Hochschule für Technik, Wirtschaft und Kultur Leipzig

Fakultät Informatik, Mathematik und Naturwissenschaften

Masterstudiengang Medieninformatik

Projektarbeit

Projektdokumentation im Modul Semantic Web

Semantische Erweiterung von Projektdokumentation aus dem SS15 mit dem Metadaten-Vokabular DCAT

Eingereicht von: Mario Reyes Napoles

Leipzig 31. Juli 2016

Erstprüfer: Prof. Dr. rer. nat. Riechert

Inhaltsverzeichnis

1. Fragestellung

Intention dieser Projektarbeit ist der Frage nachzugehen, wie die verwendeten Datensätze zum Modul Semantic Web aus dem SS2015 zu analysieren sind, um diese für eine vereinfachte Suche semantisch mit dem DCAT Vokabular aufbereiten zu können. Mit einer SPARQL-Anfrage sollen alle relevanten Information zu den Datensätzen und den dazugehörigen Projekten abrufbar gemacht werden. Als Grundlage dienen dabei die bereitgestellten Dokumentationen aus dem vorherigen Semester, welche als PDF-Format vorliegen. Im Rahmen dieser Projektarbeit werden zunächst die verwendeten Datenquellen und Projekte auf ihre Qualität hin überprüft. Um die semantische Qualität zu erhöhen, werden die einzelnen Datensätze mit dem DCAT Vokabular angereichert. DCAT ist ein RDF Vokabular mit denen Datenkataloge und Datensätze beschrieben werden können, um eine erhöhte Auffindbarkeit zu ermöglichen. Der Mehrwert der aggregierten DCAT Metadaten schaffen als Manifest-Datei eine erleichterte digitale Aufbewahrung. DCAT wird benötigt, um die Interoperabilität der Datensätze in Bezug zu den Projekten und ihren Autoren für zukünftige Kursteilnehmer verfügbar zu machen. Die im Laufe der Projektumsetzung erstellten Quell- und RDF-Dateien sind in dem folgenden GitHub-Repository untergebracht: https://github.com/MReyNap/Projekt_SemanticWeb.

2. Datenquellen

2.1. Analyse und Vorverabeitung der Datenquellen

Die vorliegenden Datenquellen liegen als PDF-Format in 23 Projektdokumentationen vor, deren Verwendung sich auf eine jeweils spezifische Fragestellung in den Projekten beziehen. Insgesamt werden im Rahmen dieses Projekts 88 Datensätze (Siehe Datei *Datensätze_SS15*) in unterschiedlichen Dateiformaten analysiert und im weiteren Verlauf mit dem DCAT-Vokabular beschrieben. Jeder Datensatz wurde in Bezug zu seinem Projekt auf seine Qualität hin überprüft. Die dabei verwendeten Datensätze werden dabei in den vorliegenden Projektdokumentationen durch folgende Attributen, wie in Abbildung 1 beispielhaft dargestellt, beschrieben:

- Link
- Datenformat
- Schnittstelle
- Lizenz
- Open Data

Link	http://leipzig-data.de/Data/
Datenformat	RDF, OWL
Schnittstelle	SPARQL, Linked Data
Lizenz	CC0 1.0 Universal
Open Data	⋆⋆⋆⋆⋆

Abbildung 1: Datensatz aus Projekt

Im Rahmen dieses Projekts sollen die einzelnen Datensätze in ihren jeweiligen Datenkatalogen analysiert und auf die von den Projektautoren definierten Bewertungen hin überprüft. Des Weiteren werden alle Datensätze mit weiteren Attributen verknüpft, um eine semantische Erweiterung durch das DCAT Vokabular zu gewährleiten. Die folgenden Attribute werden dahingehend für jeden einzelnen Datensatz zusätzlich integriert:

- Dataset ID
- Projekt
- Person
- Dataset Title
- Dataset Description
- Dataset Keyword
- Dataset Language
- Dataset Landing Page
- Distribution Access URL

Anschließend werden alle Datensätze mit allen dazugehörigen Anreicherungen in eine CSV-Datei übertragen, welche unter *Datensätze_SS15* auffindbar ist.

