Research paper

Keeping track of cleantech development using innovation clusters and member's website data: Evidence from leading energy clusters in Germany

Mahendra Singh *, Denilton Luiz Darold, Marian Klobasa, Andrea Zielinski, Rainer Frietsch

Fraunhofer Institute for Systems and Innovation Research, Breslauer Str. 48, 76139 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 31 January 2023
Received in revised form 2 May 2023
Accepted 13 July 2023
Available online xxxx

Keywords:
Innovation
Cleantech
Energy cluster
Web-scraping
Cluster member
Cluster policy

ABSTRACT

The main research question addressed in this work is how energy clusters can be evaluated and what general conclusions can be drawn out of their activities. Traditional innovation cluster analysis approaches chiefly rely on surveys, interviews, open publications, and patents—lack of using updated activities of innovation clusters. Therefore, preceding cluster analysis methodologies always lack of providing up-to-date information. In this sense, analyzing energy cluster activities is an obvious interest for policymakers, investors, companies, etc. Moreover, such assessment help to track the development of new technologies, participation of different actors in an innovation ecosystem, and emerging topics in the energy sector. This work presents the research outcomes on the leading energy innovation clusters in Germany. To this end, this paper exploits the publicly available website data from the clusters and member's web-pages to investigate their geographical distribution, key focus areas, cluster, and member activities. In the course of the project, a web-scraping tool has been developed to crawl the clusters and member's websites and scrape their text data. The tool performs systematic and guided web-scraping for searching a keyword presence on a particular web-page. In addition to this, data from commercially available company databases are used to complement the missing information from the website data. A total of 44 energy clusters along with 4524 members are taken into account in this study. The proposed methodology has shown that unstructured web-data is a valuable source for analyzing the clusters and their member's innovation activities. Results have also indicated that there is a strong correlation ($r = 0.85$) between Research and Development (R&D) expenditure and cluster count in individual federal states. The overall results have indicated that the majority of energy clusters are very specialized in certain topics, nevertheless, topics such as hydrogen, carbon, and bioenergy are getting notable attention from various stakeholders. Simultaneously, various cross-sectoral topics are also emerging due to the coupling between different sectors. Findings could help policymakers and federal innovation agencies to understand the ongoing progress in clean tech innovation activities. From the methodological point of the view, it provides an underlying ground to access the impact of cluster policies.

© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Energy innovation clusters have a significantly positive impact on the country’s energy-transition and economic development (Li and Ge, 2023). They are used as an innovation tool and are implemented to support energy transformation. Well-advanced cluster policies in Germany are playing a pivotal role in the country’s socio-economic growth (Lehmann and Menter, 2018; Heidenreich and Mattes, 2018). Indeed, Germany along with other EU member states (MS) has an ambitious goal to become climate neutral by 2050 (European Commission, 2018; Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety (BMUB), 2016). This could be only achieved by cutting greenhouse gas emissions, investing in green technology, and reducing reliance on fossil fuels. On the flip side, changing geopolitics on energy supplies are pressing public authorities, companies, civil society, and economic development agencies to come up with a long-term plan for supporting innovative cleantech solutions. Such a momentous transformation in the energy system demands rapid and scalable innovation to mitigate the challenges related to the security of supplies and the risk of being energy poor. Innovation cannot be matured in a stand-alone system. De-facto, it requires an ecosystem to reach the market and civil society. To put this into perspective, all major EU countries have

https://doi.org/10.1016/j.egyr.2023.07.026
2352-4847/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
implemented national and regional innovation cluster policies. At the EU level, various cluster-related policies are active to support the regional innovation clusters (Lucas et al., 2019). As an example, the European cluster collaboration platform (ECCP) of the leading cluster initiative funded by the EU Commission. ECCP facilitates collaboration among firms, economic actors, institutions, and public agencies to accelerate the green and digital regional economy. In a recent study, more than 1000 clusters and 1500 sector specialization regions are identified across EU 27 regions. A majority of clusters are active in digital, energy, health, food, transport, mobility and defense, etc (Susana et al., 2021).

Aligned with the EU’s cluster agenda, innovation clusters are highly represented in the German science, technology, and innovation (STI) policies. These clusters are active in various thematic topics and very well supported by the regional and federal governments.

In Germany, energy clusters represent regional energy and socio-technical transition and require a suitable approach to monitor regional advancement towards the energy transition goals and sustainable development (Hajighasemi et al., 2022). Research and innovation in the energy sector in Germany are on high priority. As a result, the Federal government spent around €1.22 billion on research, development of new technologies, and demonstration of pilot projects. A significant amount of this sum is invested in supporting cluster-related projects and innovation (Federal Ministry for Economic Affairs and Energy (BMWi), 2021; EnArgus Funding Database, 2022).

