UNIVERSITÄT LEIPZIG

Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE)

Bewertung der Datenqualität der SNIK-Ontologie

Seminararbeit

Leipzig, 19. Februar 2017

vorgelegt von:

Stefan Faulhaber

geb. am: 12.12.1992

Betreuer:

Konrad Höffner

Inhaltsverzeichnis

Abbildungsverzeichnis

Begriffs- Abkürzungsverzeichnis

SNIK ..... Semantisches Netz im Krankenhaus

DFG ..... Deutsche Forschungsgemeinschaft

RDF ..... Ressource Description Framework

XML ..... Extensible Markup Language

URI ..... Universal Resource Identifier

LD ..... Linked Data

SPARQL ..... SPARQL Protocol And RDF Query Language

HTTP ..... Hypertext Transfer Protocol

Bash ..... Bourne-again shell

1 Einleitung

1.1 Gegenstand und Motivation

1.1.1 Gegenstand

Seit dem Vorschlag des Semantic Web (Berners-Lee et al., 2001) hat sich eine Menge von mindestens 154 Milliarden[1] RDF-Tripeln angesammelt (Demter et al., 2012). Diese Menge von Linked Data ist jedoch nur dann sinnvoll nutzbar, wenn die Qualität der Anwendung entsprechend entsprechend hoch ist (vgl. Zaveri et al., 2012, S. 1).

In dieser Seminararbeit wird die Ontologie des SNIK (Semantisches Netz des Informationsmanagements im Krankenhaus) bezüglich ihrer Datenqualität untersucht. Die Aufrechterhaltung einer hohen Datenqualität ist eine Voraussetzung für die Erreichung der Ziele des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts (vgl. Drepper, 2016). Dazu zählt vor allem der Einsatz von SNIK in der Lehre und Softwareentwicklung (vgl. Schaaf et al., 2014, S. 754).

Die Studie von Zaveri et al. (2012) bildet die Grundlage dieser Ausarbeitung, da sie existierende Ansätze zum Thema „Datenqualitätsanalyse von Linked Data“ zusammenfasst und qualitativ analysiert (vgl. Zaveri et al., 2012, S. 2). Dabei werden 18 Qualitätsdimensionen und 69 dazu passende Qualitätsmetriken vorgestellt, mit denen die Datenqualität von Linked Data bewertet werden kann (vgl. Zaveri et al., 2012, S. 9, 12, 16, 17, 20). Weiterhin ist diese Studie durch die Spezialisierung auf Linked Data und den zusammenfassenden Charakter auf die SNIK-Ontologie anwendbar.

1.1.2 Problematik

Zum Zeitpunkt der Verfassung dieser Ausarbeitung liegt keine Evaluation der SNIK-Ontologie in Form einer Datenqualitätsanalyse vor. Damit basiert die Entwicklung der

[1] Bestimmt mittels LODStats (`http://stats.lod2.eu`) am 27.12.2016.

Ontologie noch nicht auf objektiven Qualitätsmetriken und Bewertungen in geeigneten Qualitätsdimensionen.

Unter anderem sind die Qualitätsdimensionen Verfügbarkeit, Lizensierung und Verständlichkeit mangelhaft von der SNIK-Ontologie erfüllt (siehe Kapitel 3.1.1 und 3.1.3). Passende Lösungsansätze dazu sind in Kapitel 5 mit einer Aufwandseinschätzung gegeben und sollten in Zukunft nach einer umfangreicheren und auf dieser Ausarbeitung aufbauenden Datenqualitätsanalyse umgesetzt werden.

1.1.3 Motivation

Die Datenqualitätsanalyse einer Ontologie ist die Grundlage für eine gezielte Weiterentwicklung (vgl. Zaveri et al., 2012, S. 2). Im Fall der SNIK-Ontologie kann diese darauf aufbauend gezielt weiterentwickelt werden, da die Analyse Schwachstellen der verlinkten Daten aufdeckt. Stärken der Ontologie werden ebenso herausgestellt, sodass die bestehende Entwicklungsmethodik bezüglich dieser Bereiche als korrekt bestätigt wird. Insgesamt betrachtet gibt die Datenqualitätsanalyse also Hinweise zur Korrektur des weiteren Entwicklungsverlaufs oder begründet diesen in seiner Richtung.

Bei der Entwicklung der SNIK-Ontologie wurden bereits objektive Datenqualitätsmethodiken verwendet. Diese wurden im Gegensatz zu den Methodiken von Zaveri et al. (2012) jedoch nicht in Bezug auf Linked Data systematisch analysiert oder bewertet.

1.2 Problemstellung

Es ist bisher keine standardisierte und objektive Datenqualitätsbewertung der SNIK-Ontologie vorhanden. Basierend auf der systematischen Bewertung existierender Ansätze von Zaveri et al. (2012) soll diese Lücke in der Evaluation geschlossen werden.

1.3 Zielsetzung

Für die Evaluation ist eine vorherige Auswahl von zur Domäne passenden Qualitätsdimensionen und -metriken von (Zaveri et al., 2012) nötig, da im konkreten Fall der SNIK-Ontologie nicht alle Qualitätsdimensionen und -metriken objektiv quantifizierbar sind. Dazu soll zu jeder Dimension und deren Metriken eine Untersuchung ihrer Relevanz erfolgen. Daraus ergibt sich eine nachvollziehbare Einschränkung des Qualitätsraums, ohne die Sinnhaftigkeit der Qualitätsanalyse der SNIK-Ontologie zu gefährden.

Anschließend sollen diese Qualitätsdimensionen und -metriken in Form einer Datenqualitätsanalyse auf die SNIK-Ontologie angewendet werden. Daraus ergibt sich ein Qualitätsprofil, das die Identifizierung von Schwachstellen ermöglicht.