2.2. Datenbeschreibung und Verlinkung

Da die verschiedenen Projekte aus dem SS15 in verschiedene Datensätzen und in unterschiedlichen Formaten vorliegen, wird sich ausschließlich auf die Datenbeschreibung und Verlinkung nach dem DCAT- Vokabular konzentriert. Beispielhaft zu nennen sind SPARQL, JSON, CSV oder HTML-Dokumente, deren

Inhalte keiner Kontrolle unterliegen. Das extrahieren mit anschließender Analyse würde den Rahmen dieser Arbeit sprengen, weshalb sich ausschließlich dieses Projekt um eine Anpassung der Datensätze mit DCAT[1] bemüht. Es ist einzig wichtig, dass die Originalquellen bereitgestellt werden und die Daten ordnungsgemäß beschrieben sind. DCAT ist ein Vokabular, um den Austausch von Datenkatalogen und Datensätzen via Metadaten zu ermöglichen. Die Verfügbarkeit der Daten im Web werden durch das RDF-Format nach den Prinzipien von Linked Data ermöglicht. Alle vorliegenden Datensätze werden nach den RDF-Modell verlinkt, wie es von W3C empfohlen wird. Das Vokabular hat, wie in Abbildung 2 dargestellt, eine Struktur welche die Beschreibung der 88 Datensätze vollumgänglich gestattet.