The members of innovation clusters play a crucial role in driving innovation and economic growth in the region. They bring their unique skills, knowledge, and resources to the cluster and collaborate with other members to develop new ideas, products, and services. The members can be individuals, businesses, research institutions, universities, and other organizations. Furthermore, innovation clusters provide members with access to a supportive ecosystem that facilitates collaboration, knowledge-sharing, and resource pooling (Fioravanti et al., 2021).

Along these dimensions, energy clusters are considered in this study to explore the current advancement in this particular sector. The available, cluster studies have been done following a cluster manager interview (Koecker, 2008; Benedikt Sedlmayr, 2019), targeted surveys, bibliometric analysis (Suominen et al., 2018), and public documents published by individual clusters (Ministerium für Wirtschaft, Bau und Tourismus Mecklenburg-Vorpommern, 2014). Lack of updated cluster and members activities together with complex innovation cluster research methodologies are the major shortcomings of the existing research. Cluster activities are very dynamic. They offer rapid collaboration, go-to-market strategies, and work on various cross-collaborative topics. From this perspective, website data can be more updated and consistent sources of cluster’s ongoing activities, events, projects, and publications. The integration of web-scraping technology with advanced data analytics holds great promise as a viable methodology and can serve as a fundamental approach in conducting such analyses.

Taking into account the discussion above the proposed work is designed to seek to answer the following questions:

- How the activities of energy-innovation clusters can be easily assessed and analyzed?
- Who are the major actors driving the innovation clusters in the energy sector and what role various participants are playing in establishing an innovation cluster?
- What are the current technology intervention, innovation actions, and topics emerging from the energy clusters, and how unstructured website data can be used in the innovation cluster-related analysis?

To provide answers to the above questions, the paper aims to analyze various activities of energy clusters within the German energy sector. The primary objective of this research is to assess the impact of diverse energy cluster policies by evaluating the activities of both clusters and their members. Additionally, the research emphasizes the aptness of unstructured data in the analysis of innovation.

The balance of the paper is organized as follows. The first section discusses the research motivation and highlights the current state-of-the-art. Subsequently, Section 3 outlines the theoretical background on the connection between existing cluster policies and lead-market in Germany. Further, Section 4 describes the methodology and data collection followed by results and discussion in Section 5. Section 6 concludes the findings and future perspectives of current work.

2. Research motivation and state of the research

In the literature, a plethora of definitions are available for innovation cluster (Granstrand and Holgersson, 2020). Indeed, the term cluster and network are interchangeably used with the term ecosystem. Nevertheless, clusters follow a system approach or network approach in which participating actors are linked and defined by certain rules. Agglomeration of actors is done in a specific geographical proximity (Autio and Thomas, 2014). In contrary to that, an ecosystem represents a complex and adaptive system. Such arrangements are non-linear in behavior and do not promise a similar outcome for the same input under varying conditions (Gobble, 2014).

For the sake of simplicity, we follow the cluster definition provided by the Federal Ministry for Economic Affairs and Climate Action (BMWK). This definition is derived from the Porter’s cluster theory (Porter, 2000) and is widely used in the context of German cluster policies.

Clusters are geographical concentrations of interconnected companies and institutions in related industries or technologies that complement each other through collaborative interaction and activities along one (or more) value chain(s). Key criteria for determining this are the regional proximity of the actors, a sufficient number and density of companies and research institutions, thematic-market proximity and at least national sales potential of the products and services. The proximity of the
actors along the value chain in terms of technology and content and their regional proximity leads to intensified innovation processes” (Cluster platform Germany, 2021).

Theory of innovation cluster analysis is very mature and widely studied by the various scholars (Zizka et al., 2021; Lucas et al., 2019). However, literature on the analyzing energy cluster is still very limited. There are different approaches been followed to analyze innovation clusters. A majority of approaches have followed qualitative surveys, interviews, and patent analysis to investigate cluster’s competitiveness, emerging innovations, composition, and internalization. For example, in Akhunzhanova et al. (2020) authors have analyzed the Silicon Saxony innovation cluster for cluster competitiveness analysis. A strategic cluster analysis tool developed by the World bank was used. It comprises four stages (Cluster definition, Cluster analysis, Institutional support, and Process control) and ten steps to perform cluster analysis (Shakya, 2009). Another study developed a visualization tool to discover clean energy innovation clusters (Lin et al., 2016) in the USA. The proposed tool uses i3-Connect company database (i3-Connect, 2023) and cluster publication, users can produce various dashboards as such as investors and technological maps. Likewise, studies done using patent data always trail behind because of significant delays in the patent application and advancement in innovation and technology development (Choi et al., 2021). In the context of the German studies on innovation clusters, they are very relevant and build on the feedback from the discussions with the cluster managers (Koecker, 2008; Benedikt Sedlmayr, 2019).