Danach sollen Verbesserungsansätze für die identifizierten Schwachstellen aufgestellt werden, um die weitere Entwicklung der SNIK-Ontologie zu unterstützen. Weiterhin soll zur einfacheren Priorisierung von Verbesserungsmöglichkeiten abschließend eine Schätzung des Verbesserungsaufwands erfolgen.

Die Zielsetzung und der damit verbundene Umfang dieser Seminararbeit sind durch den festgelegten Arbeitsaufwand des vorlesungsbegleitenden Seminars begrenzt. Erweiterungsmöglichkeiten dieser Ausarbeitung werden in der Diskussion am Ende aufgezeigt.

1.4 Aufgabenstellung

Die Aufgabe dieser Ausarbeitung ist die Erstellung einer auf die SNIK-Ontologie bezogenen und begründeten Analyse der Relevanz von Qualitätsdimensionen und -metriken von Zaveri et al. (2012) in Form einer systematischen Evaluation zur Datenqualität.

2 Grundlagen

In diesem Kapitel werden einige grundlegende Begriffe wie SNIK und Semantic Web erläutert. Diese Begriffe werden häufig in den folgenden Kapiteln verwendet und sind somit wichtig für das Verständnis des Gesamtkontextes. Anschließend wird die Relevanz von Qualitätsdimensionen für die SNIK-Ontologie untersucht.

2.1 SNIK

SNIK ist die Abkürzung für das Projekt „Semantisches Netz des Informationsmanagements im Krankenhaus“ (vgl. Jahn et al., 2014, S. 1492). Eine Zielstellung des Projektes ist die Beschreibung komplexer Zusammenhänge der verschiedenen Ansätze des Informationsmanagements (vgl. Jahn et al., 2014, S. 1492). Ein weiteres Ziel wird von Jahn et al. (2014, S. 1493) beschrieben:

> *„Auf Basis des SNIK sollen zukünftig entscheidungsunterstützende Werkzeuge für CIOs von Krankenhäusern entwickelt werden können. Derartige Werkzeuge integrieren Informationen aus dem Krankenhaus und dessen Informationssystem für die vielfältigen Teilaufgaben des Informationsmanagements und unterstützen die Kommunikation zwischen CIO und Krankenhausleitung.“*

Weiterhin ist das SNIK durch folgende drei Festlegungen spezifiziert (vgl. Jahn et al., 2014, S. 1495):

1. Zweck: Neben der Konstruktion entscheidungsunterstützender Werkzeuge für den CIO im Krankenhaus soll SNIK auch die Lehre und Forschung zum Informationsmanagement unterstützen.
2. Formalitätsgrad: Da semantische Netze nicht nur für Computer interpretierbar, sondern auch für menschliche Nutzer verständlich sind, ist im Sinne der genannten Verwendungszwecke die Entwicklung eines semantischen Netzes gegenüber einer Ontologie, die dem Prädikatenkalkül erster Stufe genügen muss, angestrebt.

3. Ein- und Ausschlusskriterien: Ein umfangreicher Fragenkatalog, an dem sich der Ein- und Ausschluss von Begriffen orientieren soll, fasst die Fragen zusammen, die mit Hilfe des semantischen Netzes bzw. darauf aufbauenden Werkzeugen für den CIO beantwortet werden sollen.

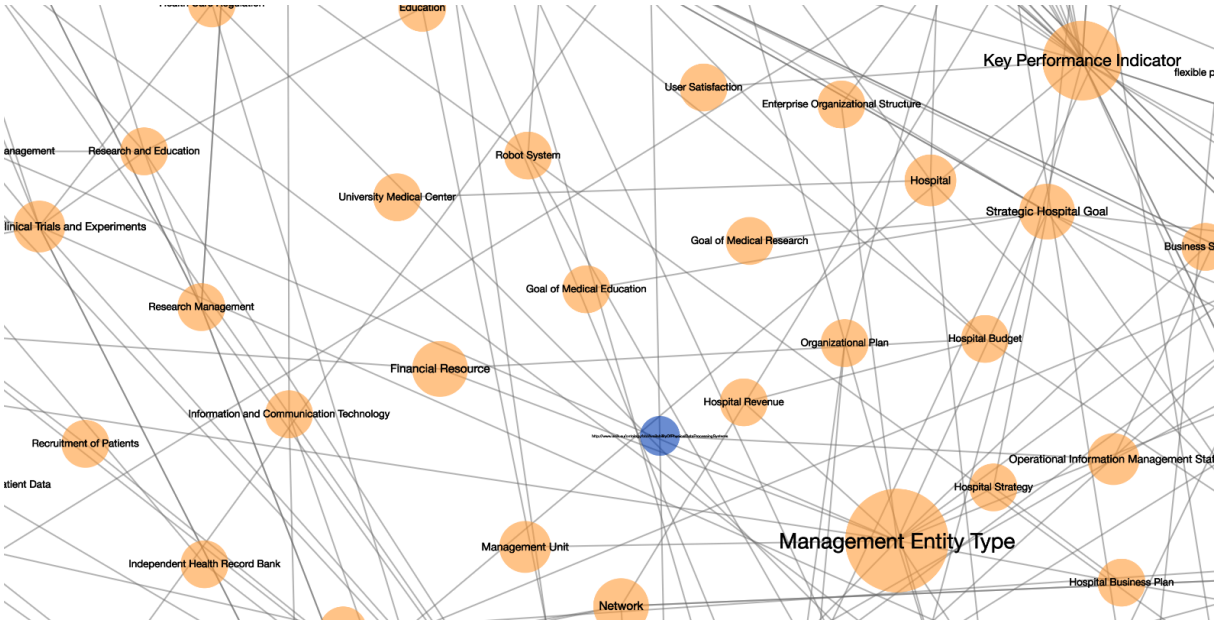

Abb. 2.1: Ausschnitt der SNIK-Visualisierung[1]

2.2 Semantic Web

Das „Semantic Web“ wurde von Berners-Lee et al. (2001) als Erweiterung des bereits existierenden Internets definiert (vgl. Berners-Lee et al., 2001, S. 3). Es gibt den vorhandenen Informationen eine klar definierte Bedeutung und ermöglicht damit eine bessere Zusammenarbeit zwischen Mensch und Computer (vgl. Berners-Lee et al., 2001, S. 4).