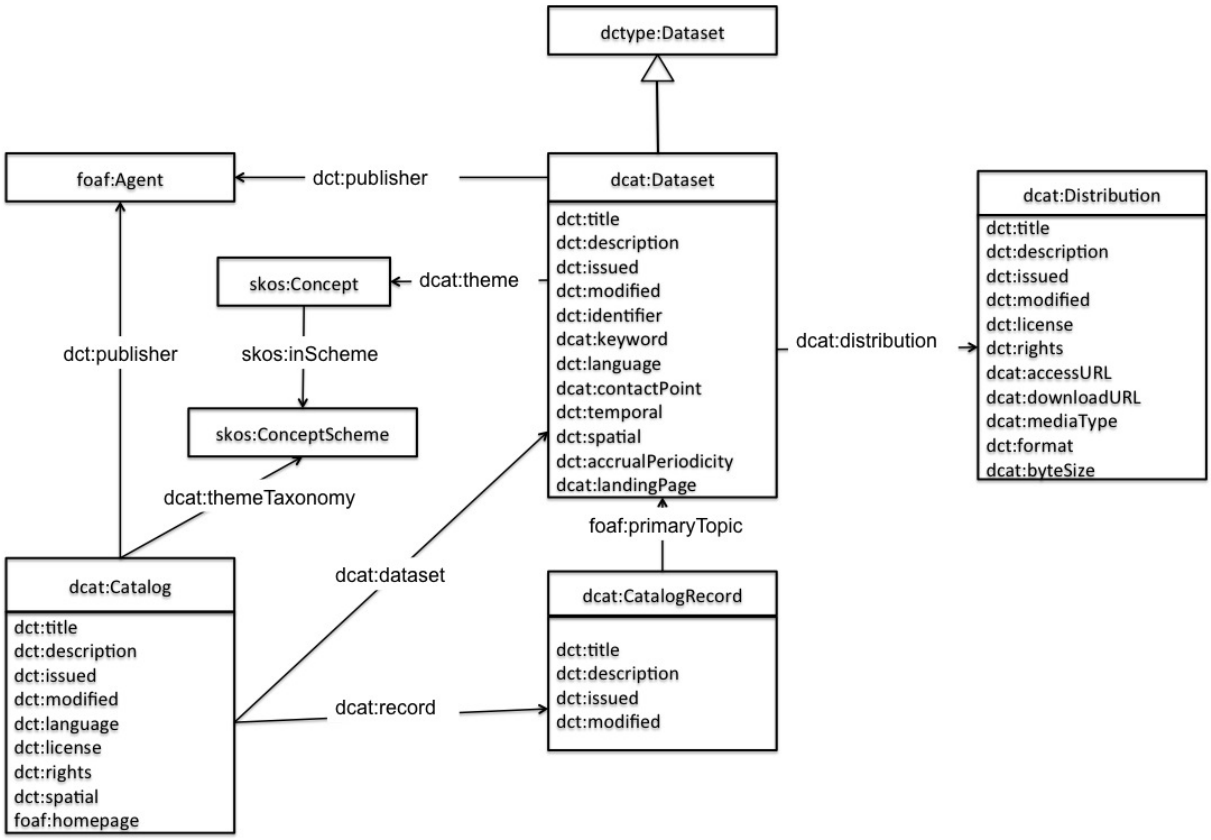


Abbildung 2: DCAT-Struktur-Modell

Im Rahmen dieser Projektarbeit werden die einzelnen Datensätze mit dem **dcat:Dataset** semantisch erweitert und mit deren jeweiligen Autoren **foaf:Agent** verlinkt. Der **dcatCatalog** wird über die Property *dcat:dataset* mit den Datensätzen verlinkt und beschreibt dabei den Datenkatalog, welche sich ausschließlich auf die jeweiligen Projekte des jeweiligen Semesters beziehen. Darüber hinaus wird die

[1] https://www.w3.org/TR/vocab-dcat/ – 29.07.2016

Verfügbarkeit der Datensätze mit der Ressource **dcat:Distribution** beschrieben, welche über die Property *dcat:distribution* mit den Datensätze verlinkt ist. Die jeweiligen Datensätze aus der Datei Datensätze_SS15 werden somit mit dem Vokabular Dcat semantisch erweitert, deren Beschreibung als RDF-Format in der *RDF-Turtle-Datei* zu finden ist.

3. Semantische Aufbereitung der Datensätze

Um einen abfragbaren RDF-Datensatz zu generieren wird auf das Programm OpenRefine (ehem. GoogleRefine) zurückgegriffen. Openrefine ist ein Open Source-Tool zur Weiterverarbeitung von Rohdaten. Für die Erzeugung von RDF-Formaten ist eine RDF-Extension nötig, welche nicht von Haus aus in OpenRefine verfügbar ist. Im Folgenden wird erläutert, wie die Datensätze in der CSV-Datei mit dem DCAT Vokabular erweitert und das in Kapitel 2.2 angedeutete DCAT-Struktur-Modell konzipiert wird.

Dafür wird zunächst ein Projekt in OpenRefine angelegt, in denen die aus den Projekten erhobenen Datensätze als CSV-Datei in OpenRefine importiert werden. Openrefine selbst ist in der Lage, viele Konfigurationen aus die Rohen Tabellen anzuwenden. Aus der importierten Tabelle kann ein RDF-Skelett definiert werden, in denen Subjekte, Prädikate und Objekte zugeordnet werden. Diese können später bei der SPARQL-Abfrage aufgerufen werden. Abbildung 3 - 5 zeigt das RDF-Skellet, mit denen die Datensätze beschrieben werden.

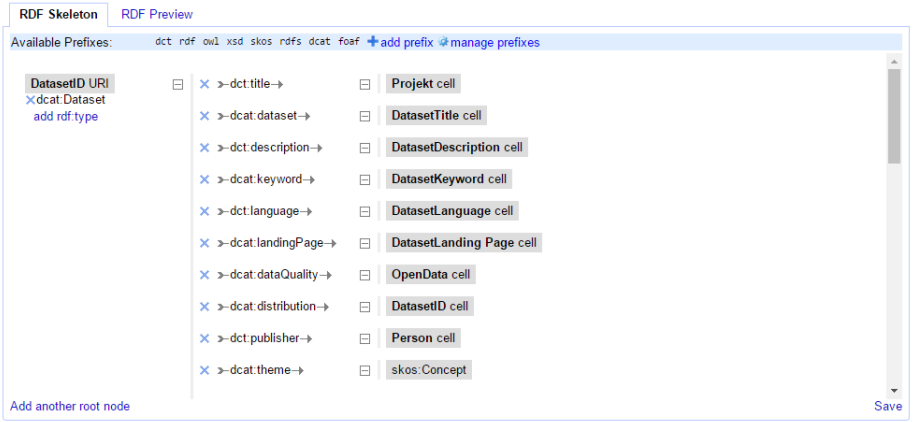


Abbildung 3: RDF Skelett dcat:Dataset

Die verschiedenen Datensätze besitzen eine individuelle DatasetID, welche als Subjekt deklariert sind. Die in der CSV befindlichen Spalteneinträge sind die Objekte, mit den dazugehörigen Propertys.

