (interoperability)
- Difficulties accessing data from some farms, etc.

The theoretical modeling of assets is a new paradigm all along their life cycle, enabling to link the two "worlds" as build and operate. Take the following examples to support this new paradigm:

- The development of the renewable assets should define precisely and realistically the layout from trusted and updated technical information. With such asset configuration management, scenario can be defined, simulated by modifying, combining design parameters, and analyzing components and their assembly to understand and assess the impact of any exogenous effect implied by layout constraints and changes to any layout.
- Thus, it is easier to consider the as-built model and update the final implementation file (Dossiers Ouvrages Executés, DOE), including underground networks and turbines.
- This asset modeling is key to compare the asset production between sites.
- This framework enables to understand how farms are operated and the root cause of observed discrepancies.
- This framework enables to calibrate the operational team onsite to better regulate and organize technicians' activities according to travel times.

This framework is a new asset-centric form of shared master data. Each equipment is linked to a model (technical description of the assets included components, physical model, localization, etc. as metadata) to ensure coherence and value of data with a "single point of truth" access.

It lays foundations for a virtual plant, ensuring continuity of data all along the life cycle and sharing in ecosystems to drive both traditional and disruptive new business models.

Problem Solved

When it comes to manage efficiently and effectively the renewable assets and all the O&M actions for the different actors (manufacturers, owners, developers, operators, etc.), the difficulties to access to data exploitable in a cross-functional way by these different profiles of actors and the different digital systems mobilized represent a real challenge.

Being able to share this same abstract representation of data for the wind and solar domain would allow a better understanding of the associated operations (asset management, RCA, structural analysis, visual inspection, monitoring, etc.) and an obvious improvement of the processes that mobilize the processing of this information.

The availability of standards (IEC 61400-25, etc.) provides a solid reference framework, but these are not transposed into semantic paradigms (ontology based), even if some initiatives exist in this transposition [1].

Concrete Benefits

The asset description model is an opportunity to build comprehensive models and analytics frameworks and improve multiparty collaboration capabilities needed to support digital ecosystems. It is a backbone for renewables operator to ensure continuity of technical data all along the life cycle.

This "single point of truth" is a real accelerator for Greenfields and brownfields assets to deliver more safely, more quickly, and more efficiently (right the first time) and with a lower total cost of ownership. It lays foundations of virtual plants with the following objectives:

- Projects designed right the first time, based on reliable and available data
- Projects delivered for operations, in compliance with specifications
- Enhanced preparation of interventions, using as operated configuration
- Break organization and information systems silos. Each business receives accurate information across the world, which makes it easier to achieve the ambition. For example, issuing an alert when a design change made in one area may have implications in another.

33.3.2.2 Renewables: Work Risk Prevention

Within the domain of safety management, it is common experience that the aim to provide a safe working environment for everyone who enters it becomes more complex when different actors, belonging to different companies, have to operate in the same industrial sites. The key aspects remain, that is, sharing risk assessments that are up to date, identifying areas with potential hazards, anticipating what could happen and what to do, and making sure that the information is known and taken into consideration.

With a large number of industrial working places/situations for solar, wind, or hydro generation, with different hazards due to height, lifting, electricity, fire, pressure, etc., safety management is achieved by numerous dedicated works prevention plans that need to be set up after contracting the work. Before starting the works, a joint inspection must be organized that requires different data on the site and the work, the industrial team, and the contractor's team. The efficiency of the plan that is produced depends on the experience on each side but also on the reliability with the exchanged data.

The difficulty to produce paper risk prevention plans for works in industrial environments leads to the perception of bureaucracy instead of responsibility. New horizons open up by digitalizing the global process with a positive aim to improve the quality, to construct a stronger capacity of hazard identification and control.

Problem Solved

Industrial works generate different situations with a high level of risk. In France, an executive Decree (N° 92-158) on specific health and security rules applying to the building, public work, and water work sectors describes the required documents and steps that should be produced by the supplier. The key component is the risk prevention plan (RPP) which includes a mutual assessment of the risks.

Similar requirements exist among industries across EU; establishing, updating, and reporting on work risk prevention require a lot of coordination between the two

parties and sometimes with the state. The information exchanged carries personal and commercial data.

Digitalizing the process in an extended Enterprise approach requires trustworthy data storage easily accessible to all kinds of companies—especially for different suppliers working on different industrial sites within the EU.

Different software packages are marketed, usually based on one industrial attempt to federate its suppliers. To go one step forward and gain time and efficiency and simplify the process, software solutions for work risk prevention need a framework in order to develop new processes across the different stakeholders.

Solution

The energy data space could accelerate the digitalization of the life cycle of risk prevention plans needed to fulfill EU/state requirements in terms of risk evaluation, risk reduction, and work authorization.

The contractor and the industry would share data between them and sometimes with the State through digital canals accessible onsite and from their respective offices, each party having access to a specific part of the information.

The data space will guarantee that personal/commercial/sensitive data can be processed and shared through different means of communication. It will offer facilities of integration with industrial IT systems as well as easy user interface for small companies.

Concrete Benefits

The main benefit is to increase safety of work, through better risk assessment, thanks to better data sharing and analyses.

To prove the value of digitalizing the process, an experiment has been carried out on several HYDRO generation units using Software as a Service with storing data on a small cloud provider. The analysis shows that with 15,000 prevention plans per year on hydro generation works, EDF has the forecast of a large saving per year (over 500 k€) on hydro maintenance. The data space is a compulsory service for the success of such a project.

The benefits at a large scale would provide gain for suppliers as well as industrial actors, resulting in less bureaucracy and more responsibility that will accompany a growing approach over risk management.

Questions/Answers Identified at This Stage

- How can actual software/cloud providers involved in Risk management become a validated GAIA-X provider? A proof of value could be carried out at a small scale in order to illustrate how complicated it would be for a given solution to become compliant. An existing cloud provider (already ISO 27001 and qualified for Health Data Storage) could join GAIA-X AISBL, and an existing software provider would implement the needed federation services.
- How can I impose the use of the data space energy to my suppliers? Gradually, suppliers will identify the benefits for themselves and will choose to comply.

Before that stage, it will be possible to take into account the added value, like respecting the semantic catalog, when contracting the works. Data space energy requirements would be added to the purchase process.

- How can I buy such a service—do I need to have many solutions in order to establish a competitive bidding? In order to set up a solution to carry out the proof of value, an innovation contract will be used through a consortium of industrials ready to make use of the solution on a number of works within a definite period. This will enable suppliers, large or small, to use the solution with no extra cost.

33.3.2.3 Renewables: Common Taxonomy Definition: IEC Standards

Solution

The objective is to propose an approach and associated services that would allow to do a job or even to provide automatic alignment systems between different information models. For this purpose, we seek to align existing models across sectors and to find transversal vectors of common information model (e.g., CIM, etc.) combined with the use of ontology building approaches. Considering the other GAIA-X CUs and focusing on transposing these CUs for the renewable sector in a CTD (Common Taxonomy Definition) would allow to link and valorize the standards around the standard description of use cases, Role Model, Canonical Data Model(s), or even architectures for the energy value chain such as SGAM (Smart Grid Architecture Model).

Problem Solved

In the energy value chain, which includes a multitude of actors and processes making it very complex, there are already reference and standardization frameworks that propose common representations to facilitate the understanding, exchange, and operation of associated systems and subsystems. The matrix model associating the major energy verticals (Generation, Transmission, Distribution, DER, Customer) and sectors (Process, Field, Station, Operation, Enterprise, Market) has an IEC core model for the energy and electricity domain, which enables the management of associated information models: CIM Data Models, COSEM Data Model, IEC 61850 Data Model, CGMES, etc.. . . but also models/ "Standard "cross sector as SAREF, CIM+, NGIS-LD, FIWARE, . . .

Nevertheless, although based on modeling standards such as UML, the semantic links between the different uses of these different standards, information models, and ontologies are not a reality. This deeply limits a higher level of transverse interoperability.

In addition, the energy sector can be broken down into several sectors composed of several fluids and the actors and infrastructures that transport them. We are referring here to three main models: electricity, gas, and heat (cf. IEC 63200).

We can then understand the extreme difficulty for systems and subsystems (increasingly digitalized) to navigate and mobilize these different representations and the complexity of aligning them.

Concrete Benefits
Although the task seems extremely ambitious, namely, to align proven standards (IEC) and Canonical Data Model (CIM, etc.) with ontologies (SAREF, OneM2M, etc.) or Linked data formats (NGIS-LD), the opportunity provided by GAIA-X to align actors, CUs, and datasets to define advanced interoperability models opens up real prospects.

It seems imperative to consider the use of ontologies to achieve these cross-sector connections. The benefits are extremely numerous in the long run, because such approaches would allow to federate different knowledge spaces and representations in the energy domain without reinventing the wheel and consolidate the years of formalization.