Over the last decade, web-data has grown extraordinarily and become an important indicator for monitoring innovation cluster activities.Clusters managers are using websites not only to promote their activities to the general public and businesses but also to promote the transparency of their activities, thus attracting new participants. However, the use of website data for innovation analysis is in its infancy, even though a number of scholarly publications are available in this realm. In Kinne and Axenbeck (2020) author has proposed a framework to analyze the innovation ecosystem using firm’s website data and argued the suitability of web-data for innovation mapping. Further in Katz and Cothey (2006) author has developed web indicators derived from the website data. These indicators help to establish the relative recolonization indicator (RRI) based on website content such as hyperlinks, text data and web documents, etc. Arora et al. (2020). Seen in this light, it is clear that website-data is an important source to monitor innovation related activities.

Following the discussion above, the current work aims to present a methodology for studying innovation clusters in the energy sector. However, before diving into methodological details, a theoretical discussion about the cluster policies in Germany has been provided in the next section.

3. Innovation cluster policies in Germany

In Germany, science technology, and innovation (STI) policies are governed by the individual Federal states (Bundesländer). All 16 states have their own agendas and policy instruments for developing regional innovation (Blümel, 2021). In this context, the formation of innovation clusters is one of the prime activities that spur open innovation and boost the regional economy in the federal states. To support, the cluster initiative different states already have implemented various cluster policies. These policies are designed to congregate diverse actors (e.g., SMEs, large companies, research institutes, and regional institutions) with similar interests in particular demographic proximity. As well as, enhance collaboration and communication between actors. Table 1 documents the major cluster policies and respective lead markets in the different federal states. These cluster policies are primarily driven by the triple-helix innovation model, which includes business, research, and government (Kalenchov and Shavina, 2018).

In view of the regional economy, infrastructures, and research activities, states have different lead markets to focus on. Indeed, these lead markets could be categorized into 16 wide categories and they are: Automotive, Life science, ICT, Energy and environment, Food, Medical research and healthcare, Chemical and Smart material, Agricultural, Media and creativity, Tourism, Maritime, Digitalization and Smart production, Industrial manufacturing, Mobility, Logistics, NanoMicro systems.

From the Table 1 it is notable that energy and related lead markets (e.g., renewable energy, environmental technology) are addressed by each federal state. Clusters are developing new energy-saving technologies, bio-based materials, energy-efficient production processes as well as sustainable products to reduce climate impact and energy consumption. In very few cases, states launch joint policies for their mutual benefit and collaboration at a higher level. One example of this, innoBB2025 is a joint policy framework between the states of Berlin and Brandenburg (Land Brandenburg and Berlin, 2019). Moreover, strategy for smart specialization (RIS3) is a European cluster policy framework that has been adopted by different federal states in Germany.

In addition to the state policies, federal ministers also initiate national-level cluster policies, funding programs, and cluster competition aligned with regional cluster policies. For example, Leading-Edge Cluster Competition (LECC) policy was initiated by the Federal Ministry of Education and Research (BMBF) and had a significant impact on regional economic development. Between the years 2007 to 2017, the program supported the strategic development of the cluster’s innovation strategy and provided 600 Million euros as financial support (Rothgang et al., 2017; BMBF, 2014).

Similarly, go-cluster initiative is a political excellence activity launched by the Federal Ministry for Economic Affairs and Climate Action (Bundesministerium für Wirtschaft und Klimageschutz, BMWK).

Clusters included in the go-cluster program are frontier of innovation and considered as most efficient and competitive national cluster management organizations. These cluster portfolios also demonstrate how highly competent Germany is in different industries and technological sectors (BMWK, 2019). Furthermore, “Zukunftscluster-Initiative” (Clusters4Future) is another national-level action endorsing highly innovative clusters in Germany. BMBF started this program in 2019 and lined up with the Federal’s Governments High-Tech strategy 2025 (The Federal Government of Germany, 2021). To be the part of Clusters4Future program clusters have to compete with each other and be evaluated against the strict criterion. Since the inception of the program, 14 clusters have been selected as Germany’s innovation networks of the future (Cluster4Future, 2022).

Cluster platform Germany (Cluster platform Germany, 2021) allows the sorting of innovative clusters based on their area of activity. For example, clusters dealing with innovation in the energy sector are listed under the ‘Energy technology’ category. Clusters active in cleantech energy technologies can be considered as an important indicator to measure the impact of cluster related policies in the energy sector.

