



Internationalisierung der GND durch das Semantic Web

Inhalt

Einleitung	3
Gemeinsame Normdatei (GND)	4
Semantic Web und Linked Data	4
RDF, RDFS, OWL und SKOS	5
Ontologien für das Bibliothekswesen	6
Generelle Schritte zur Linked-Data-Veröffentlichung	7
Anwendungsprofile als Basis für die Ontologieentwicklung?	9
Das Singapore Framework	9
Anwendbarkeit des Singapore Frameworks für die Ontologiemodellierung	10
Die GND-Ontologie	11
Ontologiemodellierung	12
GND-Klassen	12
GND-Properties	13
GND Value Vocabularies	15
Alignment der GND-Ontologie	16
Verlinkung zu externen Datensets	17
Zusammenfassung und Ausblick	20
Literatur	21

Einleitung

Seit Bestehen der Menschheit sammelt der Mensch Informationen, seit Bestehen des Internets stellt der Mensch Informationen ins Web, seit Bestehen des Semantic Webs sollen auch Maschinen in die Lage versetzt werden mit diesen Informationen umzugehen.

Das Bibliothekswesen ist einer der Sammler. Seit Jahrhunderten werden Kataloge und Bibliografien sowie Inventarnachweise geführt. Mit der Aufgabe des Zettelkatalogs hin zum Onlinekatalog wurde es Benutzern plötzlich möglich in Beständen komfortabel zu suchen. Durch die Bereitstellung von Daten aus dem Bibliothekswesen im Semantic Web sollen nicht nur die eigenen Katalogsysteme Zugriff auf diese Informationen erhalten, sondern jede beliebige Anwendung, die auf das Web zugreifen kann. Darüber hinaus ist die Vorstellung, dass sich die im Web befindenden Daten - in sofern möglich - miteinander verlinken und zu einem gigantischen semantischen Netz werden, das als ein großer Datenpool verwendet werden kann. Die Voraussetzung hierfür ist wie beim Übergang zum Onlinekatalog die Aufbereitung der Daten in einem passenden Format.

Normdaten dienen im Bibliothekswesen bereits dazu eine Vernetzung der unterschiedlichen Bestände zu erlauben. Bei der Erschließung eines Buches wird nicht bloß gesagt, dass jemand, der Thomas Mann heißt, der Autor ist - es wird eine Verknüpfung vom Katalogisat zu dem Thomas Mann erzeugt, der am 6. Juni 1875 in Lübeck geboren und am 12. August 1955 in Zürich verstorben ist. Der Vorteil von Normdateneintragen ist, dass sie zum eindeutigen Nachweis der Verfasserschaft oder Mitwirkung an einem Werk beitragen. Auch stehen Normdateneintragen bereits allen Bibliotheken für die Nachnutzung bereit - der Schritt ins Semantic Web wäre somit die Öffnung der Normdaten für alle denkbaren Nutzergruppen.

Die Gemeinsame Normdatei (GND) ist seit April 2012 die Datei, die die im deutschsprachigen Bibliothekswesen verwendeten Normdaten enthält. Folglich muss auf Basis dieser Daten eine Repräsentation für die Darstellung als Linked Data im Semantic Web etabliert werden. Neben der eigentlichen Bereitstellung von GND-Daten im Semantic Web sollen die Daten mit bereits als Linked Data vorhandenen Datenbeständen (DBpedia, VIAF etc.) verknüpft und nach Möglichkeit kompatibel sein, wodurch die GND einem internationalen und spartenübergreifenden Publikum zugänglich gemacht wird.

Dieses Dokument dient vor allem zur Beschreibung, wie die GND-Linked-Data-Repräsentation entstand und dem Weg zur Spezifikation einer eignen Ontologie.

Hierfür werden nach einer kurzen Einführung in die GND die Grundprinzipien und wichtigsten Standards für die Veröffentlichung von Linked Data im Semantic Web vorgestellt, um darauf aufbauend existierende Vokabulare und Ontologien des Bibliothekswesens betrachten zu können.

Anschließend folgt ein Exkurs in das generelle Vorgehen für die Bereitstellung von Linked Data, wobei die so oft zitierte Open World Assumption kritisch hinterfragt und damit verbundene Probleme insbesondere in Hinsicht Interoperabilität und Nachnutzbarkeit aufgedeckt werden. Um Probleme der Interoperabilität zu vermeiden, wird den Empfehlungen der Library Linked Data Incubator Group [LLD11] gefolgt.

Im Kapitel Anwendungsprofile als Basis für die Ontologieentwicklung wird die Spezifikation von Dublin Core Anwendungsprofilen kritisch betrachtet, um auszumachen wann und in welcher Form sich ihre Verwendung bei dem Vorhaben Bereitstellung von Linked Data anbietet.

In den nachfolgenden Abschnitten wird die GND-Ontologie, welche als Standard für die Serialisierung von GND-Daten im Semantic Web dient, samt Modellierungsentscheidungen näher vorgestellt. Dabei wird insbesondere der Technik des Vocabulary Alignment eine prominente Position eingeräumt, da darin ein entscheidender Mechanismus zur Steigerung der Interoperabilität und Nachnutzbarkeit gesehen wird.

Auch wird sich mit der Verlinkung zu externen Datensets intensiv beschäftigt. Hierfür wurden ausgewählte Datenbestände hinsichtlich ihrer Qualität und Aktualität untersucht und Empfehlungen für die Implementierung innerhalb des GND-Datenbestandes gegeben.

Abschließend werden eine Zusammenfassung und ein Ausblick auf weitere Schritte gegeben.

Gemeinsame Normdatei (GND)

Im Projekt Gemeinsame Normdatei (GND) wurden die bestehende Personennamendatei (PND), die Schlagwortnormdatei (SWD) und die Gemeinsame Körperschaftsdatei (GKD) sowie die Einheitssachtiteldatei (EST) des Deutschen Musikarchivs in einer gemeinsamen Normdatei (GND) zusammengeführt.

Die bestehenden Normdateien sind historisch gewachsen und in voneinander divergierenden Datenmodellen und in jeweils unterschiedlichen Formaten weiterentwickelt worden, die den heutigen Anforderungen nicht mehr gerecht werden. Im Zuge der Zusammenführung zu einer gemeinsamen Normdatei wurde die gelebte funktionale Differenzierung in Normdaten für die Formalerschließung und Normdaten für die Sacherschließung zugunsten einer objektorientierten Sicht aufgegeben [BEH11]. Folglich wurde auch eine Harmonisierung der bislang angewandten Regelwerke und Ansetzungen für die Erschließung der unterschiedlichen Normdaten vorgenommen. Darüber hinaus werden Datensätze aus GKD und SWD sowie SWD und EST-Datei, die dieselbe Entität beschreiben, systematisch zu einem GND-Datensatz vereinigt.

Die Objektorientierung in der GND und die entitätenbasierte Beschreibung von Normdaten sollen insbesondere zu einer erhöhten Verknüpfung unter den Normdatenentitäten führen. Die Inhalte der Normdatensätze werden zwischen den Bibliotheken aus Interoperabilitätsgründen im GND-MARC-Format¹ ausgetauscht, das sich an das MARC 21 Format for Authority Data² anlehnt. Allerdings muss betont werden, dass die reichhaltigen Relationen unter den Normdaten (5XX-Felder) im international eingesetzten MARC 21 Authority Format nicht abgebildet werden können und somit eine deutschsprachige Besonderheit darstellen. Grund hierfür ist vor allem die Überalterung des MARC 21 Authority Formats und der in der Vergangenheit verkannte Bedarf an Relationierung.

Durch die nachfolgend vorgestellte Repräsentation der GND im Semantic Web als Linked Data soll diese Schwäche hinsichtlich der internationalen Datennachnutzbarkeit und Interoperabilität behoben werden.

Semantic Web und Linked Data

Das Web mit seiner über 20jährigen Geschichte bietet Benutzern die Möglichkeit reichhaltige Informationen von verschiedensten Informationsanbietern zu beziehen. Die Informationen werden in einem Hypertext-Format zur Verfügung gestellt und sind über das HTTP-Protokoll abrufbar. Das primäre Ziel von Hypertext-Formaten ist die Gewährung der Informationszugänglichkeit für den Menschen. Zwar können Inhalte vernetzt bzw. verlinkt werden, gestatten aber nur bedingt eine maschinelle Interpretation und Verarbeitung.

Ziel des Semantic Web ist es, auf Basis des HTTP-Protokolls Informationen, die sich im Web befinden, auch für Maschinen verständlich zu machen. Maschinen sollen dadurch in der Lage sein, die Informationen zu interpretieren und Rückschlüsse daraus ziehen zu können. Darüber hinaus wird die stärkere Vernetzung von zu einander in Beziehung stehenden Konzepten – Aufbau semantischer Netze - forciert. Die im Semantic Web bereitgestellten Daten werden als Linked Data bezeichnet.

Die Veröffentlichung von Linked Data muss dabei den folgenden vier Prinzipien folgen [BER09] [DEL11]:

¹ http://www.dnb.de/SharedDocs/Downloads/DE/DNB/standardisierung/gnd_marc_datensatzformat_1.1

² <http://www.loc.gov/marc/authority/>

1. Für die Bezeichnung einzelner Entitäten sollen URIs (Uniform Resource Identifier) verwendet werden.
2. Die URIs sollen auf dem HTTP-Protokoll basieren, um auf einfache Art dereferenzierbar zu sein. Dieser Mechanismus kann dazu benutzt werden, um weitere Informationen über eine Entität nachzuschlagen.
3. Wenn ein URI dereferenziert wird, sollen weiterführende Informationen über die dadurch bezeichnete Entität zurückgeliefert werden. Diese Informationen sollen in standardisierten Datenformaten wie z. B. RDF, RDFS oder OWL vorliegen und nach Möglichkeit auch durch Anfragesprachen wie z. B. SPARQL abfragbar sein.
4. Die Informationen über eine Entität sollen nach Möglichkeit Links zu anderen Entitäten enthalten. Diese anderen Entitäten müssen nicht notwendigerweise aus der gleichen Datenquelle stammen. Es soll auch zu externen Datenquellen verlinkt werden.