33.3.3 Description Use Cases: Nuclear

General Context
Essential for the fight against global warming, the nuclear industry supplies energy in a sustainable and scalable manner:

1. A low carbon energy: France is today one of the countries emitting the least greenhouse gases and can rely on its electricity system to reduce CO_2 emissions in other sectors (e.g., transport, industry, construction).
2. A competitive energy which benefits all economic agents: individuals, companies of all sizes, etc.

The French nuclear industry, which controls the entire nuclear energy production chain, from uranium extraction to spent fuel reprocessing, is a benchmark worldwide, thanks to its technologies, skills, and employee know-how. It represents more than 2000 reactor-years of experience.

With its 220,000 employees and more than 3000 companies, 85% of which are VSEs and SMEs, it also contributes, through its establishments throughout the European territory, to the development of local economies.

GIFEN: A Key Player Federating and Transforming French Nuclear Industry
Due to the risks of industrial espionage and takeover bids, as well as the need to perpetuate European industrial factories, the future of the sector depends on its ability to ensure its industrial and technological sovereignty. It is critical for French and European energy sovereignty as well as for export capabilities of nuclear industry companies.

On the strength of these convictions, the companies in the French sector are united within GIFEN – by a common objective: to build together the French nuclear industry of today and tomorrow.

GIFEN brings together nuclear facilities operators (EDF, ORANO, CEA, FRAMATOME, ANDRA), large companies in construction and engineering (ENGIE, VINCI, BOUYGUES), mid-size companies in construction and maintenance service, SMEs, VSEs, IS/IT providers, software vendors, electronic manufacturers, local and professional trade organizations, and associations, covering all types of industrial activities (studies, manufacturing, construction, maintenance, etc.) as well as all areas of nuclear power generation (fuel cycle, research, power generation, equipment manufacturing, decommissioning, etc.). At the beginning of 2021, it represents more than 230 companies.

Thus, GIFEN addresses all cross-functional stakes in the service of industrial excellence and offers services of common interest. In particular:

- Map the skills of the sector to anticipate future needs
- Consolidate workload and purchasing forecast at short, mid-, and long terms, according to 18 pre-identified business families
- Analyze the supply chain workload capabilities in response to this forecast in terms of skills and industrial tools and implement actions to guarantee it
- Put mid-sized companies/SMEs in touch with the major clients, who manage the major R&D programs in the sector
- Develop collaborative platforms aimed at facilitating exchanges through the supply chain and increasing the studies and manufacturing quality
- Support companies toward better safety culture and nuclear quality (e.g., deployment of ISO 19443 norm)
- Structure French nuclear industry strategy for its international development regardless of the technology

GIFEN works in close collaboration with French and European public authorities.

Identified Opportunities Within GAIA-X

Regarding presented challenges concerning the nuclear sector, GAIA-X has the opportunity to define and implement technical, functional, organizational, and governance solutions allowing the mastery of shared data as well as the necessary support to further deploy the digital transition of the nuclear industry.

The mapping of the sector capabilities using a data-centric approach and the usage of collaborative platforms around key extended enterprise stakes illustrate the fundamental pillars of this transformation.

Considering this approach, five use cases are proposed:

1. Day-to-day collaboration capabilities within the GIFEN
2. Nuclear industry observatory: capabilities mapping and related data analytic services
3. Usage extension of the *ESPN*[1] *Digital collaborative* platform for the whole sector

[1] Equipment Sous Pression Nucléaire: Nuclear Pressure Equipment

4. Standardization and digitalization of maintenance work packages[2] (eWP/eDRT in French) through a collaborative platform, with opportunities to cover operations in non-nuclear assets
5. Optimization of nuclear waste management

The use cases put forward are not an exhaustive list of "nuclear use cases" within the framework of GAIA-X. It is rather a selection of the more advanced and significant cases, calling on different bricks, to experiment and gain maturity:

- On the efficiency of the services being defined by IT providers
- On the contribution of a common data space for energy players

Moreover, these use cases have the advantage of being under investigation from a business and IT perspective within GIFEN working groups.

33.3.3.1 Nuclear: Day-to-Day Collaboration Capabilities within GIFEN

Solution
To contribute to the operating performance and development of the nuclear industry, it is essential to have collaboration services between actors under the GIFEN governance. This means producing, storing, and exchanging information of various formats and criticalities in a secured manner by guaranteeing access control, traceability of exchanges, as well as the correct application of the usage rules. It is a prerequisite for the GIFEN, in order to succeed in the missions defined by the industrial companies of the sector.

For instance, exchange concerns industrial and technical project data, export market data, industrial feedback, studies, applied industrial and technical standards, etc.

Problem Solved
Regarding the GIFEN missions, protection of data and their sharing are a critical issue. This protection covers:

- Know-how or patents (intellectual property)
- Financial interests
- Process continuity and integrity
- People and companies' ecosystem

In addition to these items, the main requirements are:

- Authorization and accreditation management
- Access traceability
- Storage whatever the format of the information

[2] Dossier de Réalisation de Travaux électronique: electronic Work Package

- Functional modules dedicated to small entities (file sharing, chat, conference call, validation workflows, etc.)
- Interoperability between the information systems of main nuclear players and their partners

Concrete Benefits
A common language and associated collaboration solutions represent the basis for better collaboration, creating trust between stakeholders. That makes possible fostering innovation and the development of common projects. The benefits are difficult to quantify at this stage, in particular the indirect benefits linked to common projects that may be put in place. Most of operational benefits are around avoided costs: rationalization of collaborative spaces, consequences of a data leak, and data inconsistencies.

33.3.3.2 Nuclear: Industry Observatory, Capabilities Mapping and Related Data Analytic Services

Solution
The observatory would include information from various players of nuclear industry (operator, manufacturer, designer, etc.) allowing consolidation of a detailed mapping aligned with GIFEN missions. As instances, it would be valuable to gather in particular:

- Legal form of the actors
- Workforce in terms of volume, skills, and distribution on the European territory.
- Workload forecast at short, mid-, and long terms based upon main industrial project assumptions
- Breakdown over the 18 business families defined at GIFEN
- Information on international critical markets for the nuclear industry
- Innovation programs data

Mapping would be the basis of data analytics-like services enabling to understand, predict, and make decisions on the industrial system in response to main nuclear industry stakes.

Notably, the adequacy analysis of the supply chain workload capabilities regarding the operators and main project forecast in terms of skills and industrial tools represents a top priority. It will significantly ease the identification of under pressure job and competencies and ease the definition of the related action plan.

It means implementing an interoperable system for collecting, analyzing, and sharing this information in a distributed or centralized and secured manner.

Problem Solved
The setup of a "nuclear industry" data space is a prerequisite for several fundamental missions of GIFEN:

- Map the skills of the sector to anticipate future needs
- Consolidate workload and purchasing forecast at short, mid-, and long terms, according to 18 pre-identified business families
- Analyze the supply chain workload capabilities in response to this forecast in terms of skills and industrial tools and implement actions to guarantee it.
- Structure French nuclear industry strategy for its international development, whatever the technology

GIFEN has already initiated these actions with its own resources to demonstrate their value. A roadmap for industrialization is under definition. Without a trust and governed workplace nor rules for collaboration and sharing, it remains very difficult to implement this strategy.

In addition, interoperability with the upstream and downstream areas of the energy value chain, as well as related sectors, which are highly capital intensive and require common skills such as infrastructure construction, is essential at mid-term. Indeed, transversal optimization is part of European Sovereignty and, thus, deserves to be tooled.

Concrete Benefits

The benefits are difficult to quantify at this stage. They will be detailed for each added value services made possible by this observatory, notably orientations on the nuclear industrial system (manufacturing facilities, design center, etc.) taken following the analysis of the adequacy of charges/resources in the short, medium, and long term.

This will allow to strengthen the industrial policy of the nuclear industry in the medium and long term, in particular by:

- A better identification of project risks
- An eased consolidation of export opportunities

33.3.3.3 Nuclear: ESPN Digital Platform for the Nuclear Sector

Solution

Background and ambition

The "ESPN Digital" project was launched as part of the EDF SWITCH program. It is the digital extension of the industrial control strategy of the Nuclear Pressurised Equipment[3] regulation. The aim of these regulations is to guarantee a level of control of pressure equipment adapted to the nuclear safety issues at stake.

ESPN Digital is a digital platform built to facilitate the assessment of regulatory compliance of ESPNs. It complies with regulations while making exchanges more

[3] ESPN: Equipements Sous Pression Nucléaire (French)

fluid between operators, manufacturers, certification agencies, and the ASN,[4] enabling them to refocus on their core businesses, collaborate better, and become more competitive.

The first version was developed and put into service on a limited number of projects. The ambition is to extend its usage to all the industry players, for the construction and operation of new ESPN equipment intended for nuclear facilities, but also for modifications or repairs of installed equipment.

