Ein wissensbasierter Terminologie-Dienst zur Unterstützung von Konzept-getriebenen eHealth-Prozessen

Andreas Billig, Franziska Krebs

CC E-HEALTH
Fraunhofer FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin
andreas.billig@fokus.fraunhofer.de
franziska.krebs@fokus.fraunhofer.de

Abstract: Ausgehend von dem CTS2-Standard zur Verwaltung und Abfrage von Terminologien wurde der Terminologie-Dienst CTS2-LE auf Grundlage der semantischen Techniken und Kalküle des W3C aufgebaut, womit Terminologien als Teil einer Wissensbasis repräsentiert werden können. Dieser Ansatz erlaubt eine flexible Integration von Akteurs-orientierten Artefakten wie Patienten- und Fallakten sowie die semantische Bearbeitung mit Hilfe von Nutzer-definierten Regeln und komplexen SPARQL-Anfragen über den vernetzten Artefakten und Terminologien. Die Integration unterstützt damit insbesondere diejenigen Prozesse, für die der Zugriff auf die Konzepte medizinischer Terminologien notwendig ist. Wesentliche e-Health-relevante Terminologien wurden mit Hilfe von spezialisierten Modell-Korrespondenzen auf den Terminologie-Dienst CTS2-LE abgebildet.

1 Einleitung

Eine der wesentlichen Voraussetzungen für die rechnergestützte semantische Interoperabilität ist die Nutzung von kontrollierten Vokabularien (im Folgenden auch Terminologien genannt) von allen beteiligten Akteuren im e-Health-Sektor. Das technologische Rückgrat für den praktischen Einsatz bilden standardisierte Dienste zur Verwaltung und Abfrage dieser Terminologien. Einer der führenden Standards bzgl. des Informationsmodells und der Schnittstelle von Terminologie-Diensten ist der HL7/OMG-Standard Common Terminology Services Release 2 (kurz CTS2) [OMG14]. Die theoretische Grundlage für Terminologien ist gegeben durch formallogisch fundierte Sprachen zur Beschreibung von Konzepten, deren Bezeichnungen (Terme) und Beziehungen untereinander, nämlich Ontologien [Gr93].

Zum Zwecke der Interoperabilität ist es dabei von Bedeutung, allgemeingültige und international anerkannte Standards zur Repräsentation von Ontologien und zur Vernetzung von Artefakten (in unserem Fall terminologisch oder Akteurs-orientiert) zu

verwenden. Diese Standards sind mit den Semantic Web Standards des W3C gegeben (RDF, SPARQL, OWL, Linked Data) [W3C14].

Terminologie-Dienste unterstützen maßgeblich diejenigen Prozesse des elektronischen Gesundheitswesens, die eine korrekte Interpretation und die Interoperabilität der in ihnen vorkommenden Konzepte und Bezeichnungen voraussetzen. An diesen Prozessen sind als Akteure sowohl Patienten, als auch Ärzte und weitere Rollen beteiligt. Die Unterstützung kann dabei sehr vielfältig sein. Sie reicht von einfachen Grundfunktionalitäten wie die Validation bzgl. der Korrektheit medizinischer Codes bis hin zu komplexen Informationsableitungen.

Der am Fraunhofer FOKUS entwickelte Terminologie-Dienst CTS2-LE (CTS2 - linked data edition), der vorwiegend am CTS2-Standard orientiert ist, basiert auf diesen semantischen Techniken und Kalkülen, womit eine effiziente und CTS2-Modell-nahe Wissensbasis zur Verwaltung, Abfrage und Bearbeitung von Terminologien, Konzepten und Konzeptbeziehungen geschaffen wird. Ein spezieller, text-orientierter Suchdienst ergänzt die Wissensbasis um gezielte Terminologie-übergreifende Anfragemöglichkeiten und um eine Verlinkung mehrerer Sprachversionen derselben Terminologie.

Der Terminologie-Dienst CTS2-LE hält wesentliche e-Health relevante Terminologien vor. Zu diesen gehören spezialisierte Klassifikationen zu Krankheitsbildern (ICD), Prozeduren (OPS), Einheiten (UCUM) und Ordnungssysteme zur Beschreibung von Untersuchungs- und Testergebnissen (LOINC). Darüber hinaus enthält der Dienst die Medical Subject Headings (MeSH) als eher allgemein gefasstes kontrolliertes Vokabular.