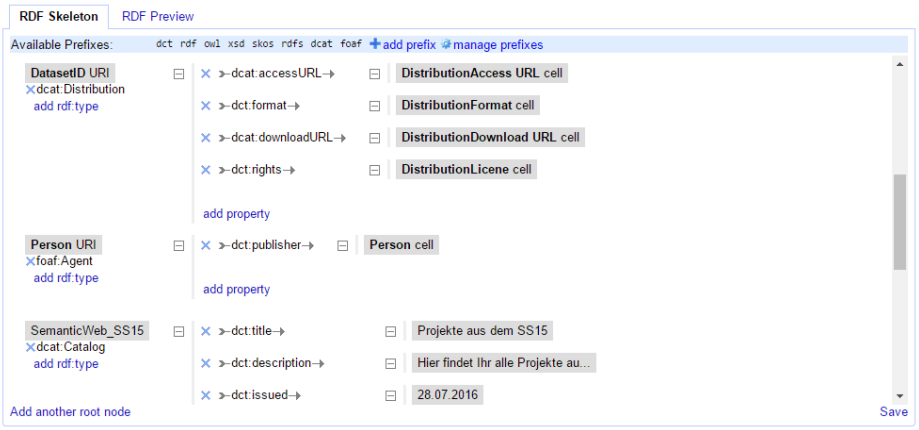


Abbildung 4: RDF Skelett dcat:Distribution, foaf:Agent, dcat:Catalog

Zur Beschreibung der Distributionsmöglichkeiten wird auf das zuvor bestimmte Subjekt verlinkt. Die Beschreibung der Distribution enthält Propertys, welche zuvor in der Analysephase recherchiert und in die CSV-Daei übertragen wurde. Die Autoren der jeweiligen Projekte sind ein eigenes Subjekt, welche den Namen des Projektbearbeiters beinhalten. Als Data Catalog wurde ein Subjekt mit dem Namen SemanticWeb_SS15 bestimmt, da sich diese Projektarbeit ausschließlich auf die Datensätze aus dem Sommersemester 2015 beziehen. Das Subjekt SemanticWeb_SS15 beschreibt alle relevante Informationen zum Modul Semantic Web an der HTWK Leipzig.

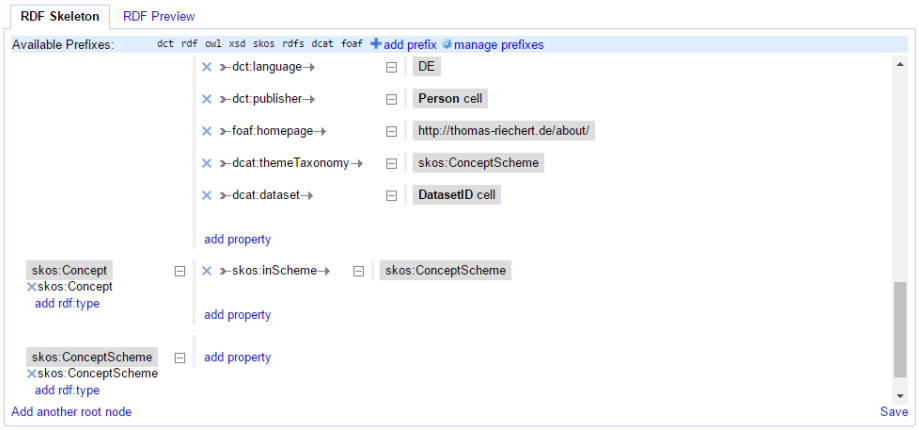


Abbildung 5: RDF Skelett skos:Concept, skos:ConceptScheme

Wie in Abbildung 3 - 5 dargestellt, werden über die Menü Knotenpunkte die Datensätze als Subjekte erstellt, deren „Propertys“ (Prädikate) zugeordnet werden. Für die Prädikate können wiederum neue Präfixe erstellt werden, die anschließend auf Attribute, die in der Tabelle hinterlegt sind, verweisen. Nach der Definition der RDF-Struktur kann das RDF-File via „RDF Preview“ vorab zur Überprüfung angezeigt werden. Nachdem keine Fehler mehr auftreten sind, ist die fertige Datei in einem gewünschten Format zu exportieren. Im Rahmen dieser Projektarbeit erfolgt die Ausgabe der strukturierten RDF-Daten als Turtle-File, welches als .txt und .ttl in der *Turtle-RDF-Datei* beigefügt ist.