Service description

The platform allows pooling and tracing all the information provided by stakeholders on ESPNs: history, provided justifications, data, workflow, etc.

Eventually, ESPN Digital will make available guidance to be followed and enable everyone to draw up with confidence all the documentation required by the regulation and progressively automate its edition, thus facilitating the work of the ASN or the authorized agency (which will eventually lead to a reduction of the time required to obtain the ESPNs certificate of conformity).

Problem Solved

GIFEN wishes to respond to the needs and difficulties raised by the actors of the sector for the application of the regulation through this project. The main difficulties put forward are the following:

- Very long and unpredictable conformity assessment times
- Heterogeneous practices of manufacturers and organizations
- A limited number of actors mastering the ESPN process

Thus, the ESPN Digital platform extended to all the industry should help meet these challenges by centralizing and making exchanges and collaboration more fluid between all conformity assessment stakeholders. The platform also has the virtue of harmonizing work methods and digitizing all processes to comply with regulations. Finally, the digital tool will offer a consolidated and 360° vision of ESPN Safety Cases and Certification processes to all stakeholders.

Concrete Benefits

ESPN Digital stakeholders expect many gains through the implementation of this project:

- Direct benefits:
  - 20% of the cost of compliance certifications
  - Estimated savings of 15 M€/year in the current configuration of the EDF fleet + FA3
  - ~80 M€ for a new project (FA3 type) over 8 years
- Major indirect benefits:

[4] Agence de Sureté Nucléaire: the French Nuclear Regulator

- Better control of project planning and potential impacts on the availability of nuclear facilities
- Lowering entry barriers linked to ESPN regulations

33.3.3.4 Nuclear: eWork Platform

Solution

As part of the digitization of the sector, one of the goals of this intermediation platform is to standardize the Work Order (eWork) and its set of associated documents constituting the electronic Work Package. This use case will facilitate B2B exchanges between the various companies performing interventions on nuclear sites.

To facilitate the adoption of this digital Work Order (eWork) by all the companies, the platform must consider the following requirements:

- Guarantee a multitenancy architecture where access to documents and data of each member of the industry is secured.
- Allow access to eWork information through mechanisms adaptable to the digital maturity of each partners
- Implement mechanisms to verify the compliance of eWork with nuclear industry standards
- Provide configurable workflow functionalities
- Offer visas/signature mechanisms for paper and digital documents such as digital procedures or QA documents.

The nuclear industry partners will have to define and agree on the standards for these digital data exchanges.

Problem Solved

The digitalization of the electronic Work Package allows the numerous companies intervening on nuclear site to have an interoperable and exchangeable format with the eWork. It can be exchanged and enriched by B2B exchanges between the players. These exchanges will be orchestrated to guarantee the best practices in the sector.

This digitalization should enable the sector to have at its disposal several software solutions compatible with the GIFEN recommendations.

For the many companies working on nuclear site, it will be possible to use, when appropriate, their own Work Force Management (WFM) tool well adapted to the company's process, by simply adapting their interfaces. These interfaces will be well defined thanks to this use case.

Concrete Benefits

The digitalization of DRT allows the ecosystem to gain in efficiency by sharing the WO schedules and by digitizing all the data relative to this eWork.

This digitization will allow efficient B2B exchanges between several hundred companies that interact on field interventions in nuclear site.

Digital continuity will reduce the breaks in the chain between the client and the contractor:

- Starting field activities quicker by transmitting information via telecommunication networks
- Completing field interventions without returning to the office, facilitating the optimization of the planning of all interventions
- Suppressing the costs of scanning and printing files; gains on this side contribute to a very short ROI.
- Offering the possibility of optimizing maintenance operations by processing data analytics on a greater quantity of data captured.

33.3.4 Description Use Cases: Low Carbon Hydrogen (H2)

33.3.4.1 H2: Import/Export International Routes Setting up

Solution
A dedicated low carbon hydrogen import/export market will connect producers with end users, similar to today's current gas market. There will be a need for local marketplaces, European marketplaces, and international marketplace. These marketplaces will integrate various forms of hydrogen (gaseous, liquefied) and will need to feature the certification of low carbon hydrogen (see specific use case).

To support the marketplace, a transparent data platform will be required to:

- Map hydrogen production based on location and relevant stakeholder
- Monitor supply and demand
- Monitor prices based on the different products (spot vs. long-term)
- Monitor management rules in case of unused capacity vs. congestion
- Enable market settlement after transactions

Problem Solved
Low carbon hydrogen production could take place next to consumption centers. However, in most cases, hydrogen will be produced where it will be the most cost-effective—because renewable electricity is cheaper and because gas infrastructure can be re-used. We are currently missing the global overview of hydrogen supply and matching demand.

Concrete Benefits
According to Fraunhofer ISE, a market for hydrogen import/export could be worth between 100 and 700 billion € per year in Europe.

33.3.4.2 H2: Station Networks Information Sharing

Solution

The emergence of these H2 mobility technologies will be linked with the ability of the sector to develop a wide range of H2 stations allowing the user to move without constraints.

We are missing a global view of existing H2 stations, with their characteristics and the global matching H2 demand for mobility, to ensure a quick development of H2 mobility.

To support the development of H2 mobility, a transparent data platform will be required to:

- Map H2 stations based on location and characteristics
- Monitor H2 stations capacities and H2 demand

Problem Solved

Mobility is today largely contributing to CO_2 emissions in Europe, and solutions rely on the development of low carbon mobility such as EV and H2 mobility. EV is developing more and more, while H2 mobility is still in the early phases. The development of H2 mobility is widely based on the capacity to propose a network of H2 stations that fits the needs of the consumer.

H2 mobility represents today a very small part of the mobility market, led by traditional carbon technologies, and emerging EV.

Concrete Benefits

The benefit for such a sharing of information is difficult to assess, but it will allow a fast development of H2 stations across Europe, optimized for the use of H2 mobility users.

33.3.4.3 H2: Mobility Asset Monitoring

Solution

The focus of this use case is to build common knowledge representation and collect and share data on operations of hydrogen refueling stations and hydrogen vehicle fleet. This could include data on main design characteristics, energy consumption of HRS and individual major component, energy efficiency of HRS and individual major components, hydrogen production rate, refueling time, vehicle efficiency, etc. A good basis of work is the listed data in "JIVE and MEHRLIN Performance Assessment" framework, and this could be extended to higher collection frequency and a wider range of HRS and vehicle.

GAIA-X enables for this use case to provide a complete functional and technical framework which for this use case will provide the means to make available throughout the value chain and between multi-actors the data and representations shared to build portable and secure services that can address a federated monitoring

by the actors concerned while respecting the rights and access to information, moreover, the availability of these data and representations to allow the creation of new services by third parties who will commit to use these open and shared repositories.

Problem Solved
The provision of data depends on the local data management, which is used by the operator onsite, although there is no integrated common infrastructure.

It will allow to develop services around maintenance, individual component improvement and development, and HRS architecture improvement.

Concrete Benefits
This demonstrator focuses on the brought usability of vehicles and refueling infrastructure data. We see a huge retention to provide infrastructure data for further purposes due to the risk of being misused. By using GAIA-X and its components of identity and trust, sovereign data exchange, and compliance third parties, it will allow to develop services around maintenance, individual component improvement and development, and HRS architecture improvement.

33.3.5 Description Use Cases: Energy Efficiency

33.3.5.1 Energy Efficiency: Energy Renovation—Map Building Potential for Renovation

Solution
2021 marks a strong acceleration of the energy-efficient building renovation policy, with very ambitious targets in terms of the numbers of buildings renovated. Concrete financing mechanisms across European member states have been put in place to subsidize residential and tertiary building renovations. The platform we would like to develop helps in identifying and prioritizing the building renovation programs to be carried out; obtaining, in a context of multiple levers to master, the right financing; and managing and monitoring the energy performance of renovated buildings over time.

Our platform offers automatic data collection and standardization flows from multiple databases, machine learning models, and data sharing principles between public and private players in the energy renovation ecosystem and therefore builds on a comprehensive and continuously updated database. This database would provide a set of detailed information on all the characteristics of the building and benchmarks for building energy consumption. It would make it possible to target the appropriate renovation work, illustrate the possible benefits, and advise on available fundings. This capacity was illustrated in the tRees project, developed by the namR start-up, which allowed them to identify all the educational buildings in the Haut de France region in France (19,734 buildings) and to characterize them with more than 150 attributes.