Zwei wichtige Technologien zur Entwicklung des Semantic Webs sind die „eXtensible Markup Language“ (XML) und das „Resource Description Framework“ (RDF), mit deren Hilfe Informationen annotiert und miteinander in Verbindung gebracht werden können (vgl. Berners-Lee et al., 2001, S. 7). Es können so Tripel (ähnlich der Satzstruktur „Subjekt – Prädikat – Objekt“ in menschlicher Sprache) erstellt werden, die von einem Computer gelesen und verarbeitet werden können (vgl. Berners-Lee et al., 2001, S. 7, 8). Über „Universal Resource Identifier“ (URI) sind Subjekte und Objekte eindeutig identifizierbar, was die Definition neuer Konzepte vereinfacht (vgl. Berners-Lee et al., 2001, S. 8).

[1] `http://www.snik.eu/graph/`

2.3 Datenqualität

Die Qualität von Daten aus dem Internet kann je nach Quelle stark variieren und ist so nicht für jeden Anwendungsfall geeignet (vgl. Zaveri et al., 2012, S. 2). Ein Datensatz mit Inkonsistenzen, sowie Fehlinterpretationen und unvollständigen Informationen wie DBpedia[2] kann zum Beispiel zur Anreicherung von Suchergebnissen beitragen, aber für medizinische Anwendungen ungeeignet sein (vgl. Zaveri et al., 2012, S. 2).

Es ist somit eine Herausforderung, die Qualität von Datensätzen zu bestimmen und explizit verfügbar zu machen (vgl. Zaveri et al., 2012, S. 2). Im Gegensatz zum traditionellen Internet, in dem die Datenqualität nur ungenau bestimmt werden kann, sind für das Semantic Web konkrete und messbare Metriken verfügbar (vgl. Zaveri et al., 2012, S. 2).

2.3.1 Qualitätsmetriken

Zaveri et al. (2012, S. 7) geben eine Definition einer Qualitätsmetrik vor:

> *„A data quality assessment metric [...] is a procedure for measuring a data quality dimension. These metrics are heuristics that are designed to fit a specific assessment situation. [...] the assessment metrics rely on quality indicators that allow the assessment of the quality of a data source [...]. An assessment score is computed from these indicators using a scoring function.“*

2.3.2 Qualitätsdimensionen

Eine Datenqualitätsdimension ist nach Zaveri et al. (2012, S. 7) eine Charakteristik eines Datensatzes, die von einer Datenqualitätsmetrik gemessen wird (siehe Kapitel 2.3.1). Qualitätsdimensionen können je nach Informationstyp in drei Kategorien eingeteilt werden: Erreichbarkeitsdimensionen, intrinsische Dimensionen, kontextbezogene Dimensionen und Repräsentationsdimensionen (vgl. Zaveri et al., 2012, S. 7).

In Kapitel 3 werden nachfolgend alle von Zaveri et al. (2012) definierten Qualitätsdimensionen vorgestellt und bezüglich ihrer Relevanz für die SNIK-Ontologie untersucht.

[2] `http://wiki.dbpedia.org/datasets`

3 Relevanz der Qualitätsdimensionen

3.1 Untersuchung der Qualitätsdimensionen

Um die Relevanz der von Zaveri et al. (2012) vorgeschlagenen Qualitätsdimensionen einzuschätzen, wird zu jeder Dimension zunächst eine Definition und einige zugehörigen Metriken gegeben. Darauf aufbauend wird dann die Relevanz für die SNIK-Ontologie festgestellt.

3.1.1 Erreichbarkeitsdimensionen

Verfügbarkeit

> *„Availability of a dataset is the extent to which data (or some portion of it) is present, obtainable and ready for use.“*
>
> Zaveri et al. (2012, S. 8)

Die Datenverfügbarkeit ist für die SNIK-Ontologie relevant, da nicht verfügbare Daten die praktische Nutzung erschweren. Die Unterstützung der Entwicklung von medizinischen Anwendungen ist ein erklärtes Ziel des SNIK-Projektes (vgl. Schaaf et al., 2014, S. 754) und hängt von der Datenverfügbarkeit ab, da fehlende Daten Lücken in der Unterstützung der Softwareentwicklung verursachen.

Lizensierung

> *„Licensing is defined as the granting of permission for a consumer to reuse a dataset under defined conditions.“*
>
> Zaveri et al. (2012, S. 8)

Die Lizensierung ist für die SNIK-Ontologie relevant. In der Vorstellung des Projektes von

Jahn et al. (2014) werden zwar keine Nutzungsbeschränkungen vorgegeben, jedoch gibt es auch für diesen Fall passende Lizenzen.

Interlinking

> *„Interlinking refers to the degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources.“*
>
> Zaveri et al. (2012, S. 10)

Die semantische Integration von Synonymen, Homonymen, Vergröberungen und Begriffsüberschneidungen ist eine Herausforderung, die mit Hilfe der SNIK-Ontologie bewältigt werden soll (vgl. Jahn et al., 2014, S. 1492). Damit ist das Interlinking eine relevante Qualitätsdimension.

Sicherheit

> *„Security is the extent to which data is protected against alteration and misuse.“*
>
> Zaveri et al. (2012, S. 10)

Ähnlich wie bei der Qualitätsdimension „Lizensierung“ sehen Jahn et al. (2014) keine Sicherheitsmerkmale wie digitale Signaturen vor. Die Sicherheit wird hier demnach als irrelevante Qualitätsdimension eingestuft.