4. Triplestore

Als Triplestore kam Apache Jena Fuseki zum Einsatz, da dieser Triplestore eine einfache und intuitive zu bedienende Weboberfläche hat. Darin wurde die zuvor erzeuge *RDF-Turtle-Date*i geladen und standen so für Abfragen via SPARQL zur Verfügung.

5. Datenabfrage

Die Abfrage der Daten zu den jeweiligen Projekten werden über einen SPARQL-Ausdruck, wie in Abbildung 6 dargestellt, ermöglicht.

```
PREFIX mreyna: <http://www.imn.htwk-leipzig.de/~srau/semanticweb/ontologie#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX fo: <http://www.w3.org/1999/XSL/Format#>

SELECT ?Projekt ?DatasetID ?Person ?DatasetDescription

WHERE {
  ?dataset a dcat:Dataset ;

   dcat:dataset ?DatasetID ;
   foaf:name ?Projekt ;
   dct:description ?DatasetDescription ;
   dct:publisher ?Person .
   }
```

Abbildung 6: SPARQL Anfrage

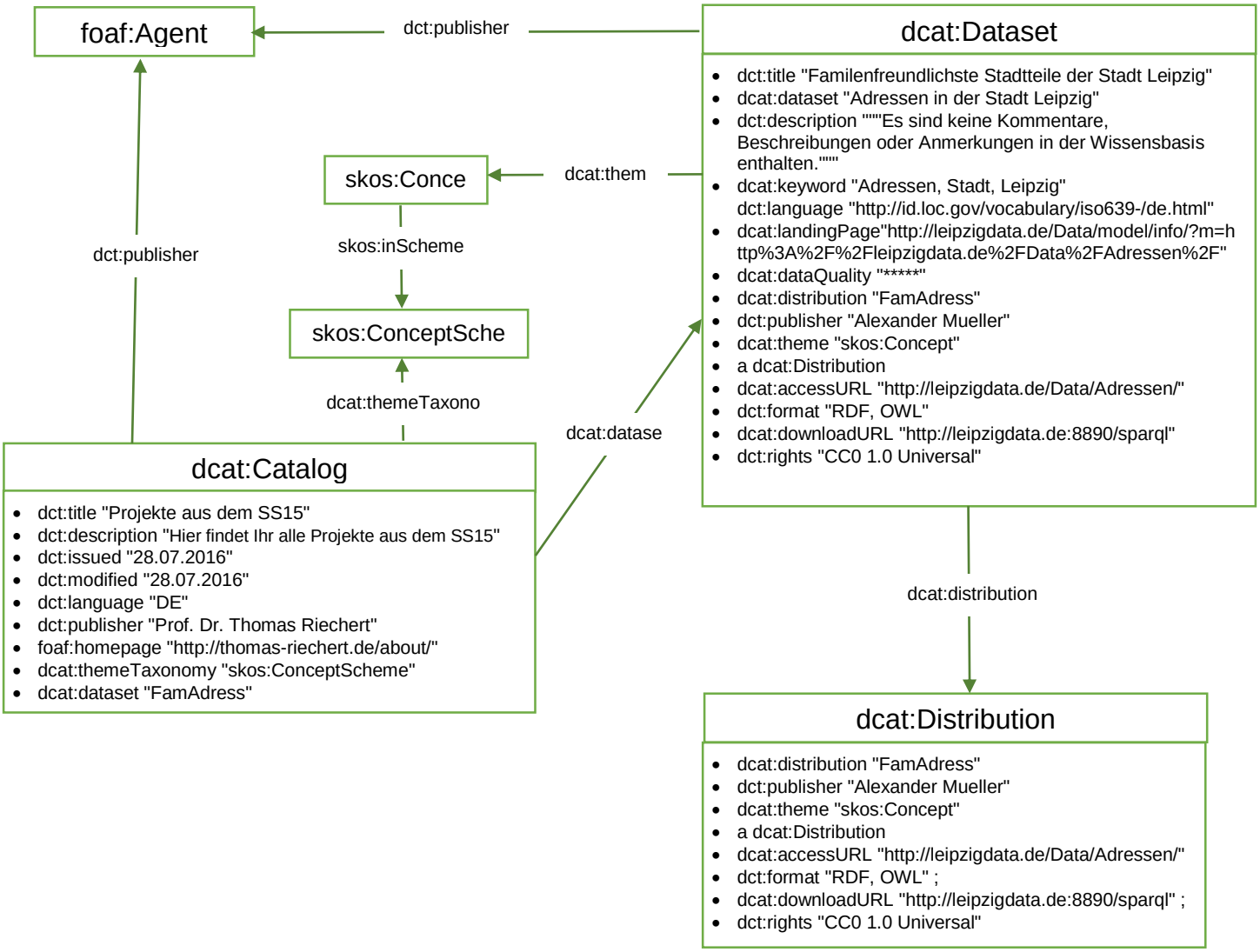
Mit Hilfe von SPARQL wurde eine Anfrage an die Forschungswissensbasis formuliert. Die Abfrage erstellt, wie in Abbildung 7 dargestellt, eine neue Tabelle in denen die Datensätze (DatasetTitle) in Relation zu den Beschreibungen (DatasetDescription) und den jeweiligen Projekten und den Personen (Autoren) abgebildet werden.