Der web-basierte Terminologie-Navigator erlaubt ein komfortables Erforschen der medizinischen Codes durch alle an den Prozessen beteiligten Akteure. Zu den Volltext-basierten Suchdiensten gehören eine Freitextsuche über eine indexierte Wissensmenge sowie eine Ähnlichkeitssuche gemäß dem Vector Space Model. Ferner erlaubt eine parametrisierbare Anfragekomponente unter Verwendung von SPARQL die Formulierung von Anfragen, welche die gesamte ontologische Struktur einbeziehen.

2 CTS2 & Semantische Technologien

Die technologische Basis zur Verwaltung und Abfrage von Terminologien ist durch den Standard CTS2 gegeben. CTS2 spezifiziert sowohl die Informationsstruktur, als auch die konkreten Dienst-Schnittstellen. Solch ein Dienst wird vor allem im Zusammenhang mit Akteurs-orientierten Artefakten wie Patienten- und Fallakten notwendig, da gerade diese Dokumente über eine Vielzahl von kodierten Attributen verfügen, d.h. Attribute, deren Wertemengen Konzepte relevanter Terminologien beinhalten. Im Kontext von HL7 wird dieser Wertebereich auch als Semantic Type bezeichnet [HL08].

Die wesentliche Aufgabe des Dienstes ist es daher, folgende Funktionalitäten zu ermöglichen:

- Rechner-Unterstützung bei der initialen Zuordnung von kodierten Attributen aus e-Health-Dokumenten zu Konzepten (wie z.B. die Zuordnung von Diagnoseattributen aus Patienten- und Fallakten zu ICD)
- Validierung von Attribut-Konzept-Zuordnungen in bestehenden e-Health-Dokumenten
- Abfrage und Anwendung einer Semantik-erhaltenden Abbildung von Konzepten unterschiedlicher Terminologien (z.B. von ICD nach MeSH)

Die CTS2-Informationsstruktur umfasst alle relevanten Informationseinheiten zur Repräsentation von Terminologien, nämlich

- Metadaten zur Terminologie (CTS2-Codesystem)
- Konzepte einer Terminologie mit ihren eindeutigen Codes, ihren Termen (designations) und ihren Assoziationen zu anderen Konzepten
- Value Sets, d.h. Mengen von Konzepten, die kodierten Attributen zugewiesen werden können
- Abbildungen zwischen Terminologien

Die theoretische Grundlage der vom CTS2 verwalteten Informationseinheiten ist durch formallogisch fundierte Sprachen gegeben, welche an die Besonderheiten des WWW als Grundlage des heutigen Informationsaustauschs adaptiert wurden [Gr93] [SS09]. Dabei kann die Basissprache RDF als einfache Repräsentationssprache für semantische Netze bzw. Frames angesehen werden. Der hier gewählte Ansatz verwendet die Techniken und Kalküle des Semantic Web zur Realisierung des CTS2-Dienstes, welcher - neben der konzeptionellen Adäquatheit aufgrund der gemeinsamen theoretischen Grundlage - die folgenden Vorteile einer weitergehenden semantischen Bearbeitung bietet:

- intrinsisches logisches Schließen
- explizites logisches Schließen via Nutzer-definierter Regeln
- komplexe Anfragen über SPARQL
- Verwaltung einer großen Menge von vernetzten Artefakten (Linked Data)

3 System- & Informationsarchitektur

Die Hauptkomponente der CTS2-LE-Architektur stellt die semantischen Funktionalitäten zur Verfügung und wurde, entsprechend unseres Ansatzes einer RDF-orientierten Wissensbasis, mit Hilfe des Jena-Frameworks [ApJ14] realisiert. Hierbei wurde die Persistenzkomponente Jena-TDB verwendet (s. Abbildung 1), welche jedoch austauschbar mit allen die Jena-Schnittstelle erfüllenden Quad-Stores (z.B. Virtuoso [OL14]) ist.