Performanz

> *„Performance refers to the efficiency of a system that binds to a large dataset, that is, the more performant a data source is the more efficiently a system can process data.“*
>
> Zaveri et al. (2012, S. 10)

Für die Nutzung der SNIK-Ontologie als entscheidungsunterstützendes System (vgl. Jahn et al., 2014, S. 1493) ist eine angemessene Performanz von Bedeutung. Weiterhin hängen akzeptable Antwortzeiten eng mit der Gesamtverfügbarkeit der Ontologie zusammen (vgl. Zaveri et al., 2012, S. 11), die für sich gesehen ebenfalls relevant ist.

3.1.2 Intrinsische Dimensionen

Syntaktische Validität

> *„Syntactic validity is defined as the degree to which an RDF document conforms to the specification of the serialization format.“*
>
> Zaveri et al. (2012, S. 11)

Die SNIK-Ontologie aggregiert verschiedene Datenquellen wie Bücher und weitere Ontologien. Um diese Ressourcen fehlerfrei miteinander zu verbinden, ist die Einhaltung syntaktischer Regeln wichtig.

Semantische Genauigkeit

> *„Semantic accuracy is defined as the degree to which data values correctly represent the real world facts.“*
>
> Zaveri et al. (2012, S. 11)

Um die Entscheidungsfindung für CIOs zu unterstützen, soll die SNIK-Ontologie vorhandene und auf unterschiedlichen Ansätzen und Frameworks basierende Werkzeuge semantisch integrieren (vgl. Jahn et al., 2014, 1492). Dafür ist die semantische Genauigkeit der Daten eine Voraussetzung.

Konsistenz

> *„Consistency means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms.“*
>
> Zaveri et al. (2012, S. 13)

Die Konsistenz ist eng mit der semantischen Genauigkeit verwandt und somit ebenfalls eine wichtige Grundlage für das Erreichen des Ziels der SNIK-Ontologie, verschiedene Werkzeuge semantisch zu integrieren (vgl. Zaveri et al., 2012, S. 15).

Prägnanz

> *„Conciseness refers to the minimization of redundancy of entities at the schema and the data level. [...]“*
>
> Zaveri et al. (2012, S. 14)

Die Prägnanz ist eine für die SNIK-Ontologie relevante Qualitätsdimension, da möglichst wenige Redundanzen im Schema und in den Daten die Ausdruckskraft der Ontologie erhöhen. Würde diese Qualitätsdimension fehlen, hätten redundante Duplikate keinen negativen Einfluss auf die Qualität der Ontologie.

Vollständigkeit

> *„Completeness refers to the degree to which all required information is present in a particular dataset. [...]“*
>
> Zaveri et al. (2012, S. 15)

Auch wenn die komplette Vollständigkeit der SNIK-Ontologie eine Utopie ist, hilft diese Dimension bei der Bestimmung der Datenqualität. Ein fehlendes Konzept, das durch die mangelhafte Vollständigkeit der Ontologie hervorgerufen wird, kann beispielsweise bei kritischen Entscheidungen wertvolle Informationen verbergen und die Ontologie in ihrer Rolle zur Entscheidungsfindungsunterstützung schwächen.

3.1.3 Kontextbezogene Dimensionen

Relevanz

> *„Relevancy refers to the provision of information which is in accordance with the task at hand and important to the users query.“*
>
> Zaveri et al. (2012, S. 17)

Die Relevanz spielt als Qualitätsdimension in Bezug auf die SNIK-Ontologie eine untergeordnete Rolle. Eine höchstmögliche Relevanz der Daten ist zwar erstrebenswert, jedoch zum Großteil schon durch den aggregierenden Charakter der SNIK-Ontologie gegeben (vgl. Jahn et al., 2014, 1492). Trotzdem ist die Relevanz wichtig für die SNIK-Ontologie.

Glaubhaftigkeit

> *„Trustworthiness is defined as the degree to which the information is accepted to be correct, true, real and credible.“*
>
> Zaveri et al. (2012, S. 17)

Die Glaubhaftigkeit der Daten ist zwar wichtig, jedoch ähnlich wie die Relevanz schon durch die glaubhaften Ressourcen gegeben, aus denen die SNIK-Ontologie besteht. Als Qualitätsdimension ist sie in Bezug auf die SNIK-Ontologie dennoch relevant.

Verständlichkeit

> *„Understandability refers to the ease with which data can be comprehended without ambiguity and be used by a human information consumer.“*
>
> Zaveri et al. (2012, S. 18)

Vor allem für die Nutzung in der Lehre und Softwareentwicklung (vgl. Jahn et al., 2014, 1492) ist die Verständlichkeit der SNIK-Ontologie ein wichtiges Merkmal. Eine unverständliche Ontologie würde die Nützlichkeit der Ontologie für menschliche Benutzer einschränken. Deswegen ist die Verständlichkeit eine für die SNIK-Ontologie relevante Qualitätsdimension.

Aktualität

> *„Timeliness measures how up-to-date data is relative to a specific task.“*
>
> Zaveri et al. (2012, S. 19)

Durch die relativ langsam verlaufende Weiterentwicklung des Wissens in der von der SNIK-Ontologie abgedeckten Domäne ist die Aktualität keine relevante Qualitätsdimension.

3.1.4 Repräsentationsdimensionen

Repräsentationsprägnanz

> *„Representational-conciseness refers to the representation of the data, which is compact and well formatted on the one hand and clear and complete on the*

other hand.“

Zaveri et al. (2012, S. 19)

Ein Aspekt der Repräsentationsprägnanz ist die Minimierung der URI-Länge einer Ressource (vgl. Zaveri et al., 2012, S. 20, 21). Eine möglichst kurze URI hilft den Nutzern bei der Verinnerlichung und Kommunikation der entsprechenden Ressource (vgl. Zaveri et al., 2012, S. 21). In Bezug auf die SNIK-Ontologie ist dieser und ähnliche Aspekte der Repräsentationsprägnanz unwichtig, was diese Qualitätsdimension irrelevant macht.