4. Methodology and data

4.1. Description of data and database

To analyze the cluster’s activities collecting the relevant dataset is a crucial step in this work. As illustrated in Fig. 1 an initial
The present work has taken advantage of the following data and platform (Germany, 2021). After having the initial list of clusters comprehensively collects important details of major innovative purposes, we have considered ‘Cluster Platform Germany’, which step is to identify the active energy cluster in Germany. For that purpose, we have considered ‘Cluster platform Germany’, which comprehensively collects important details of major innovative clusters, active in different federal states of Germany (Cluster platform Germany, 2021). After having the initial list of clusters the present work has taken advantage of the following data and databases to analyze clusters and affiliated members.

- Collection of keywords: A list of energy-innovation related keywords is stored in a separate dictionary. These keywords are collected from the S3platform. This tool searches within European Structural and Investment Funds (ESIF) operational programmes and contains a long list of keywords coming from different EU actions, projects, and funding programs. Therefore, such a tool can be used to identify regions and countries with common or interdependent interests in different energy technologies (e.g. Biomass, Hydrogen, Storage, etc.) (ESIF-Energy-Smart Specialisation Platform, 2021). The search result can be downloaded in table format. The data is grouped by region, country, activity area, and sub-activity area.

Further, SciVal uses the Scopus database and provides an overview (in the form of keyword visualization) of the most recent research activities in the field of interest (SciVal, 2021).
- Unstructured web-data: This set of data is collected from the clusters and their member’s URLs. Together with full website, data headers like about us, product and services, projects, and members are prioritized in the web-scraping for each cluster and their members. As labeled, the data is unstructured in nature and collected using an internally developed web-scraper. The text collected is stored in its raw format and processed afterward, using natural language processing techniques to mine various aspects of cluster innovation activities. In addition, web-data is always considered to be updated and kept changing which allows users to track the most recent activities.
- Amadeus company database: Amadeus is a commercial company database and contains comprehensive information on around 21 million companies across 44 European country (Amadeus, 2021). Moreover, Amadeus database also

<table>
<thead>
<tr>
<th>Federal states</th>
<th>Lead markets</th>
<th>Major cluster policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schleswig-Holstein</td>
<td>Digital economy, Food industry, <strong>Renewable energy</strong>, Life science, Maritime, Tourism</td>
<td>Regional Innovation Strategy of Schleswig-Holstein – Roads to Smart Specialization</td>
</tr>
<tr>
<td>Mecklenburg-Western Pomerania</td>
<td>Automobile manufacturing, Aerospace, Logistics, Agriculture, Food industry, <strong>Renewable energy</strong>, Information technology, Life science, Maritime industry, Industrial manufacturing and materials</td>
<td>Regional Innovation Strategy for the state of Mecklenburg-West Pomerania</td>
</tr>
<tr>
<td>Lower Saxony</td>
<td>Mobility, <strong>Energy</strong>, Life science, Health economy, Bio-economy, and Production technology</td>
<td>Innovation strategy of the Free State of Saxony</td>
</tr>
<tr>
<td>Saxony-Anhalt</td>
<td><strong>Energy</strong>, Engineering &amp; Plant Construction, Resource Efficiency, Health &amp; Medicine, Mobility &amp; Logistics, Chemistry, Bioeconomy Food &amp; Agriculture</td>
<td>Regional Innovation Strategy Saxony-Anhalt (RIS3)</td>
</tr>
<tr>
<td>Brandenburg</td>
<td>Food Industry, <strong>Energy</strong>, Plastics and, Chemistry, Metal, Tourism</td>
<td>innoBB 2025,</td>
</tr>
<tr>
<td>Hessen</td>
<td>Digitalization, Life Sciences &amp; Bio-economy, Material technologies, Mobility &amp; Resource efficiency, <strong>Environmental Technologies</strong>, Smart production</td>
<td>Hessian Innovation Strategy 2020, Technology land Hessen</td>
</tr>
<tr>
<td>Thuringia</td>
<td>Industrial production and systems, Sustainable and smart mobility, logistics, Healthcare sector, <strong>Sustainable energy supply and resource management</strong>, ICT</td>
<td>Regional Innovation Strategies for Smart Specialization (RIS3) Thuringia</td>
</tr>
<tr>
<td>Saxony</td>
<td>Life science, Automobile, Microelectronics, ICT sector, <strong>Energy technology</strong>, Materials, Logistics, Mobility, Organic and Flexible Electronics, Environmental,</td>
<td>Saxony Trade &amp; Invest, Saxony Ministry for Economic Affairs (SMWA) innovation strategy</td>
</tr>
<tr>
<td>Berlin</td>
<td><strong>Energy technology</strong>, Healthcare industries, ICT, media and creative industries, Photonics, Transport, mobility and logistics</td>
<td>innoBB 2025</td>
</tr>
<tr>
<td>Hamburg</td>
<td>Logistics, Media and creative industries, <strong>Renewable energy</strong>, ITC, Health economy, Maritime</td>
<td>InnovationAllianz Hamburg</td>
</tr>
<tr>
<td>Bremen</td>
<td><strong>Sustainable management</strong>, <strong>Green hydrogen</strong> Resource efficiency, <strong>Wind energy</strong>, Connected and adaptive industry</td>
<td>Innovation strategy for the State of Bremen, 2030</td>
</tr>
</tbody>
</table>