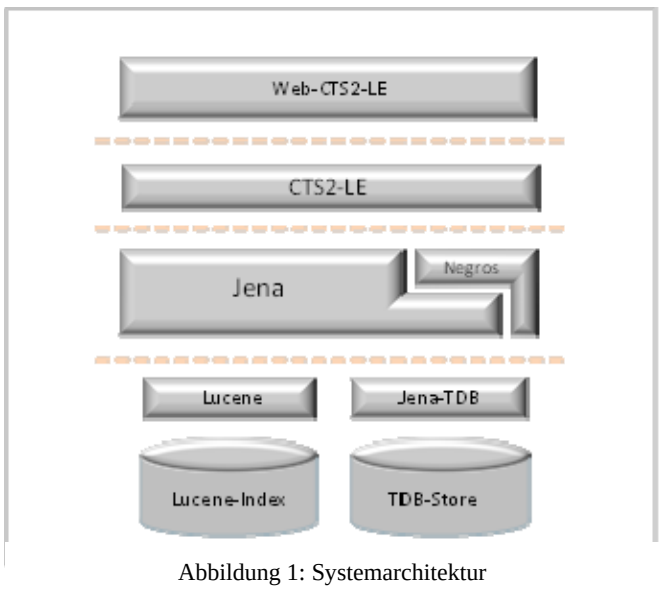

Abbildung 1: Systemarchitektur

Die Persistenzschicht verfügt als zweite Komponente über einen hochperformanten Suchindex (Lucene [ApL14])[1] , der eine Freitextsuche über der indexierten Wissensbasis sowie eine Ähnlichkeitssuche über den Konzepten gemäß dem Vector Space Model ermöglicht. Die über der Persistenzschicht angesiedelte Schicht für die reine Wissensbasis-Funktionalität und die wiederum darüber liegende Schicht der CTS2-LE-Anwendungslogik bestehen aus den Jena-Modulen sowie wesentlichen Erweiterungen zur Verwaltung von CTS2-Informationseinheiten und deren Bearbeitung.

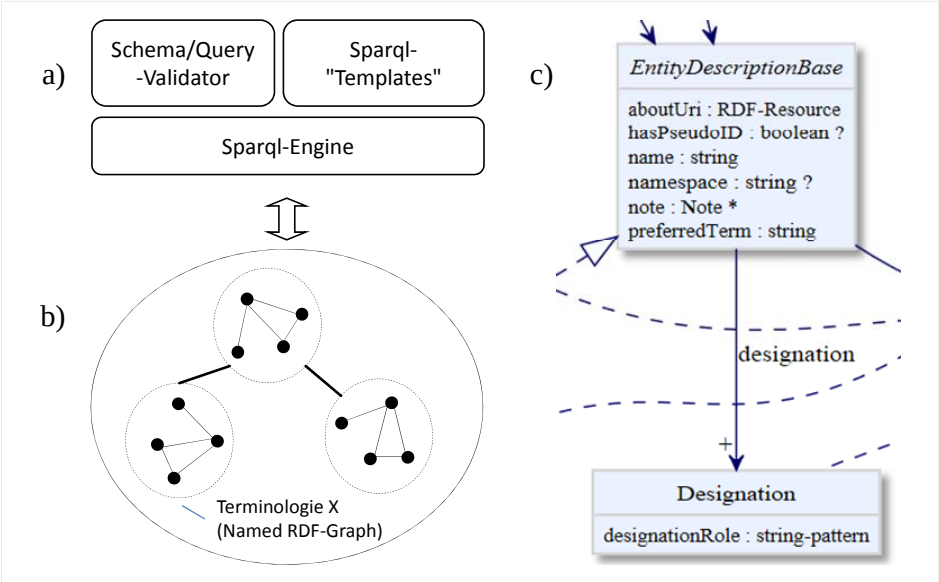

Abbildung 2 Erweiterte Funktionalität

Die CTS2-Informationseinheiten, also Terminologien, Konzepte etc. werden als RDF-Tripel repräsentiert, wobei die Partitionierung der Gesamt-Tripelmenge über jeweils einen benannten RDF-Graphen pro Terminologie erfolgt (s. Abbildung 2.b). Das Schema für die hiermit aufgebaute Wissensbasis wurde aus den Datenstrukturen des CTS2-Standards abgeleitet. Die Definition dieses Schemas erfolgte mit einer spezialisierten Schemasprache, genannt RDF-Signaturen (s. Abbildung 2.c), welche sich an F-Logic [KLW91], einer logischen Fundierung für Frame-Sprachen, orientiert und die in unserem Fall gebotene Closed World Assumption berücksichtigt [Mi82].