RDF, RDFS, OWL und SKOS

Um Informationen durch Maschinen interpretierbar zu machen, ist die Verwendung von standardisierten und etablierten Auszeichnungssprachen notwendig.

Resource Description Framework (RDF) stellt einen vom World Wide Web Consortium (W3C) veröffentlichten Standard für die Beschreibung von Ressourcen im Semantic Web bereit. Das Prinzip von und der Umgang mit RDF werden in sechs Dokumenten (Primer³, Concepts⁴, Syntax⁵, Semantics⁶, Vocabulary⁷, and Test Cases⁸) detailliert erläutert.

Wie bereits in den vier Prinzipien des Semantic Webs dargestellt, ist der Grundgedanke von RDF, dass alle Ressourcen mit Hilfe von Uniform Resource Identifiers (URI) identifiziert werden können und die Beschreibung der Ressourcen mittels einfacher Properties und Property-Values vorgenommen werden kann. Daten in RDF sind Aussagen (Statements) über Ressourcen. Diese Aussagen werden als Tripel modelliert - ein Tripel besteht dabei immer aus einer Sequenz von drei Elementen:

1. dem Subjekt: die Ressource, über die eine Aussage getroffen wird,
2. dem Prädikat: eine Eigenschaft des Subjekts (Property) und
3. dem Objekt: das Argument des Prädikats (Property Value).

Die Menge der Tripel bilden einen Graphen. Das RDF-Modell basiert dementsprechend auf gerichteten Graphen.

Durch das W3C werden eine Reihe von Repräsentationsformen für die Serialisierung von RDF-Daten, wie RDF/XML, RDF/N3 oder Turtle bereitgestellt.

Die RDF Vocabulary Description Language 1.0 entspricht dem sogenannten RDF Schema (RDFS). Mit RDFS ist es möglich, für einen bestimmten Anwendungsfall Regeln (Syntax und Semantik) für den Einsatz von bestimmten Elementen bzw. Vokabularen (Zusammenstellung von Elementen) innerhalb eines Anwendungskontextes zu definieren. Entsprechend der Ausführungen zu RDF wird festgelegt, welche Ressourcen (Subjekte) durch welche Property (Prädikat) charakterisiert werden dürfen und welche Property-Values (Objekte) in der definierten Konstellation zulässig sind. Die Spezifikation eines RDF Schema wird auch häufiger als Ontologie bezeichnet. Allerdings muss man dabei bemerken, dass dies nicht ganz korrekt ist, da ein RDF Schema im Gegensatz zu einer Ontologie keine Instanzdaten enthält.

Die Web Ontology Language (OWL) ist eine weitere Spezifikation des W3C, um Ontologien anhand einer formalen Beschreibungssprache ausdrücken zu können. Gegenüber RDFS zeichnet sich OWL durch eine höhere Ausdrucksstärke bei der Spezifikation der Semantik aus. Durch OWL kann der

³ <http://www.w3.org/TR/rdf-primer/>

⁴ <http://www.w3.org/TR/rdf-concepts/>

⁵ <http://www.w3.org/TR/rdf-syntax-grammar/>

⁶ <http://www.w3.org/TR/rdf-nt/>

⁷ <http://www.w3.org/TR/rdf-schema/>

⁸ <http://www.w3.org/TR/rdf-testcases/>

volle Umfang der Prädikatenlogik wiedergegeben werden, was es bestimmten Anwendungen (Reasoner) ermöglicht auf der Grundlage dieser Daten Entscheidungen zu treffen. Die Spezifikation liegt derzeit als Empfehlung in der Version OWL 2⁹ vor und der Umgang mit ihr wird ebenfalls durch eine Reihe von Dokumenten¹⁰ beschrieben.

Wie auch RDFS stellt OWL Möglichkeiten zur Definition von Klassen und Properties sowie Property-Values bereit. Darüber hinaus kann OWL auch die eigentlich in RDF beinhalteten Instanzdaten umfassen. Insgesamt stehen drei Versionen von OWL zur Verfügung: OWL Lite, OWL DL und OWL Full.

Für die Repräsentation von Wissensorganisationssystemen (z. B. Thesauri, kontrollierte Vokabulare) wurde durch das W3C das Simple Knowledge Organization System (SKOS)¹¹ spezifiziert. Mittels SKOS ist es möglich, in einer sehr einfachen Ausdrucksform Instanzdaten im Semantic Web zur Verfügung zu stellen. SKOS erfreut sich aufgrund seiner Einfachheit großer Beliebtheit, weist andererseits jedoch eine Vielzahl an Grenzen im praktischen Einsatz auf und besitzt bislang eine Reihe von Inkompatibilitäten beim kombinierten Einsatz mit OWL.

Ontologien für das Bibliothekswesen

Der Ansatz, Bibliotheksdaten im Web bereitzustellen, findet international bei geisteswissenschaftlichen Einrichtungen immer größeren Anklang. Mit Datenformatspezifikationen wie den DCMI Metadata Terms¹² stehen schon länger Möglichkeiten bereit, um bibliografische Daten des Bibliothekswesens im Semantic Web zur Verfügung zu stellen.

Die ersten Linked Library Data Veröffentlichungen zeigten jedoch, dass die in Bibliothekssystemen gehaltenen Daten häufig sehr speziell sind und über die Anforderungen spartenübergreifender Vertreter hinausgehen. Um Daten aus dem Bibliotheksbereich vollständig abbilden zu können, wurden durch unterschiedliche Akteure unabhängig voneinander Ontologien entwickelt.

Nachfolgend soll nur exemplarisch aufgeführt werden, welche RDF Schema bzw. OWL-Ontologien verfügbar sind.

1. Bibliographic Ontology (BIBO)¹³
2. RDA Vocabularies¹⁴
3. MARC 21 Vocabularies¹⁵
4. ISBD Vocabularies¹⁶
5. FRBRer model¹⁷
6. FRAD model¹⁸
7. FRSAD model¹⁹
8. Erlangen CRM / OWL²⁰

In den genannten Ontologien sind teilweise bereits Elemente enthalten, die eine Beschreibung von Normdaten als Linked Data ermöglichen. Normdaten dienen im Semantic Web vor allem als Bindeglied zwischen den bibliografischen Daten und tragen somit zum Aufbau eines semantischen Netzes bei.

Wie verdeutlicht, besitzen die Normdaten aus dem deutschsprachigen Raum bereits vielfältige Verknüpfungen zu bibliografischen Daten sowie untereinander. Entsprechend ist es naheliegend diese Daten als Bestandteil des Semantic Webs zu etablieren.

⁹ <http://www.w3.org/TR/owl-syntax/>

¹⁰ <http://www.w3.org/TR/owl-overview/>

¹¹ <http://www.w3.org/TR/skos-reference>

¹² <http://dublincore.org/documents/dcmi-terms/>

¹³ <http://bibliontology.com/>

¹⁴ <http://rdvocab.info/>

¹⁵ <http://marc21rdf.info/>

¹⁶ <http://iflstandards.info/ns/isbd/>

¹⁷ <http://iflstandards.info/ns/fr/frbr/frbrer/>

¹⁸ <http://iflstandards.info/ns/fr/frad/>

¹⁹ <http://iflstandards.info/ns/fr/frsad/>

²⁰ <http://erlangen-crm.org/current/>

Generelle Schritte zur Linked-Data-Veröffentlichung

Die Schwierigkeit bei der Veröffentlichung von Linked Data ist, dass das Datenkonzept von RDF nicht dem von klassischen Formaten der Bibliothekswelt gleicht. Im Zusammenhang mit dem Semantic Web wird häufig von der Open World Assumption gesprochen – alle Informationen können zusammengeführt werden und es gibt keinerlei Begrenzungen. Ist es wirklich so einfach? Dieser Frage soll im Folgenden nachgegangen werden. Darüber hinaus wird diskutiert, wie sich eine Linked-Data-Bereitstellung in der Praxis verhält und welchen Nutzen sie Dritten bringen kann.

Verfolgt man das Ziel, die eigenen Daten zu einem Teil des Semantic Webs zu machen, wird oft zuerst die Frage gestellt, wie die Daten als Linked Data abgebildet werden sollen. Dabei besteht jedoch die Gefahr einen wichtigen Aspekt für die erfolgreiche Linked-Data-Bereitstellung außen vor zu lassen: Was für Informationen will man eigentlich im Semantic Web bereitstellen?

Die Frage, welche Information man bereitstellen will, muss prinzipiell aus fachlogischer Sicht beantwortet werden und darf nur sekundär von den bislang verwendeten Datenformaten beeinflusst sein. Für die Datenmodellierung können verbale Beschreibungen, Entity-Relationship-Modelle oder auch UML-Diagramme für die formale Definition nützlich sein.

Erst wenn auf fachlicher Ebene Einigkeit herrscht, welche Informationen man als Linked Data zur Verfügung stellen will, kann man sich um das eigentliche Vorgehen Gedanken machen. Hierfür werden entweder existierende Vokabulare nachgenutzt oder neu formuliert. Die Prämisse beim Einsatz von Vokabularen (synonym verwendet zu Elementen in RDF Schemas oder Ontologien) im Semantic Web liegt auf einer hohen Interoperabilität. Unter Interoperabilität versteht man in diesem Zusammenhang die Fähigkeit des Vokabulars systemübergreifend eingesetzt und interpretiert werden zu können, um dadurch eine breite Nachnutzung der Daten sicherzustellen.