Table 1: Regional cluster policies and lead market.
supports ‘batch search’ functionality that helps to search a
list of companies. Users can assign fields like company name
and country origin to avoid companies with similar names
incorporated in other EU countries. Amadeus also provides
categorizations of companies. Four categories namely: Very
large, large, medium, and small can be assigned to each
match found. This categorization is based on the company’s
operating revenue, total assets, and number of employees.

The purpose of this data collection is to explore the cluster’s main
focus, member composition, type, and emerging innovation from
the cluster’s ecosystem. In many cases, such information is not
easily available on even the cluster’s website in an organized
manner. For instance, clusters provide very limited information
about their member’s activities. The combination of the above
data sources covers a wide range of detailed knowledge about
the energy clusters and also complements missing information
in cluster analysis.

4.2. Methodology

Different methods have been proposed in the literature to
study the industrial innovation cluster. In Tvaronaviciene et al.
(2015) author has reviewed the methodologies used in the cluster
analysis. In total, seven methodologies are identified as core
methodology: case studies, analytical analysis, surveys, research,
equation, and others. Moreover, these methodologies cover vari-
ous industry sectors and are not only limited to the energy sector.
As mentioned before, these methodologies are constrained to
individual cluster case studies, qualitative research & analysis and
count on very controlled input such as cluster manager inter-
views, surveys, hypotheses, and economic models. The challenges
that lie in previous methodologies are insubstantial coverage of
clusters, limited data, and biased input from individuals. To over-
come these issues, the current work proposes a methodological
framework to analyze the innovation clusters with the help of
website data.

The proposed methodology is depicted in Fig. 1. Aforemen-
tioned, a list of energy clusters is obtained from the Cluster
platform Germany. In some instances, the cluster’s websites are
poorly designed, i.e., not following the web standards. In other,
their contents are just very limited and therefore negligible. Such
clusters are not considered further in the analysis. Following a
preliminary assessment, a total of 44 energy clusters are selected.

In the next step, after the cluster selection, certain cluster-specific
information such as cluster web URL (Uniform Resource Locator),
year of incorporation, and list of members are collected. In a few
cases, a detailed members list is not available on the Cluster plat-
form. Nevertheless, a complete list of members is collected from
the cluster website with the help of web-scraping. A total of 4524
members are closely examined. During the project, a Python-
based web-scraper is developed to scrape the web URLs. Python
package Scrapy (Scrapy web crawling and web scraping frame-
work, 2023) is used to build the web-scraper. The web-scraper
has two main features:

- Firstly, it can locate the presence of a specific keyword
  on multiple web-pages and raise a flag (0=absence and
  1=presence). For example, if the term ‘Biocathode’ is present
  in the searched URL then web-scaper raises flag= 1.
- Secondly, it can scrape the full content from the requested
  URLs and store it in a separate database.

For the text mining purpose, energy-innovation related key-
words are selected from two distinct sources. S3platform di-
vides energy-related activities into 16 main categories. They are;
Bioenergy, Carbon Capture Storage and Use (CCS and CCU), Co-
generation/Combined heat and Power (CHP), Energy efficiency,
Energy storage, Geothermal energy, Heating and cooling (H&C),
Hydrogen & fuel-cell, Hydropower, Ocean energy, Renewable en-
ergy, Smart Cities, Smart grid, Solar energy, Wind Power, Non-
technological actions related to Energy. Moreover, considering
the cluster activities, four additional categories: Electric mobility,
Digital/Digitalization, Energy transition, and Wood industry are
also added. Sub-categories are introduced in each category. For
example, the smart grid category includes smart grid, microgrid,
smart transmission system, interoperability, network manage-
ment, etc., and their synonyms as sub-categories. However, to
have an up-to-date list of subcategories a list of keywords from the
Sciwal platform is combined with S3platform data. These
keywords are initially supplied to web-scraper to identify their
presence on clusters and member’s URLs. Later the entire data
from these URLs are scraped and stored in the database for
detailed analysis purposes. Website data is mostly unstructured
and contains significant noise. Further, a python-based natural
language toolkit (NLTK) is used to clean the data (Natural Lan-
guage Toolkit in Python, 2023). It can easily remove punctuations,
extra spaces, stopwords, and perform case normalization along
with text tokenization. Furthermore, data validation is done using Python pandas DataFrames to perform format, consistency, uniqueness, and presence checks.