Interoperabilität

> *„Interoperability is the degree to which the format and structure of the information conforms to previously returned information as well as data from other sources.“*
>
> Zaveri et al. (2012, S. 20)

Das Ziel SNIK-Ontologie, zur Unterstützung der Entscheidungsfindung im Informationsmanagement beizutragen (vgl. Jahn et al., 2014, S. 1492), impliziert die intendierte Nutzung weiterer Werkzeuge. Da die Entwicklung alternativer Werkzeuge in der Domäne nicht vorausgesagt oder abgeschätzt werden kann, ist eine hohe Interoperabilität der SNIK-Ontologie ein Weg zur Sicherstellung der Zusammenarbeit verschiedener Werkzeuge. Die Qualitätsdimension „Interoperabilität“ ist somit von Relevanz für die SNIK-Ontologie.

Interpretierbarkeit

> *„Interpretability refers to technical aspects of the data, that is, whether information is represented using an appropriate notation and whether the machine is able to process the data.“*
>
> Zaveri et al. (2012, S. 20)

Der aggregierende Charakter der SNIK-Ontologie erfordert die Einigung auf eine gemeinsame Interpretation von Daten. Nur so können die Vorteile der maschinellen Verarbeitung genutzt werden (vgl. Zaveri et al., 2012, S. 21). Für die SNIK-Ontologie ist die Interpretierbarkeit somit relevant.

Vielseitigkeit

> *„Versatility refers to the availability of the data in different representations and in an internationalized way.“*
>
> Zaveri et al. (2012, S. 21)

Die SNIK-Ontologie ist in englischer Sprache verfasst und somit international lesbar. Verschiedene Repräsentationen wären für das Erreichen der Ziele der SNIK Ontologie jedoch nicht hilfreich (ähnlich der Repräsentationsprägnanz). Trotzdem ist die Vielseitigkeit eine für die SNIK-Ontologie relevante Qualitätsdimension.

3.2 Zusammenfassung

Von den 18 von Zaveri et al. (2012) vorgestellten Qualitätsdimensionen kommen nach genauerer Betrachtung 15 für die Feststellung der Datenqualität der SNIK-Ontologie in Frage.

Aus der Kategorie Erreichbarkeitsdimensionen sind die Verfügbarkeit, die Lizensierung, das Interlinking und die Performanz relevant. Dazu kommen alle intrinsischen Qualitätsdimensionen: Syntaktische Validität, Semantische Genauigkeit, Konsistenz, Prägnanz und Vollständigkeit. Von den kontextbezogenen Dimensionen ist für die SNIK-Ontologie die Relevanz, die Glaubhaftigkeit und die Verständlichkeit von Bedeutung. Abschließend wurden die Qualitätsdimensionen Interoperabilität, Interpretierbarkeit und Vielseitigkeit als wichtig eingestuft.

Nachfolgend wird die Datenqualität der SNIK-Ontologie anhand einer Teilmenge dieser Dimensionen analysiert, um dabei potentielle Schwachstellen der Datenqualität aufzudecken.

4 Datenqualitätsanalyse der SNIK-Ontologie

In diesem Kapitel werden die als relevant eingestuften Qualitätsdimensionen nun auf die SNIK-Ontologie angewendet. Dafür werden die von Zaveri et al. (2012) gegebenen Metriken genutzt.

4.1 Verfügbarkeit

4.1.1 A1: Erreichbarkeit eines SPARQL-Endpunkts

Ein SPARQL-Endpunkt der SNIK-Ontologie ist auf der Website[1] verfügbar und kann genutzt werden.

4.1.2 A2: Erreichbarkeit eines RDF-Dumps

Ein RDF-Dump der SNIK-Ontologie ist auf Github[2] verfügbar.

4.1.3 A3: Dereferenzierbarkeit der URIs

Zur Navigation in der SNIK-Ontologie ist eine Graphenrepräsentation[3] sowie der RDF-Browser LodView[4] verfügbar. Die eigentlichen URIs (z.B. `http://www.snik.eu/ontology/bb/Quality`) liefern bei einer Anfrage jedoch den Fehler „File not found (404)“.

[1] `http://www.snik.eu/sparql`
[2] `http://www.github.com/IMISE/snik-ontology`
[3] `http://www.snik.eu/graph`
[4] `http://lodview.it`

4.1.4 A4: keine fehlerhaften Content-Types

Da URIs der SNIK-Ontologie nicht dereferenzierbar sind, liefern die HTTP-Antworten auch keine Header-Felder. Das Feld „Content-Type“ ist somit immer fehlerhaft.

4.2 Lizensierung

4.2.1 L1: maschinenlesbare Lizenzindikation

Im Datensatz der SNIK-Ontologie ist keine Lizenz spezifiziert.

4.2.2 L2: menschenlesbare Lizenzindikation

In der Dokumentation der SNIK-Ontologie ist keine Lizenz spezifiziert.

4.2.3 L3: korrekte Lizenzindikation

Da keine Lizenz vorhanden ist, kann die Lizenzindikation nicht als korrekt angesehen werden.

4.3 Interlinking

Die SNIK-Ontologie hat keine Interlinks im Sinne von Zaveri et al. (2012) mit anderen Ontologien zum Thema Informationsmanagement, vor allem weil es keine weiteren Ontologien zu diesem Thema gibt. Ersatzweise können jedoch die Interontologierelationen zwischen den beiden Hauptquellen betrachtet werden[5].

4.3.1 I3: dereferenzierte Backlinks

Von den Interontologierelationen ist nur eine Richtung gegeben: aus der bb-Ontologie[6] zur ob-Ontologie[7]. Die Rückrichtung ist nicht explizit angegeben.

[5] siehe `https://github.com/IMISE/snik-ontology/blob/master/links_bb.ttl`
[6] `https://github.com/IMISE/snik-ontology/blob/master/bb.rdf`
[7] `https://github.com/IMISE/snik-ontology/blob/master/ob.rdf`

4.4 Performanz

4.4.1 P1: Nutzung von Slash-URIs

Wie bei der Metrik A4, ist auch P1 durch die fehlende Dereferenzierbarkeit der URIs (A3) nicht erfüllt.