In den letzten Jahren ist eine Vielzahl von verschiedenen Vokabularen aus dem Boden geschossen. Trotz Open World Assumption ist keine Anwendung in der Lage, mit RDF-Daten in einer ihr unbekanntenen Beschreibung umzugehen, wodurch die tatsächliche Nachnutzbarkeit erheblich eingeschränkt ist.

Die Library Linked Data Incubator Group schlägt zwei Ansätze zur Bewältigung dieser Thematik vor [LLD11]:

1. die Nachnutzung von Linked Data Vokabular aus existierenden Standards,
2. die Definition von expliziten Beziehungen ("Alignments") bei der Spezifikation von eigenem Vokabular zu existierenden Linked Data Terms innerhalb und außerhalb der Bibliothekswelt.

Beide Ansätze haben ihre Daseinsberechtigung. Es ist immer zu untersuchen, welche Methode für die Veröffentlichung eines bestimmten Datensets die bessere ist.

Die Verwendung von bereits bestehendem Vokabular stellt dann einen optimalen Ansatz dar, wenn es bereits ein Vokabular gibt, das die zu veröffentlichenden Informationen und somit das erstellte Datenmodell vollständig abdeckt

Insbesondere im Bibliotheksumfeld zeigt sich, dass für die Abdeckung der zu veröffentlichenden Informationen oft eine Vielzahl verschiedener Vokabulare gleichzeitig zur Beschreibung einer Entität eingesetzt werden muss, um Entitäten disambiguiert darstellen zu können. Häufig ist es sogar notwendig, darüber hinaus noch eigene Elemente zu definieren. Für die Verwendung solcher Linked-Data-Repräsentationen muss eine Anwendung in der Lage sein, alle verwendeten Vokabulare zu kennen und mit ihnen umgehen zu können. In der Praxis entspricht dies zumeist nicht der Realität.

Ein schwieriges Unterfangen bei der Nachnutzung von Fremdvokabularen liegt unter anderem in der Einschätzung ihrer Stabilität. Es haben sich zwar eine Reihe von Vokabularen als stabil bewiesen – Beispiele sind das Dublin Core Metadata Element Set, oder das Friend-of-a-Friend

(FOAF) Vocabulary - allerdings gibt es bislang wenig domänenspezifische Vokabulare, von denen man dies zum augenblicklichen Zeitpunkt behaupten kann.

Wenn festgestellt wird, dass für die Abbildung des angestrebten Datenumfanges Elemente zu vieler verschiedener Vokabulare zusammen getragen werden müssen, besteht die Alternative, ein komplett eigenes Vokabular zu definieren. Wie den Empfehlungen der Library Linked Data Incubator Group zu entnehmen, ist es aber nicht allein damit getan, ein solches Vokabular zu spezifizieren. Der Maintainer ist zusätzlich in der Pflicht, Elemente – soweit wie möglich - mit Elementen aus anderen existierenden Vokabularen zu verknüpfen („Alignment“). Im Optimalfall wird ein Alignment auch nicht nur mit einem Vokabular, sondern mit mehreren vorgenommen. Der sich daraus ergebende Vorteil ist, dass Elemente, die Entsprechungen in mehreren Vokabularen gleichzeitig haben, diesen auch zugewiesen werden. Somit wäre eine Anwendung, die nur mit einem dieser Vokabulare umgehen kann, in der Lage, den für sie größtmöglichen Informationsgewinn zu erzielen.

Eine Entsprechung von Elementen muss nicht immer eine exakte Gleichheitsbeziehung sein, Elemente können auch unter- bzw. übergeordnet zueinander in Beziehung stehen, je nach Bedeutungsumfang der Elemente. Für ein Alignment gibt es daher neben der Äquivalenzausweisung weitere Konzepte wie Inverse oder Unterklasse/-property, wodurch eine differenziertere Ausweisung der Beziehung möglich ist.

Wenn Elemente, egal ob neu definiert oder bereits existent, für die Linked-Data-Repräsentation ausgewählt wurden, gilt es, eine Datentransformation der bestehenden Daten (z. B. in einer relationalen Datenbank oder in MARC 21) in die gewünschte Serialisierungsform vorzunehmen.

Nun stellt sich die Frage, wie man RDF-Daten vorhält bzw. wie man sie für die Öffentlichkeit im Web zugänglich und durchsuchbar macht. Es ist davon auszugehen, dass eine Datenrepräsentation auf der keine Suchabfragen gestellt werden kann, eine viel geringere Nachnutzung erfahren wird, als eine durchsuchbare Datensammlung. Daten die nicht über das Web durchsuchbar sind, müssen in lokalen Systemen zwischengespeichert werden, um sie prozessieren zu können. Es findet kein Datenzugriff auf dezentral organisierten Datenquellen statt, sondern es wird das Prinzip des Harvesting angewandt, bei dem alle Daten in das Verarbeitungssystem „kopiert“ werden und so eine redundante Datenhaltung entsteht.

Um die Zugänglichkeit der Daten zu garantieren, sollte nicht nur den Datenformatempfehlungen des W3C gefolgt werden, sondern auch den Standards, um auf diese Daten zuzugreifen. SPARQL²¹ ist die durch das W3C standardisierte Abfragesprache für RDF und unabhängig davon, ob die Daten in einem Triplestore oder einem anders gearteten Datenhaltungssystem abgelegt sind, sollte eine Abfrage auf RDF-Daten über diese Schnittstelle möglich sein.

Nach Berners-Lees viertem Prinzip für die Veröffentlichung von Daten im Semantic Web [BER09] besteht der Anspruch des Semantic Webs, dass Daten auch mit Daten anderer Data Provider verknüpft werden. Falls solche Beziehungen nicht existieren, ist man in der Pflicht sie zu erzeugen. Bewährte Einstiegspunkte hierfür sind beispielsweise Verknüpfungen zu DBpedia oder GeoNames.

In Abbildung 1 werden die notwendigen Schritte in Form eines Ablaufdiagramms grafisch wiedergegeben.

²¹ <http://www.w3.org/TR/rdf-sparql-query/>

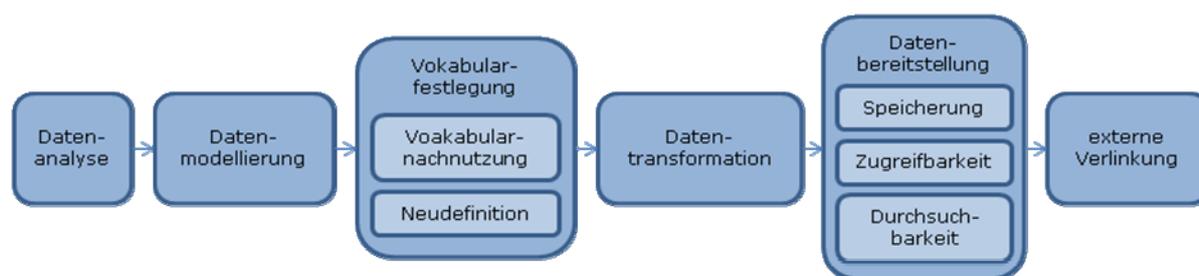


Abbildung 1 Workflow für die Veröffentlichung von Linked Data

Das hier kaskadierend dargestellte Vorgehen wird im Arbeitsschritt Datentransformation Iterationen unterzogen. D.h., die generierten Daten müssen durch eine Evaluation auf ihre Richtigkeit hin überprüft und gegebenenfalls korrigiert werden.

Auch müssen separate Workflows für die externe Verlinkung erarbeitet werden, die deren Aktualität sicherstellen. Insbesondere die Änderungen in Fremddatenbeständen können ungültige Verknüpfungen hervorrufen. Diesbezüglich empfiehlt es sich, Kooperationen mit den Datenprovidern einzugehen.

Anwendungsprofile als Basis für die Ontologieentwicklung?

Die Metadaten-Community verwendet den Begriff "Anwendungsprofil", um die Verwendung von Standards für bestimmte Anwendungen zu beschreiben. Somit lassen sich in Anwendungsprofilen Community- oder Domain-eigene Standards – oder Teile dieser Standards - kombinieren. Konkret bedeutet dies, dass Metadatenelemente aus unterschiedlichen Metadatenschemata kombiniert werden können, um die für eine spezielle Anwendung benötigten Elemente bereitzustellen. Der Ursprung von Metadatenanwendungsprofilen findet sich im Umfeld von Dublin Core.

Dieser Abschnitt diskutiert inwieweit die Konzepte auch auf die Erstellung von Ontologien gemäß dem vorgestellten Vorgehen übertragbar sind.

Das Singapore Framework

Durch das Dublin Core Singapore Framework²² werden die Komponenten eines Anwendungsprofils wie folgt beschrieben:

1. Funktionale Anforderungen (verpflichtend) beschreiben die Funktionen, für die das Anwendungsprofil entworfen wurde, sowie Funktionen, die außerhalb der Anwendung liegen. Die Funktionellen Anforderungen bilden die Grundlage für die Bewertung eines Anwendungsprofils in Hinblick auf die interne Konsistenz und in Hinblick darauf, ob das Anwendungsprofil für eine bestimmte Anwendung geeignet ist.
2. Domain-Modell (verpflichtend) definiert die Objekte, die mit dem Anwendungsprofil beschrieben werden und deren grundlegende Beziehungen zueinander. Ziel des Domain-Modells ist es, den eigentlichen Wirkungsbereich des Anwendungsprofils zu definieren. Das Domain-Modell kann in Form von Text dargestellt werden oder in Form einer formalen Darstellung, z. B. mittels der Unified Modeling Language (UML).
3. Description Set Profile (DSP) (verpflichtend) definiert die Gruppe von Metadatensätzen, die in einem Anwendungsprofil gültig sind. Das Modell des DSP wird zurzeit im Dublin Core Architecture Forum entwickelt und soll als DCMI Working Draft veröffentlicht werden. Das Modell des DSP basiert auf dem DCMI-Abstract-Model²³ und dient dazu, für die Beschreibung von Metadaten eine einfache Sprache (Constraint Language) zur Verfügung zu stellen. Das DSP bestimmt die Objekte und die Eigenschaften (Properties), die gemäß des Anwendungsprofils verwendet werden dürfen sowie die Art und Weise, in der auf einen Wert referenziert wird.