In the next step, different analyses such as co-relation analysis of topics, decile ranking, and frequency analysis are conducted by utilizing clean data. These analyses are explained in the result section in a detailed manner.

5. Results and discussion

This section highlights the key outcomes of the present work. These results are derived to accomplish the research questions mentioned in the introduction section. Fig. 2 shows the cluster and members count aggregated over the year of incorporation. In the year 2011, the adhesion reached its maximum with 6 energy clusters being added. Though there is a positive correlation between the clusters and members count, they are not strongly co-related. That means the participation of members not only depends on the number of clusters but also depends on the factors like how well clusters are developed, spillover effects of innovation clusters in the regions, attractiveness to local actors, etc. In the terms of regional distribution, clusters are distributed unevenly. In fact, regional factors such as economic support, specialization of companies, research institutes, and funding support are important factors that decide the number of innovation clusters in a particular region. In the state of Baden-Württemberg, 14 clusters are spotted in the cluster list, followed by the 6 clusters in Lower-Saxony and 5 in the Bavaria region.

Table 2 shows the correlation matrix between the Gross domestic product (GDP) of federal states, R&D expenditure and cluster count. The Pearson correlation coefficient (r) value is high (r = 0.85) between the cluster count and R&D expenditure, it indicates a strong link between R&D funding by federal states and cluster count. Furthermore, the correlation matrix also explains, the federal states with high GDP and R&D spending (for example, Baden-Württemberg, Lower-Saxony, and Bavaria) correspond to a large number of active innovation clusters in the energy field.

Each cluster has a thematic focus, and various focus areas are developed in the vicinity of such. Fig. 3 elaborates main thematic fields of energy clusters listed in this study. The size of each circle corresponds to the cluster count belonging to distinct thematic fields. Indeed, more generalized themes are renewable energy, energy technology, bioenergy, hydrogen & fuel-cell, environment, and electric mobility. In addition, a small number of clusters are specialized in very focused thematic domains such as energy storage, energy efficiency, smart city, etc.

Refereeing to the second question, we have analyzed the major actors driving the cleantech innovation clusters. To this end, the composition of different clusters has been analyzed. A list of members is prepared using the member’s details from cluster web-data and data from the Cluster platform.

Fig. 4 shows the fundamental stages of cluster formation and various member types identified in the member list. As mentioned such like innovation clusters in other fields, the key members in an energy cluster are firms (Large, Medium, Small, and Startups), research institutes, universities, public organizations, and financing actors. However, cluster composition in the energy sector is populated with several other actors (e.g., Training centers, Co-working facilities, Project management companies, etc.) and individuals. These actors play an eminent role in facilitating innovation culture. In addition, they also get benefits from the various funding projects and activities.

Following the theoretical analysis, it is clear that innovation clusters in Germany follow a top-down approach (Koschatzky et al., 2017; Koecker, 2008), in which the government takes control, federal ministries and state energy agencies set up the policy guidelines and research agenda for the particular subject or field. In the first stage, large companies, research, and educational institutes join the cluster initiative and get the capital access and regulatory framework to develop an innovative service or solution. At the next stage, startups, SMEs, banks, venture capitalists, and other financial actors join the cluster initiative. Startups get access to capital, patents, and working talents. On the other hand, financial actors get a reliable platform to invest their funds in an innovation action. The third stage brings different associations (e.g., Industry associations, Chamber of Commerce, etc.) alongside members supporting innovation acceleration. The presence of these actors strengthens the maturity of innovations. In the final stage, several other actors and individuals enter the ecosystem and participate in the cluster (see Fig. 4).
Fig. 3. Thematic focus of energy clusters in Germany (Bubble size is proportional to the cluster count).