Problem Solved

Most players at national and local levels underline a lack of reliable data to define, finance, implement, and manage their energy renovation policy and their assets strategy. The pain points identified include:

- A partial vision of their assets, difficult information to collect, limited internal human resources, and costly and numerous external diagnostics and audits to implement to gather information
- Difficulty in identifying the renovation work that could be carried out and in estimating the associated financial and energy gains in order to be able to prioritize them
- Multiple funding levers to know and master, complex paperwork to undertake, including intricate analysis to be carried out
- Insufficient energy monitoring of buildings, while local authorities are to achieve 10-year energy reduction objectives and must manage new information on a regular basis
- Segmented industry with low interaction between stakeholders
- Archaic tools for planning, construction, and operation

Concrete Benefits

Improve the energy efficiency and the energy bill of buildings, in line with the promise of a low carbon balance, to reduce the final energy consumption by at least 40% in 2030, 50% in 2040, and 60% in 2050. This will be enabled thanks to the ability to:

- Obtain/acquire quantitative and qualitative characteristics on buildings, local territory characteristics, and the energy renovation sector and compare them with other territories
- Target the buildings to be renovated and the types of renovation to be carried out and simulate the expected gains
- Determine the sources of funding available and simplify requests, in particular within the framework of the recovery plan
- Access a personalized tool library and a network of experts relevant to the renovations to be undertaken
- Estimate and monitor the financial and environmental impact of planned and completed renovations
- Cost-effectiveness ranking of various renovation technological packages

33.3.6 Description Use Cases: Electric Vehicle

33.3.6.1 Electric Vehicle: Energy Roaming

Solution
A platform that allows customers to freely move and charge between their home charger and partnering public charging networks. Allowing a view of all of their EV "fuel" in one place.

This platform will make use of online maps to help locate chargers, give an availability status, and also allow access to book a charge point, start, stop, and pay for a charge.

It will provide reporting and data. This will require the access of data from various e-Mobility Service Providers, charge point operators, and vehicle OEMs.

Problem Solved
The problem to solve is that EV drivers can find charging their cars confusing and worrying. There are many data sources separately held that could unlock solution and value if they were aggregated and centralized for use and analysis.

- Multiple smartphone apps/RFID cards are needed to access the different public charging networks
- Varied pricing rates and structure on all networks—difficult to compare competitiveness (connection charge only, connection + unit rate, unit rate only, pay for a time period, etc.)
- Difficult to plan longer journeys across multiple networks you are a member of
- Difficult to know which networks are more reliable/trusted brands for quality and service
- Users frequently find charging points not working or already in use
- EV users must plan their journeys more akin to that of a commuter using public transport. It takes time and effort and can cause worry on the road
- Limited real interoperability of networks

The situation can be exacerbated if a driver does not have a home charger or lives in an apartment block with no EV charging facilities (which is likely at the early stage of EV market development).

Concrete Benefits
GAIA X will enable the creation of a new service around the EV's owner experience. This shall occur by producing a set of standards aimed at facilitating the coherent centralization of data owned by the relevant partners involved (CPOs and EVs).

33.3.6.2 Electric Vehicle: New Services

Solution
A platform/ecosystem where:

- An EV owner can reduce their total cost of ownership by making available the EV's storage capacity (battery) for a period of time under clear conditions (e.g., minimum guaranteed level of charge of the vehicle and departure time with a sufficient level of charge)
- An operator can aggregate EV individual storage capacities in order to offer services based on a large storage capacity (Virtual Power Plant), with possibly specified grid connection points.
- An operator can actuate/manage the charge of EV currently connected to charging points in order to respond to TSO, DSO, or electricity supplier needs (peak shedding, ancillary services, voltage management at HV-MV power station perimeter or off-peak management, etc.)
- An operator can act as an active agent of the electric market making arbitrage (buying electricity at a cheap or even negative price and selling it when price goes up) or at least optimizing the charging price (knowing the purchase contract of any individual customer and its needs).
- An operator or "smart charging services provider" can also help optimize the sites' own consumption by helping it to avoid taking power from the grid when costs are higher and instead taking part of the load from the vehicle batteries (Vehicle-to-Building).

To do so, a huge amount of data needs to be shared among different stakeholders:

- Data of charging EVs: state of charge, time left until next use, minimum state of charge required by driver, etc.
- Charging point characteristics (location, power, V2G, etc.) and status (car connected or not, etc.)
- Data of network: voltage and load at HV-MV power stations level (real time and forecast), grid monitoring data, level of congestion of power stations
- Data of production: power produced (nuclear, hydro, coal, PV, solar, etc.)
- EV owner electricity purchase characteristics
- Market conditions and forecasts

And computing capacity must be available for real-time calculations of the "smart charging algorithms."

Problem Solved

This solution brings globally to the electric system the following services:

- Flexibility through load management, network equilibrium, peak shedding, and ancillary services
- Reduce total cost of ownership of EV by offering an EV owner the possibility to generate revenue from the electricity injected or off taken from the grid from the EV
- For electricity suppliers: real-time management of their portfolio in order to balance consumption and sourcing of electricity. For example, if renewable energy production is high at a moment of the day when there is not enough demand, it is better to charge the EV than to sell electricity at negative prices

Concrete Benefits
This use case brings the opportunity to share and to aggregate data from the different stakeholders of the EV value chain. These technical and personal data are the fuel of the services.

By offering this possibility, GAIA-X will enable the creation of new services around smart charging, flexibility, and V2G by offering a uniform, secure, and real-time access to critical data points.

GAIA-X would therefore play a significant role in the mainstream development of the EV ecosystem in Europe while contributing toward the optimization of the whole electricity process from production to consumption.

33.3.6.3 Electric Vehicle: CPO and DSO Investment and Planning

Solution
The solution has two parts:

- A technical part with the creation of a platform which can analyze and cross data
- A business part with the sale of studies, which use the data of the platform and personalize the result for the client

Problem Solved
The main goal of this use case is to increase the efficiency of the deployment of charging stations in Europe, thanks to the data convergence between the different operators. Finding the best place for a charging station requires having the following cross vision:

- The city's electrical plan to find out if an area can easily accommodate a fast-charging station without doing a lot of long and expensive network work (from DSOs)
- The city development plan to know how the city will evolve (from local collectivities)
- The flow of vehicles in the city, data coming from local collectivities, parking operators, and motorway companies. A special interest could be brought by data about where cars stop and where professional drivers are parking.

Vehicle flow could also be captured by passing through telephone operators, car manufacturers, or pure players such as Waze.

Concrete Benefits
A consolidated knowledge of the data of all the players would allow:

- CPOS to install the most profitable charging stations
- Local authorities to deploy charging stations where it is needed and at with an efficient price
- DSOs to plan in the medium and long term the evolution of their network

33.3.7 Description Use Cases: Local Energy Communities

33.3.7.1 Local Communities: Local Communities of Energy Setting Up and Decentralization

Solution
A local energy manager (LCE) coordinating the local energy efficiency by managing the renewable assets and energy infrastructures, by self-consuming and injecting the extra production to the distribution grids.

The business model will be consolidated by sharing economy principles. A variety of approaches to community ownership, including joint ventures, split ownership, and shared revenue could be explored in the project. Enabling clean energy ownership through community enterprise meets the twin objective of decarbonizing with cheap onshore renewables and winning the support of the local community. Paring the constraints and opportunities of energy and real estate sectors, the UC plans to develop a new design/build/own and operate (DBOO) offer for new and renovation districts co-developed with the clients and by sharing investments and profits.

Problem Solved
The LCE concept overcomes building energy seasonal peaks and provides interoperable business models and digital services based on trading of various energy carriers between the communities and the gross market and ancillary support to the distribution and transport networks. The essence of the business model is to have long-time ambitions and to share a part of the gains with the end users that become prosumers.

Concrete Benefits
- Up to 20% self-sufficiency, thanks to PV without storage
- More than 70% renewable rate, thanks to renewable production, heat recovery, DHC, etc
- More than 30% seasonal peak covered by the geothermal or hydrogen storage

33.3.7.2 Local Communities: Stadtwerke/Local Open Data for Business Models

Solution
The focus of this use case is the interdisciplinary challenge of the grid connection process for both customers and prosumers, which requires a large amount of both data and information and sub-processes from individual grid operators (typically DSO) as well as a large amount of data from various sources, especially from public geographic information systems (e.g., municipal data for the grid connection point and other data from residential registration offices). While it is possible for a resident

to initiate the grid connection process, it is yet not feasible to systematically incentivize, e.g., residents that have a heightened potential for utilizing renewable energy technologies due to a favorable combination of their location, supply contract, other already installed technologies/appliances, etc.

GAIA-X enables the central provision of needed data without media breaks and guarantees the quality assurance of the data provided. Thus, not only grid connection processes but also maintenance services as well as other integrated business cases relying on information about technology and usage pervasion can be accelerated.

Especially in the grid connection use case, data required in the grid connection process can be unified and standardized based on a common semantics and access platform with GAIA-X. In addition, the price of the required data for the customer/data receiver is set, and there is a role-based access control for various kinds of service levels (e.g., DER contractors, utility, etc.).

GAIA-X offers the required common infrastructure and enables the transparent and traceable transmission of public data from public administration to the grid operators. GAIA-X promotes the digitalization of public administration processes and enables the improvement and added value of data-intensive processes as well as the development of future business models in the energy industry.