²² <http://dublincore.org/documents/singapore-framework/>

²³ <http://dublincore.org/documents/abstract-model/>

4. Nutzungsrichtlinien (fakultativ) beschreiben, wie ein Anwendungsprofil angewendet wird, wie die verwendeten Eigenschaften im Kontext der Anwendung benutzt werden sollen usw.
5. Richtlinien für die syntaktische Kodierung (fakultativ) beschreiben die für ein Anwendungsprofil spezifische Syntax und/oder spezifischen syntaktischen Richtlinien, soweit sie vorhanden sind.

Das vollständige Modell wird in Abbildung 1 dargestellt. Die Abbildung zeigt, wie sich die Komponenten eines Dublin-Core-Anwendungsprofils in Bezug auf "Domain-Standards" – Modelle und Spezifikationen, die in den Communities verbreitet sind – sowie den W3C-Standard des Resource Description Framework (RDF) verhalten.

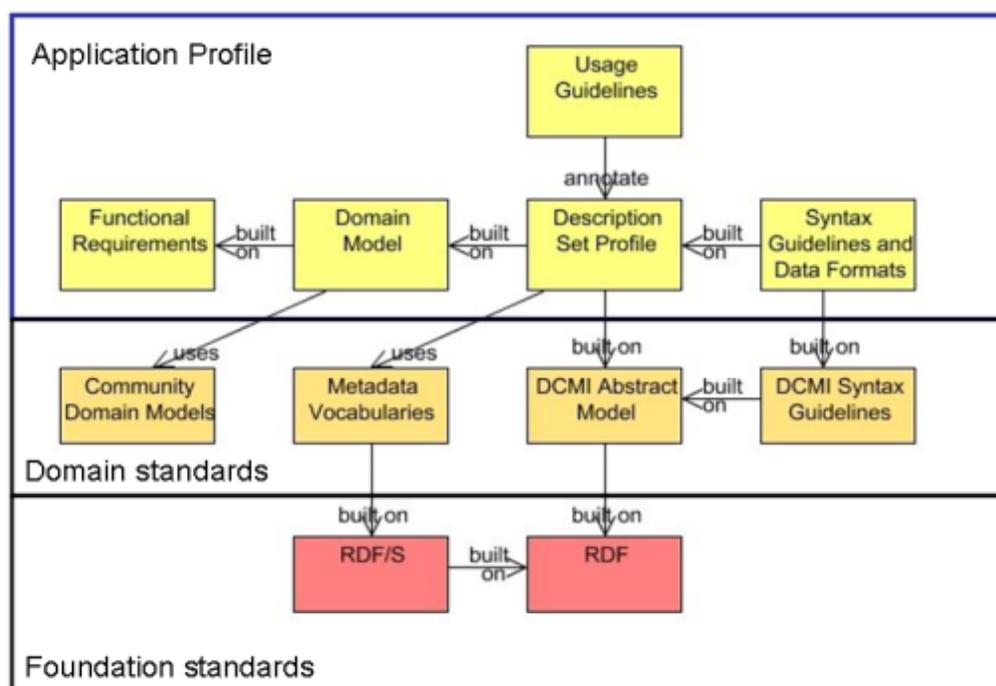


Abbildung 2 The Singapore Framework for Dublin Core Application Profiles [vgl. 22]

Anwendbarkeit des Singapore Frameworks für die Ontologiemodellierung

Das Singapore Framework erhebt den Anspruch für eine Serialisierung von Daten als Linked Data anwendbar zu sein. Dies wird nachfolgend anhand der aufgeführten Komponenten eines Anwendungsprofils betrachtet.

Die Erhebung funktionaler Anforderungen steht in der klassischen Softwareentwicklung vor allen Folgeaktivitäten. Das Ziel von funktionalen Anforderungen ist die Erstellung eines Lastenheftes. Wenn gemeinsam mit dem Auftraggeber eine Anforderungsliste erstellt werden konnte, wird diese in der Pre-Analyse-Phase auf Vollständigkeit, Widerspruchsfreiheit und Implementierbarkeit geprüft und anhand eines Fragenkatalogs abgerundet.

Sich über die funktionalen Anforderungen einer Ontologie bewusst zu sein, ist eins-zu-eins übertragbar. Die Anforderungen an die Verwendung einer Ontologie müssen im Vorfeld definiert werden, auch wenn noch nicht bekannt ist, in welchen konkreten Anwendungen die Ontologie hinterher verwendet wird. Hier ist vor allem zu beachten, dass die Anforderung besteht, den durch das W3C veröffentlichten Standards zu folgen, um eine Verwendung in Softwareprodukten, die diese Standards unterstützen, zu ermöglichen.

Die Zielsetzung des Domain-Modells lässt sich ebenfalls auf die Ontologiemodellierung übertragen, da es auch gilt, den eigentlichen Wirkungsbereich zu definieren. Letztendlich gleicht das Vorgehen dem der Datenanalyse und -modellierung, da die Objekte definiert werden, die mit der Ontologie

beschrieben und deren grundlegende Beziehungen zueinander festgehalten werden. Auch die formale Beschreibung gleicht sich.

Aussagen zum Description Set Profile (DSP) zu treffen, stellt sich als schwierig dar, da das Modell bzw. eine einfache Constraint Language zur Beschreibung von Metadaten durch das Architecture Forum noch nicht, und wahrscheinlich auch nicht mehr, bereitgestellt wird. Seit 2009 gibt es keine Fortschritte bezüglich einer Weiterentwicklung.

Die grundlegende Aussage in der Dublin Core Definition eines Description Set Profiles²⁴ lautet, dass es typischerweise Properties und Klassen aus definierten und standardisierten MetadatenVokabularen wie DCMI Metadata Terms verwendet. MetadatenVokabulare werden im Gegenzug auf Basis von RDF Vocabulary Description Language (RDF Schema oder RDFS) ausgedrückt.

Wenn man sich die grundlegende Idee von DSPs genauer betrachtet, bedeutet dies, dass Anwendungsprofile „typischerweise“ RDF-basierte Vokabulare benutzen. Ob das tatsächlich der ursprüngliche Gedanke von Anwendungsprofilen war, kann hier nicht beantwortet werden, aber der Ansatz eine neutrale Constraint Language einzuführen, lässt anderes vermuten.

Betrachtet man mit diesem Vorwissen erneut die Empfehlungen der Library Linked Data Incubator Group, so könnten Anwendungsprofile bei der Zusammenstellung verschiedener Vokabulare für die Beschreibung eines Datensets durchaus von Nutzen sein. Für die Definition eines neuen Vokabulars wäre der Ansatz wohl weniger relevant, da die zu treffenden Angaben in der Format- bzw. Ontologiespezifikation selbst hinterlegt sind. Auch werden keinerlei Aussagen getroffen, inwieweit DSPs Vocabulary Alignment unterstützen sollen.

Da bislang keine formale Sprache durch Dublin Core zur Spezifikation von DSPs vorgegeben wird, gilt es, nach Alternativen Ausschau zu halten, da es zumindest unumstritten scheint, dass man die Vokabularelemente, die man zur Serialisierung des identifizierten Datenmodells verwenden will, in einer menschen- und maschinenlesbaren Notation dokumentieren muss. Die Thematik ist zu komplex, um ihr hier in aller Vollständigkeit nachgehen zu können. Allerdings sollte durch das Architecture Forum überprüft werden, ob die Anforderungen an die Constraint Language bereits durch existierende Standards wie XSD oder OWL abgedeckt werden können.

Nutzungsrichtlinien, die beschreiben wie ein Anwendungsprofil angewendet wird bzw. wie die verwendeten Eigenschaften im Kontext der Anwendung benutzt werden, machen eine Aussage, die Interpretationsräume offen lässt. Selbstverständlich müssen einleitende Worte für die Anwendung und Nachnutzung eines jeden Metadatenformats oder Anwendungsprofils existieren. Allerdings ist die Frage, was Bestandteil der Elementspezifikation bzw. -dokumentation ist. Dieser Frage wird durch Dublin Core nicht im Konkreten nachgegangen.

Gleiches gilt für Richtlinien für die syntaktische Kodierung. Sollten spezifische syntaktische Richtlinien, wo notwendig, bereits durch den jeweiligen Vokabularanbieter definiert sein oder kann man die Ansetzung und das Vokabular voneinander losgelöst betrachten? Dublin Core lässt dies offen. Bewegungen wie die unabhängige Veröffentlichung des RDA Regelwerktextes und der RDA Element Sets zeigen jedoch, welche Probleme mit der Separation einhergehen, und dass ein enger gegenseitiger Informationsaustausch bei der parallelen Erarbeitung empfehlenswert ist. Auch muss man in Hinblick auf die Interoperabilität kritisch hinterfragen, ob eine effiziente Datennachnutzung noch möglich ist, wenn jeder syntaktische Datenkodierung nach Gutdünken vornimmt.

Die GND-Ontologie

Um eine international nachnutzbare GND-Datenrepräsentation im Web anzubieten, muss eine Serialisierungsform für diese spezifiziert werden. Für die Bereitstellung der GND wird dies durch die GND-Ontologie geleistet.

²⁴ <http://dublincore.org/documents/singapore-framework/#DSP>

Ziel der GND-Ontologie war es, eine Beschreibungssprache für die entitätenbasierte Repräsentation der GND-Daten im Semantic Web zu erschaffen. Die vorhergehende Datenanalyse und Datenmodellierung zeigten, dass keine Ontologie verfügbar ist, die die Mächtigkeit besitzt, das zugrundeliegende Datenmodell abzubilden. Von daher wurde der Entschluss gefasst, eine eigene Ontologie für Normdaten im deutschsprachigen Raum zu definieren und unter Einsatz von Vocabulary Alignment zu existierenden Vokabularen zu verlinken.