Fig. 4. Energy cluster: Diverse set of stakeholders and stages.
The next step analyzed the unstructured dataset scraped from the cluster and member’s web-pages. Above-stated in the methodology section, a long list of keywords is used for text mining. Levenshtein distance (LD) function is used to measure the text similarity (NLTK metrics distance module, 2020). In natural language processing, Levenshtein distance is a useful measure to calculate approximate string matching implying upper and lower bounds for matches. Based on the matching score (LD > 95%) each keyword is counted and assigned a decile ranking. It split frequency counts into 10 different sub-counts. Due to the limitation of visualizing all the matched keywords, Fig. 5 shows the decile ranking of the top 72 keywords. Keywords with higher frequency count fall into the high decile rank category. For example, topics like hydrogen, renewable energy, electric vehicles, and solar energy are highly outlined in the cluster and members dataset. However, topics either with a relatively low focus that are emerging (e.g., Circular economy, Hydrogen storage, Digital transformation, etc.), belong to the lower decile rank category. In fact, a number of topics are overlapping and addressed by both i.e., clusters and their members. To narrow it down, several topics of interest are analyzed in greater depth. Fig. 6 presents the relative frequency plot for the selected main and sub-focus topics. Moreover, due to the visualization issue of overlapping frequencies, certain keywords are not very clear in Fig. 6. In general, keywords with higher relative frequency are at the top edge in Fig. 6. These keywords represent generic topics, for instance, bioenergy, hydrogen, wind energy, solar energy, etc. These topics are well-acknowledged and most frequently addressed by the energy clusters.

Contrastingly, keywords with low relative frequencies (e.g., Dye-sensitized solar, Hydrogen network, Wind breaks, etc.) fall at the bottom of the figure. In fact, the majority of these keywords
appear in the Scival database that collects top scholarly addressed topics. In this setting, such topics could be considered as most compelling research development and innovation activities in the energy sector. Thus, these keywords represent the most advanced topics addressed by the energy-innovation clusters taken into account in this study.

Additionally, a large number of focus topics are evolving due to the mutual interaction of different sectors. A few examples of alike topics are:

"Biohydrogen, Organic solar cells, Solar heating, Photobioreactor, Solar cooling, fuel-cell heating system, Solar thermal energy, Geothermal heat pump, Brine-water heat pump, Power-to-gas"

These topics are emerging in the framework of sector-coupling and have strong potential to attract various stakeholders active in very specialized activities. The presence of these topics amply demonstrates that energy clusters in Germany are not only limited to particular topics, but also cover various cross-domain topics. This makes clusters a more open and attractive platform for participants looking for further collaborative activities in their field of expertise. Energy clusters are diverse in the terms of member types and count. Fig. 7a shows the percentage distribution of member count over the total sum i.e. 4524. 35.2% of the members are identified as active and profit-making firms. The matching is done with the help of the Amadeus company database (Amadeus, 2021). Out of total matched firms, 38.2% identified as medium and 30.2% as small size firms, while the share of very large and large firms are 15.9% and 15.7% respectively (see Fig. 7b). Most of these companies are incorporated in Germany, however, companies founded in neighboring countries and other EU countries are also listed, which verifies the internalization and cross-broader activities of energy-innovation clusters in Germany.

On the other hand, institutes (3.9%), universities and colleges (3.6%), city and municipalities (2.5%) are other major participants in an energy cluster identified in the member data set. Though, energy utilities are one of the key pillars in the energy value chain. Nevertheless, their share in cluster-related activities are surprisingly low i.e., (0.5%). The category ‘other’ includes all actors that are neither matched in the company databases nor belong to the given categories. Having such diverse categories within the energy cluster represents a broader interest in energy-related topics.

Innovation clusters are frontier in developing new clean technologies and innovative business models. They encompass various actors of society and represent a territorial collaboration among them. Alongside, states and federal governments encourage new energy cluster policy frameworks to flourish the innovative activities in this sector. For that matter, this work highlighted major cluster policies in different federal states. In consideration of the energy sector, all states have active cluster initiatives. However, a few federal states have very effective and impactful cluster policies that result in a larger number of clusters in those particular states. Two such examples are Baden-Württemberg (Ministerium für wirtschaft arbeit und wohnungsbau baden-württemberg, 2018) and Lower-Saxony.

Contrary to cluster analysis methodology performed in Tvaronavičiene et al. (2015), the proposed methodology is data-driven and able to perform a relatively easy analysis of innovation clusters using their URL data. Similarly, studies done in Tvaronavičiene et al. (2015), Koecker (2008), Benedikt Sedlmayr (2019) do not emphasize the role of unstructured data and company databases in cluster analysis. In this regard, the methodology proposed in the current work has proven better use of a combination of different data sources.

Websites are a proven source for analyzing cluster activities. The main advantage of website data is its openness and volume, which helps deep dive into the analysis. Noting the outcome from website data analysis, it is very clear that clusters are focusing on a certain lead market. However, detailed analysis has reflected a diverse range of activities.