4.4.2 P2: niedrige Latenz

Die folgende SPARQL-Query liefert alle Prädikate der SNIK-Ontologie:

```
select distinct ?p
where {?s ?p ?o.}
```

Um die Latenz zu bestimmen wurden fünf Mal zehn Anfragen ohne Unterbrechung an den Server gestellt. Dazu wurde folgendes Bash-Skript genutzt[8]:

```
#!/bin/bash
echo Measure response time of SPARQL query:

count=0
while (( count < 10 )); do
  time wget -q -O/dev/null "[URL]" --no-cache --no-cookies --no-dns-cache
  count=$[$count+1]
done
```

Das Ergebnis ist eine durchschnittliche Latenz von 0,073 Sekunden:

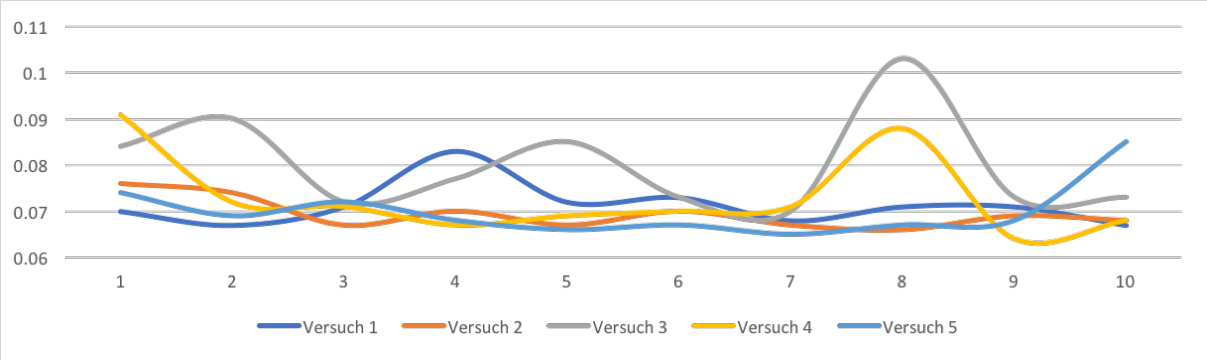

Abb. 4.1: Latenzverlauf des Servers

[8][URL]=http://www.snik.eu/sparql?default-graph-uri=&query=select+distinct+%3Fp%0D%0Awhere+%7B%0D%0A%3Fs+%3Fp+%3Fo.%0D%0A%7D

4.4.3 P3: hoher Durchsatz

Die folgende SPARQL-Query liefert ein Subjekt eines Tripels der SNIK-Ontologie:

```
select ?s
where {?s ?p ?o.}
limit 1
```

Um die maximale Anzahl an beantworteten Anfragen pro Sekunde zu bestimmen wurde fünf mal zehn Sekunden lang ohne Unterbrechung eine Anfrage nach der anderen an den Server gestellt. Die Gesamtanzahl der beantworteten Anfragen wurde dann durch zehn geteilt, um den Durchsatz näherungsweise zu bestimmen[9]:

```
#!/bin/bash

secs=10
echo -e Counting maximum responses per second for $secs seconds...$'\n'

count=0
SECONDS=0
while (( SECONDS < secs )); do
  time -p wget -q -O/dev/null "[URL]" --no-cache --no-cookies --no-dns-cache
  count=$[$count+1]
done

echo Anfragen pro Sekunde: $[$count/$secs]
```

Das Ergebnis ist ein durchschnittlicher Durchsatz von 12 Anfragen pro Sekunde.

4.4.4 P4: Skalierbarkeit der Datenquelle

Sei t_1 die Zeit zur einmaligen Beantwortung einer Anfrage q. Weiterhin sei t_{10} die Zeit, die bei der zehnmaligen Beantwortung von q vergeht. Die Skalierbarkeit kann nun bestimmt werden, indem t_1 mit t_{10} verglichen wird. Es wurde dazu die Anfrage aus P2 genutzt:

```
#!/bin/bash

echo -e Determine the scalability of the datasource: $'\n'
echo Time of one request:

time wget -q -O/dev/null "[URL]" --no-cache --no-cookies --no-dns-cache

echo -e $'\n' Time of ten requests:
```

[9] [URL]=http://www.snik.eu/sparql?default-graph-uri=&query=select+distinct+%3Fp%0D%0Awhere+%7B%0D%0A%3Fs+%3Fp+%3Fo.%0D%0A%7D

```
count=0
time while (( count < 10 )); do
  wget -q -O/dev/null "[URL]" --no-cache --no-cookies --no-dns-cache
  count=$[$count+1]
done
```

t_1 ist dabei im Schnitt höher als t_{10}, was auf eine gute Skalierbarkeit schließen lässt.

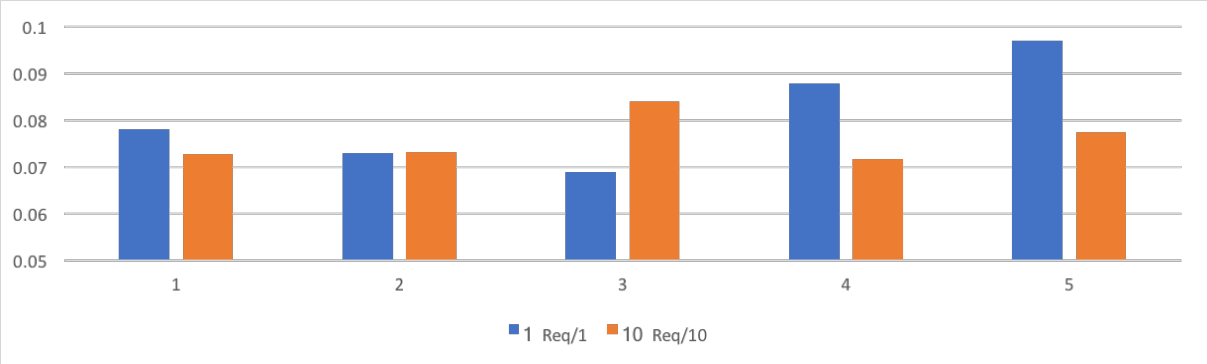

Abb. 4.2: normierte Anwortzeiten bei einer und zehn Anfragen

4.5 Syntaktische Validität

4.5.1 SV1: keine Syntax-Fehler in den Dokumenten

Eine Überprüfung der RDF-Dokumente mit *xmllint*[10] ergab keine keine Treffer. Eine Überprüfung der SNIK-Ontologie mit dem *W3C-RDF-Validator*[11] fand einen Fehler in der ob.rdf-Datei: in Zeile 869 befindet sich ein überflüssiger Punkt am Ende der Zeile.