Die GND-Ontologie findet sich unter:

<http://d-nb.info/standards/elementset/gnd#>

In den nachfolgenden Abschnitten wird die Ontologie samt Modellierungsentscheidungen näher vorgestellt.

Ontologiemodellierung

Um möglichst viele prädikatenlogische Aussagen treffen zu können, fiel von Beginn an die Entscheidung gegen RDFS und Zugunsten von OWL, da durch den Einsatz von OWL detailliertere Aussagen zur Semantik der Elemente gemacht und bei Bedarf zukünftig ergänzt werden können.

Für die Spezifikation einer Ontologie in OWL muss primär festgelegt werden, welche Klassen und Properties mit welchen Constraints benötigt werden, um das identifizierte Datenmodell in OWL wiederzugeben. Entitäten werden in OWL durch Klassen repräsentiert, Eigenschaften als Datatype-Properties und Relationen zwischen Entitäten als Object-Properties.

Für alle Properties muss festgelegt sein, welche Klassen mit ihnen beschrieben werden können – Definition der Domain. Für Object-Properties ist zusätzlich die zu relationierende Klasse und für Datatype-Properties der zulässige Wertebereich bzw. Datentyp festzulegen – Definition der Range. Aus Gründen der Nachnutzbarkeit und Minimierung von Properties existieren in der GND-Ontologie Properties, die gleichzeitig mehrere Domains oder Ranges besitzen – Definition einer Union. Eine Besonderheit bilden Properties, die als Range (Wertebereich) externes kontrolliertes Vokabular verlangen. Diese werden im Abschnitt GND- Properties näher diskutiert.

Generell erhält jede Klasse bzw. Property einen Uniform Resource Identifier (URI) für die eindeutige Identifikation. Des Weiteren sind jede Klasse und Property mit Benennungen und, soweit vorhanden, Beschreibungen in deutscher und englischer Sprache versehen. Darüber hinaus wurde eine Annotation-Property eingeführt, die Klassen und Properties ihre Entsprechung im GND-MARC-Format ausdrücken lässt.

GND-Klassen

In der klassischen MARC 21 Darstellung der GND unterscheidet man individualisierte Personen, nicht-individualisierte Personen, Körperschaften, Kongresse und Veranstaltungen, Geografika, Schlagwörter und Werke.

Alle Entitäten, außer nicht-individualisierte Personen, können einen dreistelligen Entitätencode²⁵ annehmen, um ihren Typ differenzierter auszuweisen. Beispielsweise kann ein Geografikum ein Bauwerk oder Denkmal bzw. ein Land oder Staat sein. Eine Familie, die laut weitverbreiteten Datenmodellen wie FRBR und FRAD immer eine High-Level-Entität ist, versteckt sich im GND-MARC-Format unter den individualisierten Personen.

Für die GND-Ontologie wurde der Entitätenbaum als Klassenhierarchie (vgl. Abbildung 3) umgesetzt. Jede OWL-Klasse ist somit einer Oberklasse zugeordnet.

Es sei darauf hingewiesen, dass nicht-individualisierte und individualisierte Personen als Personen betrachtet werden. Das Konzept, das der GND-Entität Werk zugrundeliegt, ist historisch bedingt und deutlich vom FRBR Konzept "Work" unterschieden. Ein Werk im Sinne der GND kann

²⁵

http://www.dnb.de/SharedDocs/Downloads/DE/DNB/standardisierung/inhaltsermittlung/entitaetencodierung.pdf;jsessionid=C1D5CA2856B4094647348079D4A0F1AC.prod-worker5?__blob=publicationFile

Unterfeld \$9, da hier unter Zuhilfenahme eines Präfixes sowie des eigentlichen Relationencodes die Beziehung ausgedrückt wird. Es existieren Beziehungen zu:

- Personen und Familien im Feld 500
- Körperschaften in Feld 510
- Konferenzen und Veranstaltungen in Feld 511
- Werken in Feld 530
- Zeitangaben in Feld 548
- Schlagwörter in Feld 550
- Geografika in Feld 551

Da alle Angaben in Feld 548 keine Relationen auf echte Normdatenentitäten enthalten, sondern Literale oder Datumsangaben, wurden diese im Gegensatz zu den restlichen 5XX-Inhalten durch Datatype-Properties abgebildet.

Die nächste Herausforderung bei der Modellierung, ergab sich aus dem Umstand, dass Felder „echte“ und „unechte“ Relationen enthalten können. Eine echte Relation wird in MARC 21 im Unterfeld \$0 abgebildet und verweist auf einen Normdatensatz. Eine unechte Relation verwendet ein MARC-Datenfeld, das für die Relation vorgesehen ist, gibt als Inhalt jedoch nur ein Literal im Unterfeld \$a an.

Am Beispiel der Property „Verfasserschaft“ lässt sich die Problematik gut verdeutlichen: Verfasser eines Werkes (Domain-Klasse) dürfen individualisierte Personen (Range-Klasse Nr. 1), Familien (Range-Klasse Nr. 2) und/oder Körperschaften (Range-Klasse Nr. 3) sein, wodurch deutlich wird, dass die Verfasserschaft als Object-Property realisiert werden muss, da die Range kein Literal ist. Allerdings existieren in den GND-Daten auch unechte Relationen, wo der Bearbeiter für den Verfasser lediglich einen Namensstring und keine Verknüpfung zu einem Normdatensatz hinterlegt hat, was den Einsatz einer Datatype-Property verlangen würde.

Um diesem Problem nicht mit der Spezifikation einer Datatype- und einer Object-Property für dieselbe Eigenschaft begegnen zu müssen, wurde entschieden, unechte Relationen mittels einer Object-Property auszudrücken, die auf einen künstlich generierten Blank-Node zeigt, der die Benennung sowie eine Typisierung beinhaltet. Somit kann auch für die unechte Beziehung die Object-Property `gnd:author` verwendet werden, die wiederum vom Werk auf einen Blank Node verweist, welcher durch eine Typisierung sagt, dass es sich beispielsweise um eine Person handelt und als bevorzugten Namen der Person den angegebenen String bereitstellt.

Die zweite Unklarheit lag darin, wie mit Properties umzugehen ist, die nur Literale aus einer bestimmten Liste enthalten dürfen, wie beispielsweise Ländercodes. Eigentlich müsste die Property als Datatype-Property definiert sein und als Range eine Aufzählungsliste erhalten. Ziel bei der GND-Ontologie Modellierung war es, kontrollierte Vokabulare wie Ländercodes getrennt zu spezifizieren, um deren Nachnutzung unabhängig von der GND-Ontologie zu ermöglichen. Daraus folgt, dass Eigenschaften, deren Werte durch kontrollierte Vokabulare ausgedrückt werden, als Object-Properties umgesetzt und ihre kontrollierten Werte in SKOS Schemes abgebildet werden. Weitere Details hierzu finden sich im Abschnitt GND Value Vocabularies.

Nach der Festlegung bei welchen Angaben es sich um Datatype- und Object-Properties handelt, muss die Zuweisung von Domain und Range vorgenommen werden. In der GND-Ontologie werden vorerst nur Zuweisungen der obersten Entitätenebene (High-Level-Entität) zugelassen, da ein differenzierteres Datenmodell derzeit nicht existiert. Um die Definitionen von mehreren Properties für dieselbe inhaltliche Eigenschaft bzw. Beziehung zu vermeiden, ist es in der GND-Ontologie erlaubt einer Property mehrere Domains bzw. Ranges zuzuordnen. Um eine solche Liste semantisch korrekt darzustellen, wird `owl:unionOf`²⁷ eingesetzt.

Um eine gewisse Ordnung in die Properties hineinzubringen, wurden zusätzlich abstrakte Properties eingeführt, die als Aggregator für ähnliche Properties dienen. Zum Beispiel gibt es eine Property

²⁷ <http://www.w3.org/TR/owl-ref/#unionOf-def>

Vorgänger (Predecessor), die Preceding conference or event, Preceding corporate body, Preceding place or geographic name, Preceding subject heading und Preceding work als ihre Sub-Properties ausweist.

Darüber hinaus ist angestrebt alle Object-Properties so zu charakterisieren, dass Aussagen zu ihrer Transitivität, Symmetrie und Reflexivität hinterlegt sind.

Auch werden Kardinalitäten in die GND-Ontologie Einzug halten. Durch Restriktionen, die auch die Häufigkeit von Properties spezifizieren, ist es leichter Instanzdaten auf ihre Korrektheit zu überprüfen. Kardinalitäten werden in OWL primär als Restriktionen in Klassen hinterlegt, können aber auch für eine Multiplizität von höchstens eins für Object-Properties direkt durch eine Typisierung als Functional-Property erreicht werden.

GND Value Vocabularies

Für die Definition von Value Vocabularies bestand die Auflage SKOS-Concept-Scheme zu verwenden. In einem SKOS-Concept-Scheme kann eine Liste von SKOS-Concepts, die die eigentlichen kontrollierten Werte bilden, hinterlegt werden. Das SKOS-Concept-Scheme dient somit als Aggregator für die Instanzen der SKOS-Concepts.

Kontrollierte Vokabulare, auf die die GND-Ontologie verweist, sind:

1. Dewey-Dezimalklassifikation (DDC) - deutsch
2. Ländercodes (Geographic Area Codes)²⁸
3. Sprachencodes²⁹
4. GND-Sachgruppen³⁰
5. Geschlechtsangaben³¹
6. Koordinatentyp³²

Für die Angabe von Sprachencodes werden die durch Library of Congress spezifizierten Codes nach ISO 639-2 genutzt.