From the policy perspective, it is vital to identify the regions and clusters facilitating strategically important topics and provide them with special assistance and funding. Such examples are the federal state of Bremen and Lower Saxony exceptionally leading the development of wind and hydrogen initiatives. Furthermore, exchange between clusters is also important for a just and fair energy transition. It is recommended that cluster evaluation policies should consider collaboration among the regional clusters and internalization of activities as a key performance indicators (KPI). Similar to the European Cluster Collaboration Platform (ECCP) a national Cluster Collaboration Platform should be also established for promoting knowledge exchange between clusters. Furthermore, energy clusters are non-homogeneously distributed in Germany, therefore new policy measures are required to create a supportive framework for the regions with relatively lower participation.

6. Conclusions and future perspectives

Analyzing energy innovation clusters is a novel way to monitor the recent progress and emerging interests of the various actors. This project was undertaken to design a cluster analysis methodology and evaluate their footprint in developing cleantech-related
Thrive in internet data could play an important role in analyzing major activities in the energy sector. In addition, the work has provided a comprehensive overview of major cluster policies in different federal states of Germany. Although some federal states have limited activities related to energy clusters, the theoretical review showed that the energy sector is a lead market for all states. In addition to the above, the following conclusions can be drawn from the present study.

- In this investigation, the aim was to gain insight from the leading energy clusters in Germany. These insights are also an indicative measure of the success of regional energy policies and transition agendas.
- This project was undertaken to design a novel methodology that can analyze energy clusters by combining the different sources of structured and unstructured data. It could be extended and applied to evaluate clusters active in other regions.
- Though, energy clusters in Germany are organized in a top to bottom approach. In reality, the representation of actors is very diversified and includes actors from civil society, public bodies, universities, research institutes, companies, etc. Hence, it is evident that participation in the innovation cluster has become an obvious interest for different actors in society.
- This paper has argued that energy clusters and their member website data serve as an important source for measuring their activities. On the grounds of this, a tool for guided web-scraping is introduced. It uses a large set of keywords to mine the data from the selected web-links. This reduces the volume of unstructured data which is often a challenging job to handle.
- Taken together, results suggest that energy-innovation clusters are reasonably active in conventional topics (e.g., solar, wind, renewable energy), and to some extent they are also well engaged with nascent topics like biohydrogen, carbon nanofibers, dye-sensitized solar cell, etc. Furthermore, various cross-collaborative activities are also identified in cluster analysis.

The future prospect of this work is to develop an innovation indicator from the cluster and member web-data. Such an indicator could help to measure the impact of cluster policies, cleantech financing, and investment in emerging innovative technologies. Another direction of this work is aligned with the techno-economic analysis of innovative technologies and processes identified in the proposed analysis.

Methodologically speaking, matching and categorizing cluster members is a major limitation of this work. Name-matching via Python scripts on company database information often leads to various unmatched member categories. Hence, future work will focus on implementing an improved method of category matching using website text. In regarding the text content analysis, semantic similarity techniques could be applied, substantially expanding the reach of the current keyword-based method. Additionally, adding other data sources like social media, news articles could also provide more substance on clusters activity in particular topics. Another research stream of this work is to analyze the firm’s performance by investigating the impact of joining an innovation cluster through time-series analysis on company’s financial indicators.

CRediT authorship contribution statement

**Mahendra Singh:** Conceptualization, Methodology, Writing – original draft, Data analysis. **Denilton Luiz Darold:** Data curation, Methodology, Data collection, Software. **Marian Klobasa:** Editing, Investigation, Methodology. **Andrea Zielinski:** Validation, Review & editing. **Rainer Frietsch:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Mahendra Singh reports financial support was provided by Fraunhofer Institute for Systems and Innovation Research ISI. Mahendra Singh reports a relationship with Fraunhofer Institute for Systems and Innovation Research ISI that includes: employment.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to acknowledge Fraunhofer society and Innovation System Data Excellence Center (ISDEC) program for the financial support to the project ‘Data driven assessment of innovation clusters in the energy sector’.
Appendix. Web-scraper

A web-scraper was developed using python and the widely used scraping package Scrapy. In order to have a comprehensive data collection, a crawler is also implemented, allowing identify subpages and even subdomains, according to some criteria, to finally look through the keywords of interest. That allowed us to go deep into the cluster and member’s websites scanning for the keywords. Fig. A8 summarized the flow diagram with the basic steps of the web-scraper.

We have defined a threshold of five levels and fifty thousand pages per domain. Given the scale, multiprocessing techniques were used to allow parallel scraping and speed up the data collection. Lastly, the output is exported into CSV files containing metadata attributes like URL address, page title, and language, along with the text extracted. The text is split by a pre-defined separator, and only sentences larger than 30 characters are included in order to avoid labels and text related to the page layout but without meaningful relevance. Regarding legal affairs, the tool respects the rules present in robots.txt file, the de-facto standard.

References