Problem Solved

The provision of data depends on the local data management, which is used by the offices onsite, although there is no integrated common infrastructure. There is currently a technical and organizational challenge to transfer public data in a way that they are digitized as a public administration process and are available to the grid operators even though there is a lack of uniform semantics and access platform.

Further challenges are the identification of the required data sources, the digitalization of analog datasets, as well as the quality assurance and development of price models for additional services that can be offered. In addition, concepts for organizational governance and data ownership must be clarified, and data maintenance must be guaranteed.

Concrete Benefits

Open data has great economic potential: For Germany, the economic value of open data is estimated at around 12 billion euros annually. Positive effects of open data result from the use of data in the economy, its potential for innovation, increased transparency, and its potential for cost savings. In addition, open data plays an important role for future business models, especially in data-intensive processes such as those found in the energy industry.

33.3.8 Description Use Cases: Networks

33.3.8.1 Networks: Long-Term Scenarios

Solution

Creating long-term scenarios for energy transition requires running mathematical optimization models which are to be fed by numerous data.

By creation of long-term scenarios, we understand optimum energy mixes (installed capacity and storages, expansion of networks) from now to 2050. There exist today many different initiatives and projects within the European modelling community, but it appears that models and input data are not fully open and suffer a lack of transparency, which makes the studies' results difficult to understand and analyze. Most of the needed data are often not available, non-consistent (coming from various sources), and not transparent.

The openENTRANCE project (started 2019, running until 2023) proposes a first step for more transparency and more data available. This project has the following main deliverables:

- A database offering access to modelling results and inputs
- A nomenclature of data (i.e., an accurate description of all different variables which are available on the database)
- A series of open models (made open during the project), connected to the platform (i.e., able to be fed with data based on a common data format)
- Long-term energy scenarios computed from open data, with an open-source model
- A series of case studies (inputs and results being available on the database) focused on some topics of interest.

The objective of the GAIA-X use case would be to:

- Share and extend the data nomenclature
- Facilitate access to various sources of data (which would be made consistent with the nomenclature)
- Access to consistent data
- Run advanced functions:
  - Selection/upload of consistent data
  - Various treatments of data
  - Allow users to run the open models. Those models require IT systems which are often not available for the teams willing to run the models. GAIA-X could offer this facility (containerize models so that they can run on any IT system + rent high-performance computing resources)
  - Benchmark functions making it possible to re-run a specific study with the same model but different inputs/assumptions or the same inputs and a different model
  - Visualization of inputs and outputs + statistical analysis

Problem Solved
Decision-makers, stakeholders, and modelling teams, at European, national, and local levels, underline a lack of reliable, consistent, transparent data to build relevant long-term energy scenarios. Many studies are published by many different teams which the latter cannot easily interpret nor challenge, due to the lack of transparency both in input data, modelling assumptions and algorithms, and output data. Moreover, modelling team often cannot access to some of the data (including the fact that a high percentage of modelling resources is used to look for data or replace unavailable data) or only has access to low-quality data in their studies, and it is very difficult to assess the impact of this lack of quality in data on the results.

The approach proposed by openENTRANCE includes most of the advanced functions but relies on a database which cannot be scaled up, as it was designed for small volumes of data. Moreover, modelling teams do not have access to the adequate IT resources to be able to run big cases.

Concrete Benefits
- Improve the European ability to build relevant and feasible long-term energy scenarios
- Enhance the reliability of scenarios
- Enable benchmarking of scenarios
- Increase actor's confidence in published scenarios
- Enable different kinds of actors to run their own energy system modelling studies:
  - Actors with access to data but not to state-of-the-art model
  - Actors with access and experience on models but lacking consistent and high-quality data
  - Actors without access to HPC resources, etc.

33.3.8.2 Networks: OrtoPhotos

Solution
This use case aims to digitalize the mapping of existing networks to be more precise and have the possibility to create new services and serve new clients on this basis. It would be made possible by acquiring high-precision aerial images (from 20 to 5 cm/pixel) and then recover, transform, store, and disseminate the aerial images acquired.

This use case is part of the New Large Scale Mapping project from Enedis and is currently facing the following major issues:

- Human factor: Transfer between regional direction and supplier by physical exchange of hard drives
- Integration and injection into the IS: Data processing for customization
- Storage: Large volume of data (75 TB)
- Service exposition: Consumption/data visualization/orthophotos via APIs

Problem Solved
Today, this network mapping is done manually and not precise enough.

Concrete Benefits
- Improve the process to secure and track photo embedding
- Set up a scalable, reliable, robust, and autonomous orthophoto service
- Ultimately allow new services to be opened and new users to be served

33.3.8.3 Networks: Real-Time Data for EaaS Market Design Cross-Border

Solution
The residential end consumer wants to use its own produced energy in the way he wants. The national grids are slowly evolving to a more interconnected European grid enabling cross-border flows. This would allow the end consumer to charge his electric vehicle in another country when going on a vacation or business trip by using the excess of produced electricity, of course in consideration of congestion constraints.

Besides the necessary needed changes in market design and legislation, there is also a need for exchanging real-time data, first locally (between DSO, TSO, and other market players) and second cross border between different parties (DSO, suppliers, charging pole operators) so that excess electricity in real time can be used for charging the electric vehicle by a EaaS (Energy-as-a-Service) provider. GAIA-X can provide the technology for the required platforms to enable these data exchange.

- Consumer data: name, address, EAN, supplier, production
- Charging pole operator data: name, supplier, charging consumption, location
- Electric vehicle data: charging consumption, location

If there is an excess in production of the consumer, this value will be subtracted from the charged electricity leaving the consumer only to pay for the infrastructure of the charging pole operator and the remaining charged electricity that is not covered by his own production. Of course, this means that the produced electricity is not remunerated by the consumer's supplier since it is already consumed and that the transferred energy is not charged by the CPO's supplier. The local DSO will give a green light if this transaction does not cause additional congestion problems, while the TSO will monitor if sufficient cross-border capacity is available for the transaction.

A centralized platform for these EaaS providers based on the GAIA-X technology can further facilitate the coordination between these EaaS providers when the same data is used by different EaaS providers. For example, you want to share your excess of electricity with your neighbor while at the same time charging your EV somewhere else. While at the same time, the output of certain services can be used as input

for other services, for example, I want to know the average CO2 emissions of the electricity consumed by my EV for a better insurance.

Problem Solved
The traditional energy system stores many data like consumption, supplier information, and EAN about the consumer that is only shared between a limited number of parties within a country (DSO, TSO, suppliers). GAIA-X can provide a foundation to enable an easy access to this kind of data for new EaaS providers targeting a European client base.

Not only traditional energy players have many data about the consumers; also, new big players like OEM (Tesla, Viessmann) will have data about the end consumer. However, these data are available with different players, in different formats, but very useful for new EaaS providers. GAIA-X can provide guidelines that would foster standardization and data interoperability.

And last but not least, GAIA-X can create a more consumer-centric system. GAIA-X will not only enable the consumer to solve his pains in a climate-friendly way by use of EaaS (in this use case, charging his EV cross border with his solar production). GAIA-X can provide the foundation of putting data sharing more in the hands of the consumer instead of the big companies.

Concrete Benefits
- Increased consumer experience around energy transition with monetary and non-monetary value streams
- Lower entry barrier for new EaaS providers due to standardization and easier access to data

33.3.8.4 Networks: Congestion Management Through TSO-DSO Traffic Light

Solution
The increase of decentralized renewables and the electrification (of the mobility, of the heating, etc.) will increase the pressure on the distribution grid that has not necessary been initially built to withstand such energy flows and variation. At the same time, the decrease of central power plant to balance the grid coupled to the uptake of renewables will require to leverage more and more the decentralized demand side asset for the grid services including balancing purposes activated by TSO. However, the activation of such flexibility should take into account the local congestion impact and therefore be conditioned by an agreement from the local DS0.

For that reason, DSOs should share "traffic light" to TSO in order to signal congestion risk in all steps of congestion management: planning, forecasting of the need, market auction, and finally real-time activation.

As presented in the *TSO-DSO report: an integrated approach to active system management with the focus on TSO-DSO coordination in congestion management*

and balancing, the traffic light concept will be used to coordinate the flexibility activation.

- If the traffic light is green, the TSO can use the flexibility without restriction
- If the traffic light is orange, the DSO will ask Flexibility Service Providers (FSP) to resolve a planned congestion which could limit flexibility activation
- If the congestion is not resolve in right before the activation

The DSO could also incentivize the participation to the market for resolution of congestion through the exposition of the orange traffic light state.

In order to get more accurate response of the flexibility portfolio at local level (e.g., heat pump, electric car, etc.), the incentivization could be based on artificial model calculating best dynamic tariff price based on previous reaction. This could shorten the reaction and maximize value for the flexibility provider.

State is seen as out of the scope of this report.