Über diese Schemasprache ist es möglich, beliebige weitere Schemata zu definieren, in unserem Kontext beispielsweise für Akteurs-orientierte Artefakte wie Patienten- und Fallakten. CTS2-LE dient hierbei als Integrationskomponente, da jedes neu

[1] Die in Abschnitt 2 beschriebene Attribut-Validierung von Dokumenten lässt sich somit in einem Bereich von unter 50 Millisekunden pro durchschnittlich attribuiertem Dokument realisieren.

hinzukommende Artefakt mit Teilen einer Terminologie zum Zwecke der konzeptuellen Einordnung in Beziehung gesetzt werden kann. Des Weiteren lassen sich logische Regeln zur Ableitung neuer Informationsverknüpfungen erstellen, so beispielsweise die Zuordnung eines Patienten zu Konzepten, die nur in indirektem Zusammenhang zu seiner medizinischen Falldokumentation stehen (z.B. zu denjenigen Medikamenten, welche die gleichen Nebenwirkungen haben wie die von ihm eingenommenen Medikamente).

Zu den zusätzlichen Erweiterungen der Wissensbasis-Grundfunktionalität (s. Abbildung 2.a) gehören

- eine SPARQL-Template-Engine zur Formulierung und Auswertung von parametrisierbaren Anfragen
- ein Validator für Anfragen hinsichtlich der CTS2-RDF-Signaturen

Schließlich zeigt Abbildung 1 die oberste Schicht Web-CTS2-LE, in der alle Komponenten, die den Web-Navigator realisieren, angesiedelt sind. Der Navigator erlaubt ein komfortables Erforschen der medizinischen Codes, eine Cross-Terminologie-Suche und eine Konzept-Ähnlichkeitssuche. Ferner erlaubt eine Anfragekomponente unter Verwendung von SPARQL die Formulierung von Anfragen, welche die gesamte ontologische Struktur einbeziehen.

4 Abbildung von Terminologien

4.1 Importprozess

Zu Beginn des Importprozesses steht die inhaltliche und strukturelle Analyse eines in den Dienst zu integrierenden Codesystems. Zu diesem Zwecke werden die zur Verfügung gestellten Dokumentationen und Quell-Dateien zum Codesystem recherchiert. Im Ergebnis entsteht ein

I. konzeptionelles Modell, welches logische und strukturelle Zusammenhänge einer Terminologie beschreibt.

Grundlegend beschreibt ein solches Modell die in einer Terminologie enthaltenen Konzepte und das relationale Geflecht. Darüber hinaus legt es taxonomische Besonderheiten wie Polyhierarchien offen. Auf der Basis des Modells erfolgt die Definition einer

II. Abbildungsvorschrift, welche Korrespondenzen zwischen dem Modell der Terminologie und dem durch den CTS2-Standard gegebenen Datenmodell spezifiziert.

Dieses Mapping wird im letzten Schritt in Form eines

III. Terminologie-Loaders

implementiert.

Nach diesem Vorgehen wurden sowohl Klassifikationen von Krankheitsbildern (ICD[2]), Prozeduren (OPS[3]) und Einheiten (UCUM[4]), als auch Ordnungssysteme zur Codierung von Labor- und Testergebnissen (LOINC[5]) importiert. Der Bestand der durch den CTS2-Dienst zugreifbaren Wissensbasis umfasst neben diesen – besonders im klinischen Alltag relevanten – Standard-Codesystemen auch Vokabularien, welche sich für Aufgaben des Dokumentenmanagements eignen (z.B. allgemeine medizinische Terminologien).

Die Implementierung (Schritt III) entfällt für terminologische Inhalte, welche bereits CTS2-konform im RDF-Format oder im ClaML-Format (ein speziell für medizinische Klassifikationen geeignetes XML-Datenformat [ISO14]) vorliegen. Ferner kann ein automatisierter Import für FHIR[6] Value Sets erfolgen, wenn diese in Form von FHIR-Value-Set Definition Files existieren.

Im Folgenden wird am Beispiel von Terminologien aus der Domäne des Infektionsschutzes die Integration in den Dienst beschrieben. Es werden die wesentlichen Konzepte und Konzeptstrukturen identifiziert und die Abbildung auf die durch den CTS2-Standard spezifizierten Entitäten gezeigt.

4.2 Abbildung von Terminologien aus dem Bereich des Infektionsschutzes

Die Begriffssysteme wurden vom Robert Koch-Institut entworfen und in Form von XML-Dateien bereitgestellt.