4.5.2 SV2: syntaktisch korrekte Werte

Eine manuelle Überprüfung der Ontologie ergab unter Anderem die falsche Verwendung des page-Datentyps. Definiert ist dieser Datentyp als nicht-negative ganze Zahl:

```
<owl:DatatypeProperty rdf:about="page">
  <rdfs:label xml:lang="en">page</rdfs:label>
  <rdfs:label xml:lang="de">Seite</rdfs:label>
  <rdfs:range rdf:resource="&xsd;nonNegativeInteger"/>
  [...]
</owl:DatatypeProperty>
```

[10] http://xmlsoft.org/xmllint.html
[11] https://validator.w3.org/#validate_by_uri

Verwendet wird er jedoch teilweise als Tupel ganzer Zahlen:

```
<owl:Class rdf:about="ADT">
  [...]
  <page>150, 294</page>
  [...]
</owl:Class>
```

4.5.3 SV3: keine falschen Datentypliterale

Mit folgender SPARQL-Query konnte festgestellt werden, dass die Datentypliterale *page* und *cardinality* falsch verwendet wurden:

```
select ?p count(?o)
{
  ?p rdfs:range xsd:nonNegativeInteger.
  ?s ?p ?o.
  filter(regex(str(?o),"^[0-9]+$"))
}
```

4.6 Glaubhaftigkeit

Durch den aggregierenden Charakter der SNIK-Ontologie ist die Glaubhaftigkeit der Ontologie im Allgemeinen auf die Glaubhaftigkeit der Quellen zurückzuführen. Die beiden Hauptquellen sind Fachbücher von Winter et al. (2010) und Ammenwerth and Haux (2005).

4.7 Verständlichkeit

4.7.1 U1: menschenlesbare Klassennamen

Folgende SPARQL-Query gibt eine Liste aller URIs aus, deren rdfs:Label-Property leer ist oder fehlt:

```
select ?uri ?label
where
{
  ?uri rdfs:label ?label .
  filter(?label = "")
}
```

In der SNIK-Ontologie konnten mit dieser Query keine URIs gefunden werden, deren Klassenname fehlt oder leer ist.

4.7.2 U2: Angabe von Beispiel-URIs

Es ist keine Beispiel-URI vorhanden, die das URI-Muster verdeutlicht.

4.7.3 U3: Angabe eines regulären Ausdrucks, der mit den URIs übereinstimmt

Es ist kein regulärer Ausdruck vorhanden, der mit allen URIs der SNIK-Ontologie übereinstimmt.

4.7.4 U4: Angabe einer Beispiel-Query

Es ist keine SPARQL-Query vorhanden, die die Nutzung des SPARQL-Endpunkts exemplarisch demonstriert.

4.7.5 U5: Angabe der genutzten Vokabulare

Die genutzten Vokabulare sind nur implizit in den RDF-Dateien[12] vorhanden.

4.7.6 U6: Bereitstellung eines Forums und einer Mailingliste

Es ist kein Forum vorhanden, jedoch sind alle E-Mail Adressen des Projektteams auf der Website[13] vorhanden.

[12] z.B. `https://github.com/IMISE/snik-ontology/blob/master/meta.rdf#L12`
[13] `http://www.snik.eu/de/Team/index.jsp`

4.8 Interoperabilität

4.8.1 IO1: Verwendung existierender Terme

An geeigneten Stellen werden vorhandene Terme verwendet, zum Beispiel „label“.

4.8.2 IO2: Verwendung existierender Vokabulare

An geeigneten Stellen werden vorhandene Vokabulare verwendet, zum Beispiel „rdfs“.

4.9 Interpretierbarkeit

4.9.1 IN3: keine falsche Nutzung undefinierter Klassen und Properties

Mit folgender SPARQL-Query kann überprüft werden, ob definierte Klassen falsch verwendet werden:

```
select ?c
{
 ?s ?c ?o.
 ?c a owl:Class.
}
```

Es konnten damit keine falsch verwendeten Klassen gefunden werden. Weiterhin muss die SNIK-Ontologie nun undefinierte Klassen und Properties untersucht werden. Die folgende SPARQL-Query findet 17 undefinierte Properties:

```
select distinct(?p)
{
 ?s ?p ?o.
 filter not exists {?p ?x ?y.}
 filter(regex(str(?p),"http://www.snik.eu/ontology/"))
}
```

Die ähnliche SPARQL-Query zur Identifizierung undefinierter Klassen liefert keine Ergebnisse:

```
select distinct(?c)
{
```

```
 ?s a ?c.
 filter not exists {?c ?x ?y.}
 filter(regex(str(?c),"http://www.snik.eu/ontology/"))
}
```

4.10 Vielseitigkeit

4.10.1 V1: Bereitstellung der Daten in verschiedenen Serialisierungsformaten

Die SNIK-Ontologie liegt nur in Form von RDF-Dokumenten vor.