Problem Solved

The main challenge to deal with the congestion management will be related to the exchange of data in a standard and near-real-time manner between TSO and DSO.

The use of standardized semantic should facilitate the translation of the traffic light between the DSO and the TSO and therefore enable near-real-time activation of flexibility respecting the constraint of the congestion.

Concrete Benefits

- The benefit of GAIA-X will be translated in the easiness to connect the different stakeholders Tso/DSO/OeM, etc. This should then relate to low cost of implementation
- In parallel, European standard for TSO-DSO interface could offer to market players easy business model development at European scale (e.g., flexibility aggregation, optimization, etc.)

33.3.8.5 Networks: Cross-TSO Failure or Labelled Image Database for Predictive Maintenance Training

Solution

As the share of renewables is increasing in the energy mix, the volatility and uncertainty are making the management of the system more and more complex. Therefore, the high availability of the grid is more important than ever cause each failure could lead to more congestion risk and therefore more system operational risk and ultimately affect market conditions in case of high-scale impact.

At the same time, the grid aging is increasing in Europe requiring more and more maintenance which will imply more risk of congestion in case of unplanned outage and increase of the maintenance cost. On the other side, maintaining an asset too early while it is still healthy is also leading to new cost that could be avoided. Therefore, it is very critical to get a deep understanding of the best moment for

maintaining the critical asset to maximize the availability and minimize the overall maintenance cost.

For that reason, multiple players start to test and implement first condition-based maintenance and then predictive maintenance.

The latter is based on an artificial intelligence model that will be trained based on previous failures data. These data could contain information about occurrences of specific failures (e.g., broken connector switches or explosions of transformer equipment) or some pictures taken generally by drones that are used to detect defect on pylons and lines (e.g., as rust).

However, as the grid is one of the most critical infrastructures, the maintenance is generally followed very closely, and the number of unplanned failures is limited. For that reason, it is often difficult to get proper data to train predictive maintenance or image recognition algorithm.

In parallel, the value of the AI algorithm is sitting more in the data used to train, and therefore by using multiple suppliers, companies are generally reinforcing indirectly the product of suppliers without benefiting from the potential commercial impact of it.

For that reason, it would be beneficial to put in common failures and/or pictures collected among the different system operators which could be used to train new algorithms. Later, the development of shared sovereign AI model for predictive maintenance could be envisaged.

However, to be efficient, such cross-company database would need to be coordinated among the different system operators, and the data semantics/labels should be harmonized among the system operator to make sure the predictive maintenance solution can effectively be trained. That requires a strong alignment between the system operators and strict common rules for data exchange.

Problem Solved

- *Unified dataset*: The standardization of data semantics and exchange through GAIA-X rules could highly facilitate the integration of new data as training set for predictive maintenance will be easier.
- *Efficiency and portability*: On top, the use of efficient sovereign cloud infrastructure will facilitate the treatment of this cross-company data for the asset management analytics services.
- *Integration*: The use of unified API will also facilitate the integration
- *Sovereignty*: Finally the use of a co-built dataset ensures a sovereignty of the data among the TSO (and potentially DSO) avoiding that analytics supplier keep the knowledge through training their model in blackbox and being the only one to benefit from the commercial added value of these dataset. The setup of a unified cross-country dataset could bring more power to negotiate with the analytics supplier or even to develop own analytics services on top of the unified dataset

Concrete Benefits

- Better-trained predictive maintenance
- Sovereignty of data and potentially analytics model leading to more weight in the negotiation with service provider for predictive maintenance or even the development of sovereign analytics services.

33.3.8.6 Networks: Energy Data-X

Solution

The energy data-X initiative applying for the BMWi GAIA-X Funding-competition is aiming to define the "data space energy" for Germany. In Germany, as it is in Europe, the energy infrastructure belongs to the most critical infrastructure in our economy. This also applies to the data created and used in the energy domain, for example, the misuse of which could cause great social and economic damage. For this reason, the data obtained may only be used and processed in a way that is relevant to the matter in hand. Data exchange in the current market model is highly complex and decentralized to meet the high data sovereignty and security requirements. This works in the classic energy market model where a few large producers guarantee the security of supply. In particular, the shift toward a decentralized and decarbonized energy market thereby requires major changes in order to meet the increased demands on data quality (e.g., highly granular or real-time smart meter data) and data transmission to ensure the security of supply. The currently used technologies as well as the given design of the market communication, where the (smart) meter values are transmitted every 15 minutes on the following day, will reach its limits in the near future. Furthermore, players such as e-mobility, prosumers, or mass-market heat pumps are not yet integrated into the current system but will take on a core function in the energy market of the future.

Energy data-X uses the GAIA-X Data Space infrastructure to define and test a possible solution by creating an energy data space where smart meter and sensor data are connected and made available for identified and authorized participants in the energy market. By "Sensor Data" we mean all data that is not collected via BSI-compliant measuring systems (e.g., in PV inverters, from uncalibrated meters of charging stations). The energy data space then builds the basis of a digital energy economy that efficiently reduces the complexity of the future renewable energy system. This goal will be reached by simple and fast provision and post processing of required information through a common energy data space. This enables a more efficient response to fluctuating feed-in; serves as a basis for new, cross-sector applications; and enables the development and provision of smart services by and for the large number of players in the energy market. Enabling data-sovereign use of cross-company data for concrete analyses and decisions is the basis for the use of artificial intelligence. Energy data-X will define the integration of selected master data as well as sensor data via an energy data space enabling efficient, sovereign, and transparent use of exchanged data for system operation as well as value-added

services ("smart services") and innovations. Furthermore, certain process can be developed and provided centralized through the energy data space.

Problem Solved

The current energy market model for data exchange is not able to meet the future requirements for data transmission, data quality, and data security. The energy-data x initiative will define a solution for this problem by using the GAIA-X infrastructure, thereby realizing economic potentials and synergies, enabling an increase in innovative capacity, clarification of the economic and technological benefits of GAIA-X for the German energy market, and realization of competitive advantages with the GAIA-X data infrastructure.

Furthermore, energy data-X addresses the need in the market to include "sensor data" in certain processes. This sensor data could then be used by MPOs or grid operators for the purpose of substitute value formation. The research focus lays on the following key elements:

1. *Applicability of energy data space to crosslink smart meter and sensor data:* Enable interoperability and portability of data and data-driven applications within and across sectors. Data spaces aim to create an ecosystem (including companies, organizations, and individuals) that generates new products, business models, and services based on more and more accessible data.
2. *Advanced smart energy data services which are based on the energy data space* ("Innovative Smart Applications"): Include data-based business solutions that use, for example, AI, the Internet of Things (IoT), or big data. The GAIA-X ecosystem is intended to provide a marketplace for data monetization and incentivize trusted data sharing across different actors in the ecosystem.

The result will be a demonstrator of technical maturity level 4–6 that shows the suitability of the Data Space approach for a future data exchange model in the future German energy market model.

Concrete Benefits

Establishing a large and comprehensive energy data space offering various data-driven use cases for players in the energy market.

- Consumer: Simplified accessibility of data and service offerings, improved data security and data sovereignty
- TSO: Improved system knowledge and system operations, e.g., replacement value creation, forecast
- DSO: automated settlement processes, process efficiency, e.g., supplier switching
- Service provider: Accessibility of data common data source for new and existing services
- OEM: Additional data value streams

33.3.9 Description Use Cases: Compliance and Traceability

33.3.9.1 Compliance and Traceability: Green Certifications

Solution

Within the GAIA-X Energy Space, our solution provides a certification service for green energy. The goal is to issue automated, timely, and government-approved sustainability certificates for energy, cross sector and along the entire value chain, covering certification of, e.g., electricity, gas hydrogen, green fuels, and other green goods.

The initial solution scope could be as follows:

- Defining and implementing governance structure for partner ecosystem
- Defining certificate standards for different energy sources
- Building decentral partner ecosystem and providing decentral digital identities for assets producing or consuming green energy (e.g., wind turbine, PV-plant, electrolyzer, methanol synthesis plant, steel plant)
- Connecting decentral digital identities with asset sensors
- Implementing proven standards into the certification management scheme
- Proving and confirming trustworthiness of implementation
- Issuing certificates related to asset sensor data in asset-specific wallets
- Transferring certificates between wallets (market participants) based on defined standards
- Devaluating certificates based on timely (e.g., 15 min) asset sensor data
- Defining and implementing automated payment scheme for certificate management

Problem Solved

Provide a working example. Prove that trustworthy and automated issuing and management of sustainability certificates work across multiple energy sources and industrial sectors as well as across EU countries. Demonstrate that decentral data ecosystems can connect different industries to prove the sustainability of produced goods with regard to energy.

Concrete Benefits

Get a demonstrator off the ground quickly. Solve all initial issues with applying GAIA-X concepts and architecture for a real-world sustainability certification application. Thereafter, use this demonstrator for training and onboarding of new stakeholders.