I konzeptionelles Modell der Terminologien

Die Terminologien umfassen meldepflichtige Krankheiten, deren Krankheitsformen, Erreger, Nachweismethoden, Symptome und Impfstoffe. Im Zusammenspiel ergeben sie ein ganzheitliches Informationsnetz, in dessen Zentrum eine Krankheit durch

i. die Attribute Krankheitskürzel, Krankheitsbezeichnung, Erregerbezeichnung, Kodierung der Krankheit gemäß ICD-10-Codesystem und

ii. die Relationen zu Erregerlisten, Krankheitsformen, Impfstoffen und Bundesländern, in denen die Krankheit meldepflichtig ist,

beschrieben wird.

Innerhalb der Konzeptmengen der Krankheiten, Krankheitsformen und Impfstoffe existieren keine Hierarchie-Relationen. Die Mengen der Erreger, Nachweismethoden und Symptome werden durch Subsumtionsbeziehungen strukturiert.

---

[2] Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme
[3] Operationen- und Prozedurenschlüssel
[4] Unified Code for Units of Measure
[5] Logical Observation Identifiers Names and Codes
[6] HL7/FHIR: http://www.hl7.org/implement/standards/fhir/ (zuletzt besucht: 8.9.2014)

II Abbildungsvorschrift

CTS2 spezifiziert zur Repräsentation von Terminologien ein gemeinsames strukturelles Modell, welches Technologie-unabhängige Elemente beinhaltet. Tabelle 1 zeigt die Teilmenge, die benötigt wird, um die wesentlichen Entitäten wie beispielsweise Konzepte, deren Bezeichnungen und Relationen zwischen Konzepten abzubilden.

Tabelle 1 Terminologische Elemente und deren Repräsentation gemäß CTS2

Element einer Terminologie	**Repräsentation durch CTS2-Standard**	**Attribute (Attributbezeichnung in Klammern)**
Konzept	*ClassDescription*	Identifikator (*name*) Namensraumangabe (*namespace*) Vorzugsbegriff (*preferredTerm*) Systemobjekt (*about*)
Relation	*PredicateDescription*	Identifikator (*name*) Namensraumangabe (*namespace*) Vorzugsbegriff (*preferredTerm*) Systemobjekt (*about*)
Bezeichner	*Designation*	Bezeichnung (*value*) Verwendungshinweise (mittels *designationRole* Typisierung als PREFERRED, ALTERNATIVE oder HIDDEN)
Erklärungen	*Definition*	Erklärung (*value*) Verwendungshinweise (mittels *definitionRole* Typisierung als INFORMATIVE oder NORMATIVE)
Assoziation (zur Verbindung zweier Konzepte mit Hilfe einer Relation)	*Association*	Subjekt (*subject*) Relation (*predicate*) Objekt (*target*) Ableitung (*derivation* zur Kennzeichnung einer hinzugefügten (*ASSERTED*) oder geschlussfolgerten (*INFERRED*) Aussage)

Jedes Konzept wird so mittels einer *ClassDescription* eindeutig identifiziert und durch den Einsatz von *Designations* mit unterschiedlich typisierten Bezeichnern versehen. Die Abbildung 3 zeigt die in CTS2-LE umgesetzte Repräsentation einer Krankheit.

Gemäß einem auf semantischen Technologien aufgebauten Wissensnetz beschreibt eine *ClassDescription* eine RDF-Resource (*about*). Die Krankheitsbezeichnung wird angesehen als das primäre Beschreibungsmittel und wird daher dem CTS2-Attribut *preferredTerm*, als auch einer als PREFERRED typisierten *Designation* zugeordnet. Eine eindeutige Identifizierung ergibt sich durch die Verwendung des Krankheitskürzels

als *name*, in Verbindung mit einer Namensraumangabe (*namespace*). Die Angabe des ICD-Codes erfüllt den Charakter einer als *NORMATIVE* deklarierten *Definition*, der Name des Erregers wird als eine alternative Bezeichnung mittels einer *Designation* festgelegt.