	Projekt	DatasetID	Person	DatasetDescription
1	" Familenfreundlichste Stadtteile der Stadt Leipzig"	"Adressen in der Stadt Leipzig"	"Alexander Mueller"	"Es sind keine Kommentare, Beschreibungen oder Anmerkungen in der Wissensbasis enthalten."
2	" Familenfreundlichste Stadtteile der Stadt Leipzig"	" Schulen der Stadt Leipzig,"	"Alexander Mueller"	"ohne beruflische Schulen und Förderschulen. 2014-07-20:Abgleich mit den Stadtseiten."
3	" Familenfreundlichste Stadtteile der Stadt Leipzig"	" Schulen der Stadt Leipzig,"	"Alexander Mueller"	"ohne beruflische Schulen und Förderschulen. 2014-07-20: Abgleich mit den Stadtseiten."
4	" Familenfreundlichste Stadtteile der Stadt Leipzig"	" Stadtbezirke und Ortsteile"	"Alexander Mueller"	"Es sind keine Kommentare, Beschreibungen oder Anmerkungen in der Wissensbasis enthalten."
5	" Familenfreundlichste Stadtteile der Stadt Leipzig"	"Adressdatenbank auf der Website der Stadt Leipzig "	"Alexander Mueller"	"Es sind keine Kommentare, Beschreibungen oder Anmerkungen in der Wissensbasis enthalten."
6	"Regattatätigkeit der deutschen Segelvereine in Abhängigkeit der Gewässergröße"	"Liste von Seen in Deutschland/"	"Anja Rommel"	"Der Artikel „Liste von Seen in Deutschland/“ existiert in der deutschsprachigen Wikipedia nicht. Du kannst den Artikel erstellen (Anleitung). Wenn dir die folgenden Suchergebnisse nicht weiterhelfen, wende dich bitte an die Suchhilfe oder suche nach „Liste von Seen in Deutschland/“ in anderssprachigen Wikipedias."

Abbildung 7: SPARQL Ausgabe

6. Interpretation der Ergebnisse

Das Ergebnis dieser vorliegenden Projektarbeit hat gezeigt, dass die beschriebenen Datensätze aus dem Sommersemester 2015 mit dem DCAT Vokabular erfolgreich erweitert werden konnten. Zukünftige Projekte können somit um Datenkataloge aus vergangenen Semestern erweitert werden. Aus folgender grafischen Visualisierung soll beispielhaft am ersten Projekt abschließend aufgezeigt werden, wie das DCAT-Struktur-Modell auf die jeweiligen Datensätze angewendet werden.



Abbildungsverzeichnis