33.3.9.2 Compliance and Traceability: Infrastructure Data for New Business Models

Solution

Energy infrastructure belongs to the categories of most critical infrastructure in our economy. Critical infrastructures are those that ensure the supply of essential goods and services. They form the nerve cords of our modern society. Due to their importance for the interaction of all segments of society, these infrastructures require special protection. This also applies to data in the energy supply sector, for example, the misuse of which could cause great social and economic damage. For this reason, the data obtained may only be used and processed in a way that is relevant to the matter on hand.

At the same time, digitization has also arrived in the energy industry. It is driving forward the process of restructuring the energy system initiated with the energy turnaround in the form of more efficient processes and new business models. Several hundred energy start-ups are already supplying the energy turnaround in Germany with innovations such as virtual power plants, i.e., physical power plants that are interconnected via a platform. A "dedicated turnstile" now provides them with even more support. Business models that drive the energy turnaround can and should also be implemented using infrastructure data on a simple and secure way.

Problem Solved

The challenge is to reconcile the objectives and principles of using data from critical infrastructures and the need to use data for new business models. To this end, regulatory issues should also be addressed: Is all data classified as critical or is there any differentiation? If so, may only a certain type of, e.g., certified data centers be used for processing? Which market players would have access to the data they could use to transform value chains? How would data sovereignty finally be guaranteed?

In short, the provision and secure use of infrastructure data must be clearly regulated. One option could be to use only digital data twins, as proposed in the present use case. In this case, it would also be possible for energy suppliers or third parties to develop new business models based on this data. Some operators of critical information already provide such data twins which can be used by others for their own business models.

Concrete Benefits

This demonstrator focuses on the brought usability of infrastructure data without endangering the critical infrastructure itself. We see a huge retention to provide infrastructure data for further purposes due to the risk of being misused. By using GAIA-X and its components of identity and trust, sovereign data exchange and compliance third parties will become providers of distinguishing data products and software appliances.

33.3.9.3 Compliance and Traceability: Existing Standards Integration to GAIA-X

Solution

Gaia-X will be a non-domain-specific platform and will be used in various contexts and use cases. However, those domains have a specific vocabulary as well as corresponding basic communication stacks and data models. The use cases already envision common semantics as data models for communication or automation and control of, e.g., DER/DES, but there is a strong need to foster the use of domain-specific standards whenever possible and be open to incorporate and mediate to those legacy systems and technologies. From the UC point of view, one particular issue will be the convergence of both OT (operational technology) and IT (information technology) in the future smart grid. As data from IT and historical archives will be used for real-time optimization in OT (grid control contingency, etc.), there is the danger of format and semantics transformations which should be prevented. Given approaches like OPC which helps use domain-specific data models as address spaces, such an open approach shall be considered and documented for GAIA-X infrastructure services. Existing standards shall be screened for compatibility with the envisioned GAIA-X infrastructure and fallacies and gaps documented.

IT/OT.

Problem Solved

NIST identified 75 existing standards and 15 high-priority gaps in support of smart grid interoperability, in addition to cybersecurity issues, as a starting point for standards development and harmonization by standards setting and development organizations (SDOs like IEC and CEN/CENELEC). Sixteen Priority Action Programs (PAPs) have been initiated by NIST to address areas in which standards need revision or development to complete the standards framework according to their smart grid vision. In addition to the US perspective, the IEC Standardization Management Board (SMB) of Technical Committee (TC) 57 identified over 100 standards and standard parts in a strategic review of power system information exchange. Both of these studies concluded however that only a small number of standards lie at the core of smart grid interoperability, and they can be organized into a corresponding layered reference architecture described in IEC/TR 62357—the so-called SIA—Seamless Integration Architecture. The evolution of IEC/TR 62357 reflects the broadening scope of TC 57 in step with smart grid use cases from its original charter of "Power System Control and Associated Telecommunications" to "Power System Management and Associated Information Exchange." Generally this change reflects the shift in emphasis from lower-level interconnection automation OT protocols to abstract information models in the higher levels of the architecture in IT as the number of business functions needing to interoperate with PSAs has increased with smart grid evolution. The TC57 architecture generally follows the form of the GWAC Stack layers 1–7, as it ascends from standards concerned with communications relating to the connectivity of field devices through to information exchanges to support business processes and enterprise objectives.

This reference SOA blueprint shows how these standards relate to each other, require harmonization, and present the gaps where further standards development work is required. In general, all standards setting and development organizations advocate a collaborative approach to the development of open standards for the smart grid, with the reuse of existing standards as far as possible. GAIA-X will be a non-domain-specific platform and will be used in various contexts and use cases. However, those domains have a specific vocabulary as well as corresponding basic communication stacks and data models.

Concrete Benefits
The cost to fix a software defect varies according to how far along you are in the cycle, according to authors Pressman/Grady. One of the main costs drivers is the integration of components and system from various heterogeneous software or system vendors. Integrating based in standards lowers the amount of coordination and integration tests needed and fosters faster integration with less errors. As integration occurs in the later stages of a software project, costs of failed early interface semantics will cause a high maintenance and integration problem in the later stages.

33.3.9.4 Compliance and Traceability: Trusted HUB

Solution
This use case focuses on the design and implementation of a managed service to address the demand of privacy-preserving machine learning and multiparty computation in the GAIA-X ecosystem. Aggregating, combining, and analyzing data—including data analysis, machine learning (ML), artificial intelligence (AI), and decision-making—from different sources are becoming increasingly important almost in each and every domain from Energy 4.0 to Industry 4.0, mobility, and financial sector. At the same time, this process is often complicated as relevant data is often privacy sensitive and created and owned by many different data owners that do not tend to share with others since when exchanging their data with third parties, they may not only reveal their business secrets but also lose control over their data and for what it is used. Many data owners are aware of the technical and economic potential that is realized by analyzing their data, in particular, in combination with data obtained from other data owners. The process of combining and collaboratively analyzing different data sources results in new insights, better AI/ML models, better decision-making, as well as new/improved data-driven products and services. Considering this fact, this use case geared toward the integration of "Trusted Data Hub," a hardware-software privacy-preserving ML and multiparty computation solution, enabling several different parties to jointly analyze data, just as if they have a shared database without ever revealing those data. In other words, sensitive data sources held by multiple parties can be linked together in a secure manner, while parties gain no additional information about each other's sensitive data, except what can be learned from the output of data analysis.

Problem Solved
AI/ML is widely used in many areas of the energy domain, from energy fraud detection to theft detection, anomaly detection of energy consumption, energy demand prediction, demand response management, renewable energy forecasting, planned/unplanned disruptions forecasting in the power grid, outage detection and prediction, predictive/preventive equipment maintenance, and energy trading, among others. Machine learning and big data solutions enable energy and utility companies to optimize their resources, improve energy flows, manage the grid, schedule energy, and prevent mistakes. Unfortunately, the utility of AI/ML solutions is currently hindered by limited data availability for algorithm training and validation due to the absence of standardized data sharing/exchange as well as the requirements and concerns to protect the privacy of data owners and parties participating in the energy ecosystem. Although International Data Space (IDS) and GAIA-X Data Space can partially address the first issue, the development of new solutions to concurrently address the demands for privacy and ML utilization is a necessity. Trusted Data Hub aims to bridge this gap by providing a secure, privacy-preserving, and multiparty platform preventing data owner's privacy compromise and protecting data leakage.

Concrete Benefits
Privacy-preserving and multiparty computation play an important role in the data economy and the spark of innovative new business models. It bridges the gap between the utilization of AI/ML services and the privacy of data owners enabling transparent aggregation, trustworthy refining, and collaborative analysis of data sources to be provided as a new product/application on the energy markets.

33.3.10 Maturity Indication of the Data Space and Current Health Status

33.3.10.1 Addressing the Demand Side

The demand side is already well represented, and the ambition is to expand further:

- There is a broad representation of energy companies from Germany, Belgium, and France at the moment
- There are companies and use cases in all segments of the energy value chain: production (renewables, hydrogen and nuclear), networks (electricity, gas, heat, etc.), marketing, sales, and compliance

This must be completed by each hub/country if necessary.

33.3.10.2 Representing the Supply Side

These energy companies are already able to develop smart services (AI, IOT, machine learning, blockchain, etc.) and to integrate suppliers to develop these services.

Technology companies able to work within GAIA-X to provide infrastructures and data protection services are also stakeholders of the data space.

33.3.10.3 Creating a Sustainable Business Model in the Data Space

There is a very good equilibrium between demand and supply side in the data space energy, ensuring the achievement of the roadmap with sustainable business models.

33.3.10.4 Ramping up the TRL: From Prototype to Operation

The business model of each us case is or will be described by each use case working group.