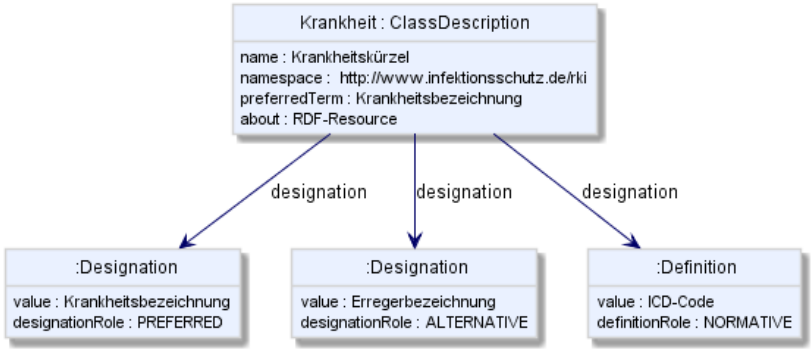

Abbildung 3 CTS2-konforme Abbildung einer Krankheit mit Attribut-Notation <<CTS2-Attributname : Attributname der Krankheit>>[7]

Zur Abbildung der hierarchischen Strukturen und der Konzeptbeziehungen erfolgt die Definition mehrerer *PredicateDescriptions* (beispielsweise subClassOf, hatKrankheitsform, hatSymptom, meldepflichtigIn), welche in einer *Association* das Bindeglied zwischen 2 Konzepten (*ClassDescriptions*) darstellen.

5. Zusammenfassung

CTS2-LE gilt für Arzt und Patient gleichermaßen als Informationsmedium, in welchem eindeutig identifizierbare Konzepte in einem vernetzten Begriffsraum nach internationalen Standards verwaltet werden. In Einrichtungen des Gesundheitswesens unterstützt ein solcher Terminologie-Dienst medizinische Prozesse und Entscheidungsträger. Beispielsweise ließe sich durch die vorherig dargestellte Integration eine gemäß ICD kodierte Krankheit direkt als meldepflichtig erkennen. Ein weiterer Benefit ergibt sich durch die Möglichkeit, das formallogisch an F-Logic orientierte Schema für beliebige Akteurs-orientierte Szenarien anzupassen. Ein so verwaltetes Wissensnetz kann mit Hilfe von Sparql-Queries angefragt und komfortabel im Terminologie-Navigator erforscht werden.

[7] Gemäß der Umsetzung des Standards mit Hilfe von Semantic Web Technologien entsprechen die Entitäten Klassen im Sinne von *rdfs:Class* und die Kanten sowie Attribute einer Entität Relationen im Sinne von *rdf:Properties.*

Literaturverzeichnis

[ApJ14] Jena Framework, Apache, http://jena.apache.org (zuletzt besucht: 9.9.2014).

[ApL14] Lucene Framework, Apache, http://lucene.apache.org (zuletzt besucht: 9.9.2014).

[Gr93] Gruber, T. R.: A Translation Approach to Portable Ontology Specifications. In: Knowledge Acquisition 5(2), London 1993; S. 199–220.

[HL08] Core Principles and Properties of HL7 Version 3 Models, HL7® Version 3 Standard, Health Level Seven®, 2008.

[ISO14] ISO 13120:2013 Health informatics -- Syntax to represent the content of healthcare classification systems -- Classification Markup Language (ClaML). International Organization for Standardization ISO, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52952 (zuletzt besucht: 9.9.2014).

[KLW91] Kifer, M; Lausen, G; Wu, J.: Logical Foundations of Object-Oriented and Frame-Based Languages. In: Journal of ACM 42(4), 1995; S. 741–843.

[Mc14] McDonald, C. et al: Logical Observation Identifiers Names and Codes (LOINC) User's Guide, Regenstrief Institute, Inc. and the Logical Observation Identifiers and Codes (LOINC) Committee.

[Mi82] Minker, J.: On indefinite databases and the closed world assumption. In: Proc. 6th Conference on Automated Deduction, London 1982. Springer; S. 292–308.

[OL14] Virtuoso, OpenLink, http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main (zuletzt besucht: 9.9.2014).

[OMG14] Common Terminology Services 2, Version 1.1, OMG, http://www.omg.org/spec/CTS2/1.1/ (zuletzt besucht: 9.9.2014).

[Re14] Logical Observation Identifiers Names and Code, Regenstrief Institute, Inc., http://loinc.org/ (zuletzt besucht: 9.9.2014).

[SS09] Staab, S.; Studer, R. (Eds.): Handbook on Ontologies, Springer, 2009.

[Vr10] Vreeman, D.J.: Clinical LOINC Tutorial, Documents, Indiana University, 2010 http://de.slideshare.net/dvreeman/2010-07-15-clinical-loinc-tutorial-documents# (zuletzt besucht: 9.9.2014).

[W3C14] Semantic Web, W3C, http://www.w3.org/standards/semanticweb (zuletzt besucht: 9.9.2014).