Für „Form of work and expression“ sowie für „Medium of performance“ wurden bislang keine eigenen kontrollierten Vokabulare angelegt, da es hierzu noch keine endgültigen Festlegungen innerhalb der GND gibt. Unabhängig davon muss bei beiden Properties auf Normdateneintragen verwiesen werden.

Content Type, Carrier Type und Media Type sind im GND-MARC-Format bereits vorgesehen, jedoch in der GND-Ontologie zurückgestellt, da sie bislang in den GND-Daten nicht eingesetzt werden. Nach Einführung dieser Properties wird hier ebenfalls ein kontrolliertes Vokabular bereitstehen.

Im vorangegangenen Abschnitt wurde erläutert, dass bestimmte Properties wie beispielsweise Ländercodes als Object-Properties realisiert wurden, um auf SKOS-Concepts als Wert zu verweisen. Die Herausforderung in der Modellierung bestand darin, in einer Object-Property zu definieren, dass alle SKOS-Concepts eines bestimmten SKOS-Concept-Scheme als Wertebereich dieser Object-Property auftreten können. Das bedeutet, dass bei der Range-Angabe einer Object-Property folglich eine Klasse referenziert werden muss, die ausdrückt, dass alle SKOS-Concepts eines bestimmten Schemas als Werte (Instanzen) zulässig sind. Eine solche Hilfsklasse ist notwendig, da eine direkte Einbindung des SKOS-Concept-Scheme semantisch inkorrekt wäre.

Für die praktische Realisierung stellte sich die Frage, ob eine solche Hilfsklasse Teil der GND-Ontologie sein soll. Um diesen Mechanismus einer breiteren Öffentlichkeit zur Verfügung zu stellen, wurde die Entscheidung getroffen, die Hilfsklasse direkt in das SKOS-Concept-Scheme zu integrieren.

²⁸ <http://d-nb.info/standards/vocab/gnd/geographic-area-code>

²⁹ <http://id.loc.gov/vocabulary/iso639-2>

³⁰ <http://d-nb.info/standards/vocab/gnd/gnd-sc>

³¹ <http://d-nb.info/standards/vocab/gnd/gender>

³² <http://d-nb.info/standards/vocab/gnd/type-of-coordinates>

Für Geographic Area Code wurde folgende Lösung gewählt:

```
<!-- Scheme: GND Geographic Area Codes -->
<owl:Class rdf:about="http://d-nb.info/standards/vocab/gnd/geographic-area-
          code#GeographicAreaCodeValue">
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <owl:equivalentClass>
    <owl:Restriction>
      <owl:onProperty rdf:resource="http://www.w3.org/2004/02/skos/core#inScheme"/>
      <owl:hasValue rdf:resource="http://d-nb.info/standards/vocab/gnd/geographic-area-code"/>
    </owl:Restriction>
  </owl:equivalentClass>
</owl:Class>

<!--utility class -->
<skos:ConceptScheme rdf:about="http://d-nb.info/standards/vocab/gnd/geographic-area-code">
  <dc:title xml:lang="de">GND-Ländercodes</dc:title>
  <dc:title xml:lang="en">GND Geographic Area Codes</dc:title>
</skos:ConceptScheme>

<!--example for SKOS Concept -->
<skos:Concept rdf:about="http://d-nb.info/standards/vocab/gnd/geographic-area-code#XA-DE">
  <skos:inScheme rdf:resource="http://d-nb.info/standards/vocab/gnd/geographic-area-code"/>
  <rdfs:sameAs rdf:resource="http://id.loc.gov/vocabulary/geographicAreas/e-gx"/>
  <rdfs:sameAs rdf:resource="http://id.loc.gov/vocabulary/countries/gw"/>
  <rdfs:seeAlso rdf:resource="http://d-nb.info/gnd/4011882-4 "/>
  <skos:prefLabel xml:lang="de">Deutschland</skos:prefLabel>
  <skos:prefLabel xml:lang="en">Germany</skos:prefLabel>
</skos:Concept>
```

Codebeispiel 1 GND-SKOS-Concept-Scheme mit Hilfsklasse

Das Code-Beispiel zeigt an oberster Stelle die Hilfsklasse, danach die Definition des SKOS-Concept-Scheme und abschließend ein exemplarisches SKOS-Concept. Jedes SKOS-Concept im Scheme drückt seine Schemazugehörigkeit durch `skos:inScheme` aus. Dadurch ist es der Hilfsklasse möglich eine Restriktion zu definieren, die alle SKOS-Concepts, die für `skos:inScheme` den Wert `http://d-nb.info/standards/vocab/gnd/geographic-area-code` haben, als potentielle Instanzen der Hilfsklasse definiert.

Alignment der GND-Ontologie

Das Ziel des Vocabulary Alignments der GND-Ontologie ist es, eine möglichst unkomplizierte Unterstützung für Anwendungen, die mit anderen „aligneden“ Datenformaten bzw. Ontologien arbeiten, zu erreichen. Für eine Datennachnutzung ohne Vocabulary Alignment ist vor der Verwendung in anderen Anwendungskontexten zuerst ein hoher Aufwand in Mappingaktivitäten zu investieren. Durch die Ausweisung von in der Ontologie integrierten Alignments muss dieser Aufwand nur einmal durch den Ontology-Maintainer (möglichst in Kooperation mit den Maintainer der zu verknüpfenden Ontologie) geleistet werden, um dann der Öffentlichkeit zur Verfügung gestellt zu werden.

Darüberhinaus ist damit die Hoffnung verbunden, dass das Vocabulary Alignment auch zur Steigerung der Akzeptanz und Verbreitung der GND-Ontologie beiträgt.

Im Konkreten bedeutet Vocabulary Alignment, eine Zuordnung von Classes, Properties und Values anderer Ontologien zu denen in der GND-Ontologie (und GND Value Vocabularies) vorzunehmen.

Im Falle der GND-Ontologie wurde in einem ersten Schritt eine Verlinkung zu:

- den RDA Element Sets und
- dem Friend of a Friend (FOAF) Vocabulary

vorgenommen.

Die GND-Ontologie enthält 49 Klassen, 162 Object-Properties und 54 Datatype-Properties. Insgesamt konnten 14 Alignments zu FOAF und 90 zu den RDA Element Sets vorgenommen werden.

Durch die Maintainer der GND-Ontologie wurde bereits angekündigt, dass in Zusammenarbeit mit der Community, Vocabulary Alignments zu weiteren Vokabularen und Ontologien folgen sollen, um die Interoperabilität der GND-Ontologie noch stärker zu erhöhen.

Vocabulary Alignment wird in der GND-Ontologie durch

- `rdfs:subClassOf`
- `owl:equivalentClass`
- `rdfs:subPropertyOf`
- `owl:equivalentProperty`
- `owl:inverseOf`

realisiert. Zusätzlich wurde durch die GND-Ontologie eine neue Property für das Vocabulary Alignment eingeführt - Super-property of.

URI	http://d-nb.info/standards/elementset/gnd#superPropertyOf
Label (english)	Super-property of
Label (german)	Super-Property von
Comment (english)	P1 <code>gnd:superPropertyOf</code> P2 states that P2 is a subproperty of P1. The <code>gnd:superPropertyOf</code> property is transitive.
Comment (german)	P1 <code>gnd:superPropertyOf</code> P2 sagt aus, dass P2 eine Subproperty von P1 ist. Die <code>gnd:superPropertyOf</code> Property ist transitiv.
Inverse-Property of (Alignment)	http://www.w3.org/2000/01/rdf-schema#subPropertyOf

Tabelle 1 Spezifikation der Object-Property Super-Property of

Super-Property of wurde als Bestandteil der GND-Ontologie spezifiziert, damit es nicht notwendig ist, Properties anderer Ontologien in die GND-Ontologie aufzunehmen, um von diesen aus `rdfs:subPropertyOf` zu vergeben. Durch `gnd:superPropertyOf` ist ein direktes Alignment aus der GND-Ontologie zu externen Properties möglich.

Verlinkung zu externen Datensets

Die Verlinkung zu externen Datensets berührt die Ontologiemodellierung nur indirekt. Die eigentliche Verlinkung von Entitäten geschieht zwischen Instanzdaten. Für die Ontologie selbst ist nur von Wichtigkeit, ob bereits Properties existieren, die die Art der Beziehung zwischen den Instanzdaten ausdrücken können.

Um eine Entscheidung über gegebenenfalls neu zu definierende Properties oder gar koordinierte Konzepte [DEL10] zu treffen, ist es notwendig, die zu verlinkenden Daten in Hinsicht auf die GND-Instanzdaten näher zu betrachten.

Der Workflow für die Cross-Konkordanz-Ausweisung im GND-Linked-Data-Service befindet sich derzeit noch im Aufbau, daher können hier nur die vorgesehenen Datenquellen für die Verlinkung mit der GND gelistet werden.

- Dewey-Dezimalklassifikation (DDC) - deutsch
- Virtual International Authority File (VIAF)
- Library of Congress Subject Headings (LCSH)
- Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU)
- Standard-Thesaurus Wirtschaft (STW)
- Thesaurus Sozialwissenschaften (TheSoz)
- Wikipedia
- DBpedia
- GeoNames
- MusicBrainz

Oft sind die Herkunft, Pflege und damit verbundene Aktualität von Cross-Konkordanzen schwer einzuschätzen. Aus diesem Grund ist zu klären, ob die Erstellung der jeweiligen Cross-

Konkordanzen sich in einem Dauerbetrieb befindet, sprich Änderungen der Datenbestände berücksichtigt werden und wenn ja, durch wen.