4.10.2 V2: Bereitstellung der Daten in verschiedenen Sprachen

Mit folgender SPARQL-Query wird die Sprachindikation der Ressourcenlabels gezählt:

```
select distinct(lang(?l)) count(?x)
{
 ?x rdfs:label ?l.
} group by lang(?l)
```

Dabei ergibt sich folgende Verteilung:

- englisch (en): 2166
- deutsch (de): 658
- ohne Sprachindikation: 289
- vermutlich Tippfehler (ce): 1

Nach der Qualitätsanalyse der SNIK-Ontologie werden im folgenden Kapitel Verbesserungsmöglichkeiten aufgezeigt. Dabei wird bei jeder Verbesserung auch eine kurze Aufwandseinschätzung gegeben, um die Priorisierung der Umsetzung zu unterstützen.

5 Verbesserungspotential der SNIK-Ontologie

5.1 Verfügbarkeit

Zur Erfüllung der Metriken A3 und A4 müssen die URIs der Ontologie eingerichtet werden. Aktuell kann zwar per lodview durch die Ontologie navigiert werden, jedoch ist die Dereferenzierbarkeit der eigentlich URIs für das Semantic Web von Bedeutung. Dabei ist auch auf die Korrektheit der Header-Felder zu achten, die bei einer Anfrage zurückgegeben werden. Der Aufwand für dieser Verbesserungen wird auf mittel bis hoch geschätzt.

5.2 Lizensierung

Die Einbindung einer Lizenz (z.B. MIT[1]) auf der Website, im Repository auf Github und im Datensatz selbst ist eine mit geringem Umfang verbundene Verbesserung bezüglich der Metriken L1, L2 und L3.

5.3 Syntaktische Validität

Die Behebung des Syntaxfehlers ist mit einem sehr geringen Aufwand verbunden und bringt die Erfüllung der SV1-Metrik mit sich.

Die Erfüllung von SV2 und SV3 ist mit einem mittleren bis hohen Aufwand verbunden, da beispielsweise entweder der Datentyp *page* angepasst werden muss oder alle Verwendungen dieses Datentyps in der gesamten Ontologie angepasst werden muss.

[1] `https://opensource.org/licenses/MIT`

5.4 Verständlichkeit

Die Erweiterung der Website des SNIK-Projektes kann die Metriken U2 bis U6 erfüllen. Dazu ist eine neue Unterseite geeignet, die neuen Nutzern der Ontologie die Einarbeitung erleichtert. Diese Unterseite umfasst dann Beispiel-URIs, einen regulären Ausdruck, mehrere Beispiel-Queries, erweiterte Kontaktmöglichkeiten für den Support und eine klare Angabe genutzter Vokabulare. Eine solche Websiteerweiterung ist mit einem mittleren Aufwand verbunden.

5.5 Vielseitigkeit

Eine mit hohem Aufwand verbundene Verbesserung zur leichteren internationalen Nutzung der SNIK-Ontologie ist die Übersetzung der Labels in verschiedene Sprachen. Aktuell liegt die Ontologie hauptsächlich in englischer und deutscher Sprache vor. Eine vollständige Übersetzung auf Deutsch und Englisch wäre ein realistischer erster Schritt.

6 Zusammenfassung

In dieser Ausarbeitung wurde nach der Einführung wichtiger Begriffe wie SNIK, Semantic Web und Datenqualität die Relevanz der von Zaveri et al. (2012) definierten Qualitätsdimensionen untersucht. Dabei wurden drei Qualitätsdimensionen als für die SNIK-Ontologie irrelevant befunden. Anschließend wurden zehn der relevanten Dimensionen anhand der SNIK-Ontologie mit den entsprechenden Qualitätsmetriken von Zaveri et al. (2012) untersucht. Abschließend wurden Vorschläge zur Behebung von Schwachstellen der SNIK-Ontologie aufgezeigt und jeweils eine grobe Aufwandseinschätzung gegeben.

7 Diskussion und Ausblick

Die Erkenntnisse dieser Arbeit liegen zum Einen in der Relevanzuntersuchung und zum Anderen in der Anwendung der Qualitätsdimensionen auf die SNIK-Ontologie. Der begrenzte Zeitrahmen lies jedoch keine komplette Untersuchung aller 15 relevanten Dimensionen zu. Diese Datenqualitätsbewertung ist somit ein Einstieg und kann in Zukunft als Grundlage für eine vollständige und umfangreichere Bewertung der Datenqualität der SNIK-Ontologie dienen. Dabei sollten dann alle Dimensionen mit Hilfe von allen dazugehörigen Qualitätsmetriken untersucht werden.

Literaturverzeichnis

Ammenwerth, E. and Haux, R. (2005). IT-Projektmanagement in Krankenhaus und Gesundheitswesen. *Schattauer*.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific american*, 284(5):28–37.

Demter, J., Auer, S., Martin, M., and Lehmann, J. (2012). LODStats — An Extensible Framework for High-performance Dataset Analytics. In *International Conference on Knowledge Engineering and Knowledge Management*, Lecture Notes in Computer Science (LNCS) 7603, pages 353–362. Springer.

Drepper, J. (2016). Standards und Werkzeuge zur Beurteilung der Datenqualität in komplexen epidemiologischen Studien. URL: `http://gepris.dfg.de/gepris/projekt/315057723`. Abgerufen am 19.02.2017.

Jahn, F., Schaaf, M., Paech, B., and Winter, A. (2014). Ein semantisches Netz des Informationsmanagements im Krankenhaus. In *GI-Jahrestagung*, pages 1491–1498.

Schaaf, M., Jahn, F., Tahar, K., Kücherer, C., Winter, A., and Paech, B. (2014). Eine Ontologie für die Unterstützung der Lehre und des Informationsmanagements im Gesundheitswesen. Universität Leipzig, Leipzig Research Festival for Life Sciences.

Winter, A., Haux, R., Ammenwerth, E., Brigl, B., Hellrung, N., and Jahn, F. (2010). Health Information Systems. *Springer*.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2012). Quality Assessment for Linked Data: A Survey. volume 7, pages 63–93. IOS Press.