A pipeline of use cases will be organized into four different steps:

1. Use cases project mapping
2. Use cases projects prioritization
3. Use cases implementation—phase 1, for use cases already close to implementation
4. Use cases Implementation—phase 2, for use cases needing R&D

Several use cases are already detailed, and a first prioritization has been done, identifying "quick wins" and "high-value use cases." New use cases will enter this pipeline in an iterative way, when data space's stakeholders identify new subjects or are joined by new participants and experts.

The business value and the potential for adoption and further scaling are key criteria of prioritization.

Before implementation, each use case will be further deepened, especially on two aspects:

- Consortium building if needed, with relationships between all parties definition and necessary resources identification
- Potential funding identification

The use cases pipeline will be developed in the "roadmap of the use cases" of the data space energy, defining priorities and timing of implementation of use cases (see Sect. 4.1).

33.3.10.5 Needed Component Certification from the GAIA-X Federation Services

All the main component of GAIA-X will be certified:

- Identity and trust
  - Federated identity management
  - Trust management
  - Federated access
- Federated catalogue
  - Self-description
  - Service governance
  - Monitoring and metering
- Sovereign data exchange
  - Policies and usage control
  - Usage control for data protection
  - Security concepts
- Compliance
  - Relation between service providers and consumers
  - Rights and obligations of participants
  - Onboarding and certification

33.4 Evolution of the Energy Data Space

33.4.1 Roadmap of the Evolution

The roadmap of the data space energy will include two roadmaps, which will be built and actualized in consistency:

- The roadmap of the use cases (based on the pipeline of use cases described above) – that will need to be detailed in 2021
- The technical roadmap of GAIA-X infrastructures and services labeled by GAIA-X, ensuring use cases development; there will be a common technical layer of infrastructures and services ensuring secured data exchange and interoperability.

33.4.2 Quick Wins (for 2021)

- Define the data space governance (how to enter the core team and the data space? What are the rights and duties of the core team and attendants?)

- Build a community of cross-border European actors
- Define the data space roadmaps, as described above
- Qualify first use cases and build some demonstrators for implementation in 2022.
- Lay the foundations around existing shared semantics

Use cases quick wins will be detailed in the soon-to-come roadmap.

33.4.3 Mid-Term Benefits (2022–2023) Building on Already Launched or Soon-to-Be-Launched Projects

Among the already or soon-to-be-launched projects, we identified:

- Renewable asset description model
- Renewable works—risk prevention
- Hydrogen station networks information sharing
- Nuclear equipment regulation
- Local communities of energy setting up and decentralization
- EV assets investments
- DSO network mapping
- Energy renovation
- "Green" certifications

List to be completed with the soon-to-come roadmap.

33.4.4 Long-Term Benefits Requiring Significant Investments on the 2021–2025 Period

- Contribute to the energy transition and carbon neutral economy, through new services and digitalization. Consistency between the energy data space, the Green Deal data space, or others (e.g., mobility) will be ensured
- Contribute to the development of European large cross-border infrastructures and data protection services
- Develop large cross-border energy services and actors in Europe, able to be developed all around the world
- Benefit to European citizens with better services, personal data protection, and job development in Europe.

Reference

1. C. Yang, Vyatkin, V., Mousavi, A., & Dubinin, V. (2014). On automatic generation of IEC61850/IEC61499 substation automation systems enabled by ontology. In IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society, Dallas, TX, USA, pp. 3577–3583. https://doi.org/10.1109/IECON.2014.7049030.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Glossary

Artificial intelligence Artificial intelligence is an area of computer science devoted to the demonstration of intelligent behavior in machines and especially on computers. Beside machine learning, knowledge-based systems and robotics are subfields of artificial intelligence.

Artificial neural network Artificial neural networks are statistical models inspired by biological brains. They consist of multiple layers of connected nodes, where the nodes contain a function and the connections carry weights to be optimized by deep learning algorithms. The input that reaches the function in a node through each connection is modified with its weight.

Big Data Big data is a term that describes the large volume of data—both structured and unstructured—that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

BSI C5 The cloud computing compliance criteria catalogue (C5) defines a baseline security level for cloud computing. The Federal Office for Information Security in Germany (BSI Germany) initially introduced C5 in 2016 and updated it in 2020.

Certification Authority (CA) An entity authorized to create, sign, issue, and revoke public key certificates.

Certification of trustworthy AI The development of a catalog and framework for auditing an AI application as trustworthy is an efforts in several recent projects. The goal is to establish auditable standards for conformance to ethical principles such as human control and autonomy, fairness, privacy and transparency as well as for security and safety.

Data mining, knowledge discovery Data mining and knowledge discovery used synonymously in practice to denote the entire process of detecting patterns, trends or relations in existing data sets with statistical methods and machine learning. CRISP-DM is an established process model for data mining and knowledge discovery.

© The Author(s) 2022

B. Otto et al. (eds.), *Designing Data Spaces*,
https://doi.org/10.1007/978-3-030-93975-5

Data science, data scientist Data science is an interdisciplinary area concerned with data mining and knowledge discovery. Data scientists may be specialists in a business domain, in data engineering, data analysis and machine learning, or application development.

DataOps DataOps (data operations) is an emerging discipline that brings together DevOps teams with data engineer and data scientist roles to provide the tools, processes and organizational structures to support the data-focused enterprise.

Deep Learning Deep learning is machine learning in artificial neural networks with many layers of connected nodes. It is responsible for the success of automated image, audio, text processing and generation in the last decade.

DevOps DevOps is a set of practices that works to automate and integrate the processes between software development and IT teams, so they can build, test, and release software faster and more reliably.

Digital Twin Computing Building a twin (representation) of a real person or thing in a digital world.

Dynamic Attribute Provisioning Service (DAPS) DAPS is an attribute server that issues OAuth2 access tokens to International Data Spaces connectors. The connectors need these to access the services and data of other connectors.

ECLASS ECLASS is a classification system for industrial products and services. The ECLASS product classes and properties allow a standardized information management for procurement, storage, production, and distribution activities in and between companies in multiple sectors, countries and languages.

FAIR dataset In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in Scientific Data. The authors intended to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets.

GDPR Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) https://eur-lex.europa.eu/eli/reg/2016/679/oj

Learning algorithm In computer science, an algorithm is an unambiguous sequence of instructions to solve a particular type of task. In machine learning, the task consists of optimizing a model using training data.

Machine learning Machine learning is an area of artificial intelligence concerned with optimizing statistical models with historic data. The optimization, called training, is done by a learning algorithm. The trained model can be applied to solve the type of task covered by the training data. Thus, machine learning is an alternative to explicit coding of a function.

Machine learning on the edge, distributed machine learning When local data must not be transmitted to a central server for machine learning, distributed machine learning allows models to be trained locally and transmit the models occasionally to the central server, where they are aggregated and retransmitted to the local nodes.

Message Queuing Telemetry Transport Protocol (MQTT) MQTT is an OASIS standard messaging protocol for the Internet of Things (IoT)

Metadata Data which describe one or more aspects of certain data. This could be, e.g. time and date, where, from whom were the data were created

MLOps Machine Learning Model Operationalization Management (MLOps), wants to provide an end-to-end machine learning development process to design, build and manage reproducible, testable, and evolvable ML-powered software.

Model (in machine learning) A model is an abstraction of reality. In machine learning, a learning algorithm optimizes a statistical model that generalizes the training data. The model can be applied to new data and computes a function. Decision trees, regression curves and artificial neural networks are popular types of models.

NAMUR NAMUR is an international user association of automation technology and digitalization in process industries.

NIS-Directive Directive (EU) 2016/1148 concerning measures for a high common level of security of network and information systems across the Union https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.194.01.0001.01.ENG&toc=OJ:L:2016:194:TOC

PID framework PID is a development methodology for C/C++ projects. It is used to homogenize projects development process of such projects and to automate operations to perform during their lifecycle.

Policy Decision Point (PDP) A system entity that makes authorization decisions for itself or for other system entities that request such decisions.

RDF Schema RDF Schema provides a data-modelling vocabulary for RDF data. RDF Schema is an extension of the basic RDF vocabulary.

Resource Description Framework (RDF) RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.

SALUS Healthcare Software

SCORVoc The Supply Chain Operation Reference (SCOR) is a cross-industry approach to lay the groundwork for more efficient and effective information exchange in supply networks. SCORVoc is an OWL vocabulary which fully formalizes the latest SCOR standard, while overcoming identified limitations of existing formalizations.

SOLID Solid (derived from "social linked data") is a proposed set of conventions and tools for building decentralized social applications based on Linked Data principles. Solid is modular and extensible and it relies as much as possible on existing W3C standards and protocols.

Spring Cloud Spring Cloud provides tools for developers to quickly build some of the common patterns in distributed systems

Web Ontology Language (OWL) The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. OWL is a computational

logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit. OWL documents, known as ontologies, can be published in the World Wide Web and may refer to or be referred from other OWL ontologies. OWL is part of the W3C's Semantic Web technology stack, which includes RDF, RDFS, SPARQL, etc.