Die Dewey-Dezimalklassifikation (Dewey Decimal Classification, DDC) ist die international am weitesten verbreitete Universalklassifikation. Seit 2003 existiert eine Übersetzung der DDC ins Deutsche. Cross-Konkordanzen von Datensätzen der GND zur DDC Deutsch³³ werden wie die beiden Bestände selbst, von der Deutschen Nationalbibliothek (DNB) gepflegt. Folglich ist für diese externe Verlinkung ein hohes Maß an Aktualität und Stabilität gewahrt. Die Besonderheit an der Verlinkung der DDC Deutsch war, dass die durch das CrissCross-Projekt³⁴ geprägten Determiniertheitsgrade zur Ausweisung der unidirektionalen Beziehung von GND-Schlagwort zu DDC-Notation in der Form noch nicht als Properties für die Verwendung in Linked Data vorlagen. Um die Semantik der vier Determiniertheitsgrade entsprechend wiederzugeben, wurden in der GND-Ontologie vier Object-Properties für die Ausweisung der Beziehungen eingeführt.

Das Virtual International Authority File (VIAF) ist ein Gemeinschaftsprojekt (Laufzeit: 2003 - 2011) verschiedener Nationalbibliotheken, das von OCLC implementiert und gehostet wird. Projektziel war es, eine Kostensenkung bei der Erstellung bibliothekarischer Normdaten bei gleichzeitiger Steigerung ihres Nutzens zu erreichen, indem die nationalen Normdateien abgeglichen und verlinkt werden. Bislang geschieht dieser Abgleich nur für Personennormdaten, soll aber auf andere Entitäten erweitert werden. In wöchentlichen Abständen werden die Normdaten von 22 Normdateien geharvestet, um rein maschinelle Abgleichverfahren durchzuführen. Eine Bereitstellung der identifizierten Cross-Konkordanzen erfolgt auf der VIAF-Website³⁵. Für die Integration von Verlinkungen von GND-Daten zu VIAF muss auf den Dump zurückgegriffen werden, da VIAF kein Suchinterface bereitstellt, durch welches auf Basis eines GND-Identifiers das zugehörige VIAF-Äquivalent ermittelt werden könnte. VIAF wird ab 2012 als reguläres Serviceangebot von OCLC in Zusammenarbeit mit der Deutschen Nationalbibliothek (DNB), der Library of Congress (LOC) und der Bibliothèque Nationale de France (BnF) betrieben. Die stetige Erweiterung ist geplant. Bislang gibt es keine Informationen, ob das Abgleichverfahren gegebenenfalls um intellektuelles Eingreifen (für Fehlerkorrektur) ergänzt werden soll, da das momentan das einzige erkennbare Defizit des Service bezüglich Nachnutzbarkeit und Vertrauenswürdigkeit ausmacht.

Durch das Projekt Multilingual Access to Subjects (MACS) wurde eine Verbindung zwischen den Schlagwörtern der drei umfassenden Schlagwortnormdateien RAMEAU, Library of Congress Subject Headings (LCSH) und Schlagwortnormdatei (SWD) – jetzt Bestandteil von GND – hergestellt und damit der Grundstein für eine Sprachgrenzen überschreitende sachliche Suche in deutsch, englisch und französisch gelegt. Hierzu wurde zunächst ein Prototyp entwickelt, der es Bibliothekaren mittels eines Link-Management-Interface (LMI) ermöglicht, die Cross-Konkordanzen intellektuell zu pflegen. Nachdem die Etablierung des Service zwischenzeitlich unklar war, wird MACS jetzt im Rahmen von The European Library (TEL) betrieben, wobei TEL die Cross-Konkordanzen auch beim Retrieval in ihrem Katalog verwendet. Die Cross-Konkordanzen werden derzeit durch Mitarbeiter der Schweizerischen Nationalbibliothek (NB) und der Deutschen Nationalbibliothek (DNB) im LMI gepflegt. Für die Aktualisierung der LCSH und RAMEAU wird ein monatliches automatisches Update (Import von Vorzugsbenennung und Identifier) vorgenommen, ein entsprechender Prozess für die Integration der GND-Daten ist in Planung. Die Sachschlagwörter der GND werden bislang nicht in MACS aktualisiert. Für die Nachnutzung der Cross-Konkordanzen steht eine OAI-Schnittstelle (vgl. Beispiel³⁶) bereit. Die 1-zu-1-Mappings werden wie die der DDC Deutsch fester Bestandteil der GND-Daten. Die Beziehungen der postkombinierten Sachschlagwörter der GND zu den präkombinierten Schlagwörtern sollen nachfolgen. Bezüglich einer Nachnutzung innerhalb des GND-Linked-Data-Service muss sichergestellt werden, dass ein Updatemechanismus auch für GND-Daten in MACS realisiert wird. Des Weiteren ist abzuwägen, ob nur die 1-zu-1-Beziehungen aus

³³ <http://www.ddc-deutsch.de/>

³⁴ <http://linux2.fbi.fh-koeln.de/crisscross/>

³⁵ <http://viaf.org/viaf/data/>

³⁶

<http://www.theeuropeanlibrary.org/proxy/oai/handler?verb=ListRecords&set=macs&metadataPrefix=macs&key=8a3c5a1d6786786ec6201addc605cfb4>

dem Lokalsystem der DNB in den Linked-Data-Service aufgenommen oder auf Basis der durch OAI bereitgestellten Daten sämtliche Cross-Konkordanzen als Linked Data abgebildet werden.

Eine Verlinkung von GND-Entitäten zum Standard-Thesaurus Wirtschaft (STW) und zum Thesaurus Sozialwissenschaften (TheSoz) sollte auf Basis der in der DNB befindlichen Cross-Konkordanz-Datensätze vorgenommen werden. Die Erstellung und Pflege der Cross-Konkordanzen ist mittels eines vertraglichen Abkommens mit dem Leibniz-Institut für Sozialwissenschaften (GESIS) als Maintainer von TheSoz und der Zentralbibliothek für Wirtschaftswissenschaften (ZBW) als Maintainer vom STW mit der DNB geregelt. Sobald ein neuer Deskriptor durch einen DNB-Mitarbeiter in der GND angelegt wird, der ein Äquivalent im STW oder im TheSoz besitzen könnte, wird nach einem solchen im jeweiligen Online-Katalog gesucht, um die gegenseitige Beziehung in Form eines Cross-Konkordanz-Datensatzes als Bestandteil der zentral gehaltenen GND zu hinterlegen. Ein solcher Cross-Konkordanz-Datensatz enthält jeweils die eindeutigen Identifier der zueinander gehörenden Normdateneintragen sowie deren Vorzugsbenennungen. Wenn an der ZBW oder bei GESIS ein neuer Deskriptor angelegt wird, können die dortigen Mitarbeiter über eine Online-Schnittstelle direkt im GND-Bestand suchen und gegebenenfalls Änderungen vornehmen. Das heißt auch diese Mitarbeiter würden Cross-Konkordanz-Datensätze direkt in die GND einpflegen können, wodurch die Aktualität der Daten immer gewährleistet ist. Cross-Konkordanz-Datensätze können für die Verwendung in eigenen Systemen wie alle anderen Datensätze der GND über OAI-PMH geharvestet werden. Selbst für die Nachnutzung der Daten durch die ZBW und GESIS muss auf diesen Mechanismus zurückgegriffen werden. Auch wenn die Cross-Konkordanzen zukünftig parallel als Linked Data bereitstehen, wurde durch die Community bereits angeregt, diese Informationen darüber hinaus in einem einfachen Format wie beispielsweise BEACON³⁷ zur Verfügung zu stellen.

Durch Wikipedia-Autoren können GND-Identifier unter Verwendung einer Vorlage³⁸ in einzelnen Artikeln eingebunden werden. In der Vergangenheit wurde dies primär für Personenartikel gemacht, um Äquivalente in der PND, aber auch in VIAF oder den Normdaten der Library of Congress nachzuweisen. Die externe Verlinkung wird durch nicht-professionelle Benutzer, aber mit einer erstaunlich hohen Qualität erstellt. Fehler und Wartungsbedarf können gemeldet, aber auch die Information über das Nicht-Vorhandensein eines individualisierten Personeneintrags kann hinterlegt werden. Für den GND-Linked-Data-Service ist es ebenso von Interesse die Rückverlinkung auf Wikipedia (z. B. unter Verwendung der foaf:page Property) auszuweisen. Es muss allerdings geklärt werden, wie die Cross-Konkordanzen mit hoher Aktualität bereitgestellt werden können. Hierfür gibt es mehrere Möglichkeiten: Sämtliche PND-Verknüpfungen der deutschsprachigen Wikipedia stehen im BEACON-Format zum Download³⁹ bereit. Die Datei wird einmal täglich aktualisiert. Des Weiteren ist eine Extraktion der Personendaten⁴⁰ also der PND-Links aus einer SQL-Datenbank oder direkt aus dem XML-Dump möglich. Aus Performanzgründen empfiehlt es sich, für den GND-Linked-Data-Service alle interessanten Links aus dem jeweils aktuellen Dump selbst zu extrahieren. Interessante Links wären selbstverständlich alle deutschsprachigen Artikel in denen ein GND-Identifier hinterlegt ist, sowie das englischsprachige Wikipedia-Pendant dazu. Die Extraktion sollte nach Möglichkeit mittels eines Skripts realisiert werden. Um nicht auf neue Gesamtabzüge der Wikipedia warten zu müssen, kann die Dumpverarbeitung auch nur initial durchgeführt werden und anschließend auf den RSS-Feed⁴¹ zurückgegriffen werden.

DBpedia ist ein Gemeinschaftsprojekt der Universität Leipzig, der Freien Universität Berlin und OpenLink Software, um strukturierte Informationen aus Wikipedia zu extrahieren und Web-Anwendungen zugänglich zu machen. DBpedia ermöglicht es weiterhin, diese Daten mit Informationen aus anderen Web-Anwendungen zu verbinden. Als Standard für die Daten wird das Resource Description Framework (RDF) benutzt. Als Quelle wurden unter anderem die englische, deutsche, französische, spanische, italienische, portugiesische, polnische, schwedische,

³⁷ <http://gbv.github.com/beaconspec/beacon.html>

³⁸ <http://de.wikipedia.org/wiki/Vorlage:Normdaten>

³⁹ <http://toolserver.org/~apper/pnd.txt>

⁴⁰ <http://de.wikipedia.org/wiki/Hilfe:Personendaten/Datenextraktion>

⁴¹ <http://dumps.wikimedia.org/dewiki/latest/dewiki-latest-pages-articles.xml.bz2-rss.xml>

niederländische, japanische, chinesische, russische, finnische und norwegische Wikipedia verwendet. DBpedia erlaubt erstmalig komplexe Anfragen auf den in Wikipedia befindlichen Inhalten zu stellen. Durch die DBpedia-Live System Architecture ist eine Synchronisation mit Wikipedia gegeben. Bei Änderungen innerhalb von Wikipedia werden diese mittels OAI-PMH von DBpedia geladen und durch Extraktionswerkzeuge in die interne Triple-Repräsentation überführt. Ausgewertet werden u. a. Infoboxes, Labels, Abstracts, sprachübergreifende Verlinkung und Angaben zur Disambiguierung. Alle in DBpedia enthaltenen Daten können für jeden zugänglich über einen SPARQL-Endpoint⁴² abgefragt werden. Falls Änderungen an einem Artikel und somit an der resultierenden Linked-Data-Repräsentation vorgenommen wurden, werden diese durch ein Change-Tracking ausgewiesen. DBpedia ist derzeit der Datenbestand mit den meisten Verlinkungen auf externe Bestände⁴³. DBpedia als Teil des W3C Linking Open Data Community Project⁴⁴ erbrachte beachtenswerte Bemühungen, um Interlinks zwischen verschiedenen Datensets auszumachen und als Bestandteil von DBpedia zu veröffentlichen. Mittels des Silk Link Discovery Framework können außerdem auf einfachem Wege Links zu DBpedia durch Datenanbieter hinzugefügt werden. Die Cross-Konkordanzen müssen dafür in der Silk Link Specification Language (Silk-LSL)⁴⁵ abgebildet sein. Im DBpedia Download-Bereich⁴⁶ stehen neben dem gesamten DBpedia-Datenbestand auch spezielle Datasets bereit. Für die Cross-Konkordanzen von DBpedia zu GND ist aktuell nur das Datenset mit den PND-Identifiern relevant. Allerdings könnten über die zugehörigen DBpedia-URIs eine Reihe weiterer Cross-Konkordanzen zur jeweiligen Normdateneintragung gefunden werden. Da die Quelldaten von Wikipedia stammen, sollte deren Qualität folglich ähnlich sein, allerdings ist nicht deutlich ersichtlich, in welchen Abständen die Synchronisation mit Wikipedia stattfindet.

GeoNames als ein Aggregator international agierender öffentlicher Einrichtungen (Ämter für Vermessung, Statistik etc.) weist rund 8 Millionen Einträge zu geografischen Ressourcen nach. In einem ersten maschinellen Versuch konnten über 41.000 Orte und Länder der GND den in GeoNames befindlichen Eintragungen eindeutig zugeordnet werden. Tests haben sogar gezeigt, dass ein Großteil der darüber hinaus in der GND aufgeführten Geografika (z. B. Bauwerke) ebenfalls eine Zuordnung finden kann.

Durch ein weiteres maschinelles Abgleichverfahren konnten fast 46.000 Künstler der GND in MusicBrainz nachgewiesen werden. Die Verknüpfung zu MusicBrainz nimmt eine besondere Rolle ein, da die Plattform auch granular erschlossene Discografienachweise führt.

Die Abgleichverfahren zu GeoNames und MusicBrainz sind jeweils zu einem konkreten Zeitpunkt entstanden und vernachlässigen Änderungen im Bestand der GND sowie in denen der zu verknüpfenden Datensets. Dadurch entsteht ein verhältnismäßig hohes Risiko der Ausweisung falscher Cross-Konkordanzen, auch da nie eine intellektuelle Überprüfung stattgefunden hat.

Zusammenfassung und Ausblick

Am Beispiel der GND wurde in diesem Bericht verdeutlicht, wie eine Institution ihre über Jahrzehnte aufgebauten Datenbestände Teil des Semantic Web werden ließ. Es wurde versucht, aufkommende Fragen und Entscheidungsfindungsprozesse darzustellen, um mit der Arbeit anderen Interessierten den Weg zur Veröffentlichung ihrer eigenen Daten zu erleichtern. Andererseits soll das Dokument Anwendern der GND-Linked-Data-Repräsentation ein Verständnis für ihren Aufbau und ihre Verwendungsweise bieten.

Insbesondere mit der GND-Ontologie wurde ein erster aber großer Schritt in Richtung „deutschsprachige Normdaten als fester Bestandteil des Semantic Web“ gemacht. Es wurde verdeutlicht, wie bei der Erstellung der GND-Ontologie vorgegangen wurde und warum es sich

⁴² <http://live.dbpedia.org/sparql>

⁴³ <http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets>

⁴⁴ <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

⁴⁵ http://www.assembla.com/wiki/show/silk/Link_Specification_Language

⁴⁶ <http://wiki.dbpedia.org/Downloads>

nicht empfahl existierende Vokabulare und Ontologien nachzunutzen und somit nur einen Teilausschnitt des Datenmodells abzubilden. Darüber hinaus wurde die Notwendigkeit des Vocabulary Alignments hervorgehoben, um die resultierende Serialisierung auch zu Datenrepräsentationen in anderen Dialekten kompatibel zu halten.

Durch die GND-Ontologie soll es Reasonern zukünftig möglich sein, Instanzdaten auf ihre semantische Korrektheit hin zu überprüfen, aber auch auf Basis getroffener Aussagen neue Informationen abzuleiten. Durch den gewählten Ansatz des Vocabulary Alignments steht die Verarbeitung der Linked-Data-Repräsentation auch Anwendungen und Reasonern offen, die nicht mit dem GND-Format an und für sich, sondern ausschließlich mit den verknüpften Formaten umgehen können. Aus diesem Grund wird das Vocabulary Alignment der GND-Ontologie in der weiteren Entwicklung der GND-Ontologie auch eine wichtige Rolle spielen.

Ebenso muss die Weiterentwicklung bezüglich der Integration von neuen Klassen und Properties kontinuierlich betrieben werden. Hierfür gilt es vor allem, innerhalb der Community aufkommende Anforderungen zu erkennen und über eine Aufnahme in der GND-Ontologie zu entscheiden. Dies soll auch Dinge betreffen, die nicht Bestandteil des GND-MARC-Formats sind. Der ausschlaggebende Punkt sollte ein Konsens über den Bedarf innerhalb der Anwender-Community sein - Mehrheitsprinzip. Die DINI AG KIM wird sich hierbei weiterhin engagiert einbringen.

Ein noch offener Punkt ist, wie Änderungen an der GND-Ontologie verwaltet und kommuniziert werden. Allgemeine Praxis ist, die Verlinkung auf die vorherige und nachfolgende bzw. die jeweils aktuelle Version sowie ein Change-Tracking anzubieten.

Für die permanente Gewährleistung der Vertrauenswürdigkeit müssen perspektivisch Informationen zur Provenienz in die GND-Daten Einzug finden. Dabei ist es nicht zwingend notwendig, dass Elemente zum Provenienznachweis Bestandteil der GND-Ontologie werden. Falls es Fortschritte diesbezüglich innerhalb von OWL oder in anderen Standards gibt, sollten entsprechende Elemente übernommen und in den Instanzdaten ausgewiesen werden

Ein letzter Punkt der anzumerken ist, ist die fehlende, aber durch das W3C empfohlene Abfragbarkeit der Daten durch SPARQL. Für einen Suchanfragen-gestützten Datenbezug muss derzeit auf Search/Retrieval via URL (SRU) zurückgegriffen werden. Bei dieser Art des Datenbezuges werden aber lediglich Datensätze ausgewählt und kein „Antwort-Graph“ auf Statement-Level erzeugt.

Unabhängig von den noch offenen Aspekten kann festgehalten werden, dass mit der GND als Linked Data Anwendungen und Suchmaschinen im Semantic Web eine zeitgemäße Art der Datenbereitstellung eröffnet wurde. Jetzt gilt es, auch Anwendungen auf Basis der GND-Daten zu entwickeln.

Literatur

[BEH11] Behrens-Neumann, Renate; Pfeifer, Barbara: Die Gemeinsame Normdatei - ein Kooperationsprojekt. Dialog mit Bibliotheken, 2011/1, S.37-40, <http://d-nb.info/101606361X/34> (letzter Zugriff 9. Juli 2012), 2011.

[BER09] Berners-Lee, Tim: Linked Data Design Issues. <http://www.w3.org/DesignIssues/LinkedData> (letzter Zugriff 9. Juli 2012), 18. Juni 2009.

[DEL10] Dellschaft, Klaas: Darstellung der CrissCross-Mappingrelationen im Rahmen des Semantic Web. Universität Koblenz-Landau, 2010.

[DEL11] Dellschaft, Klaas; Hachenberg, Christian: Repräsentation von Wissensorganisationssystemen (KOS) im Semantic Web: Ein Best Practice Guide. Universität Koblenz-Landau, <http://www.kim-forum.org/Subsites/kim/SharedDocs/Downloads/DE/Berichte/repraesentationVonKosimSemanticWeb> (letzter Zugriff 9. Juli 2012), 30. September 2011

[LLD11] Baker, Thomas et al.: Library Linked Data Incubator Group Final Report. <http://www.w3.org/2005/Incubator/ld/XGR-ld/> (letzter Zugriff 9. Juli 2012), 25 October 2011.