Die nachhaltige Bewahrung einer Forschungsdatenbank durch Linked Data. Laut welchem Vokabular?

Francesco Gelati

Linked Data bietet die Möglichkeit nicht nur normkonforme Datensätze zur Verfügung zu stellen sonder auch nachhaltig und langfristig Metadaten zu bewahren, wenn diese zu Linked Data Vokabularen verknüpft sind. Aber welches Vokabular soll man nutzen? Der Anwendungsfall einer Forschungsdatenbank des Instituts für Zeitgeschichte München - Berlin zeigt die Vorteile, Wikidata für die Erstellung von Linked-Data-Beständen zu benutzen und Wikidata/DBpedia als Backup-Vokabular zu verwenden.

Normgerechte Metadaten als Linked Data

Nur normgerechte und FAIR[1]-Metadaten sind wert, langfristig gespeichert zu werden. Zugleich erscheinen aber immer wieder neue Standards! Das bedeutet, dass aktuell normkonforme Metadaten in der Zukunft veraltet und kaum wiederverwendbar sein können. Das Dublin-Core-Datenformat, das im ersten Jahrzehnt des 21. Jahrhunderts sehr erfolgreich war, wird beispielsweise in Europa bereits heute immer seltener genutzt, da sich das Europeana-Datenmodell durchsetzt. Mit Linked Data kann man Datensätze erstellen, die später flexibel quasi als Informationsatome auch nach neuen Standards verknüpft und weiterverarbeitet werden können.

Das *Institut für Zeitgeschichte München - Berlin*[2] betreibt eine Forschungsdatenbank von personenbezogenen Daten mit mehr als 3000 Normdaten (Authority Records), die aus der proprietären Software für Sammlungsmanagement[3] als XML-Bestände exportiert werden können. Damit die Daten nachhaltig bewahrt werden, arbeitet das Institut an der Transformation der Einträge in Linked Data (Gelati 2019). Dafür wurde bereits eine Mapping-Musterdatei getestet, sodass man in den Quelldateien relevante Informationen selektieren und sie in die geeigneten Felder der Zieldateien überführen kann.

Da Linked-Data-konforme N-Triples-Bestände aus Subjekt-Prädikat-Objekt-Sätzen bestehen, ist es erforderlich zu bestimmen, welche Prädikate von welchen Linked Data-Vokabularen benutzt werden sollen.

Mit Wikidata, dem Normdaten-Projekt der Wikimedia Foundation, kann man beispielsweise den folgenden Satz erstellen:

[1] FAIR Principles: https://www.go-fair.org/fair-principles/.
[2] https://www.ifz-muenchen.de/
[3] FAUST (c) by Land Software-Entwicklung

 ISSN: 1860-7950

```
http://example.com/33
https://www.wikidata.org/wiki/Property:P20
https://www.wikidata.org/wiki/Q3004.
```

Dieser Satz bedeutet: Jene Person, die über den persistenten Identifier `example.com/33` identifiziert wird, hat Ingolstadt als Sterbeort.

Linked Data Vokabulare

Doch warum soll man sich für Wikidata entscheiden? Das Repositorium https://lov.linkeddata.es/ weist mehr als 100 Linked-Data-Vokabulare nach, die oft gleichwertige Einträge haben. Lassen wir uns das vorgenannte N-Triple-Beispiel weiter untersuchen. Mehrere Vokabulare enthalten die Eigenschaft „Sterbeort“:

```
https://www.wikidata.org/wiki/Property:P20

https://d-nb.info/standards/elementset/gnd#placeOfDeath

http://id.loc.gov/ontologies/madsrdf/v1.html#deathPlace

http://dbpedia.org/ontology/deathPlace

http://sparql.cwrc.ca/ontology/cwrc.html#hasDeathPlace

http://www.rdaregistry.info/Elements/u/#P60592

https://schema.org/deathPlace
```

Jedes dieser Vokabulare kann als der zweite Teil des Beispielsatzes (statt https://www.wikidata.org/wiki/Property:P20) Anwendung finden, um gleichlautende Resultate zu erzielen.

Flexibilität gibt es auch mit den Objekten: Lassen wir uns das Sterbedatum prüfen. Meines Wissens bietet nur Wikidata eine exakte Entität für das Jahr 1920. Aber das Jahr 1920 kann sowohl als die volle Entität als auch als Kombination von

`integer("1920")` und `class(https://www.wikidata.org/wiki/Q577)`

ausgedrückt werden.

```
http://example.com/33
https://www.wikidata.org/wiki/Property:P570
https://www.wikidata.org/wiki/Q2155 .

http://example.com/33
https://www.wikidata.org/wiki/Property:P570
"1920"^^https://www.wikidata.org/wiki/Q577 .
```

 ISSN: 1860-7950

Um einen Kalendertag auszudrücken, bleibt nur die letztere Option, denn Wikidata bietet beispielsweise für den Tag „04.01.1988“ keinen eigenen persistenten Identifier.

```
http://example.com/33
https://www.wikidata.org/wiki/Property:P570
"04.01.1988"^^https://www.wikidata.org/wiki/Q205892 .
```

Einrichtungen, die Linked Data generieren, können ihr eigenes Vokabular entwickeln sowie bereits bestehende Ontologien nutzen (Guernaccini, Mazzini, Bruno 2019). Die meisten arbeiten mit beiden Optionen: Die Deutsche Nationalbibliothek[4] erstellt Linked Data sowohl mit ihrer eigenen „GND Ontology“[5] als auch mit den berühmten „FOAF (Friend Of A Friend) Vocabulary Specification“[6] und „Dublin Core Metadata Terms“[7].

Sind der große Erfolg im deutschsprachigen Raum der Gemeinsamen Normdatei[8] (aber momentan nicht der GND Ontology) sowie die weltweite Nutzung der oben genannten Vokabulare ein ausreichender Grund dafür, sie heute als nachhaltige Lösung für die Langzeitarchivierung zu schätzen? Wenn derselbe Satz mit verschiedenen Vokabularen zweimal formuliert würde, würde das unerwartete Ende der Weiterpflege (oder Migration, oder Änderung) eines Vokabulars keinen Informationsverlust verursachen.

Wikidata als Linked Data-Mittelpunkt

Bedeutet die nachhaltige Datenarchivierung von Linked Data nicht, wenn möglich, immer zwei Vokabulare zu nutzen? Das heißt, ein „Backup-Vokabular“ für den Fall fragmentierter, instabiler und wechselhafter Ontologien zu haben? Wenn ja, welches?

Dafür scheint Wikidata die Lösung zu sein. Mehrere wissenschaftliche Beiträge haben „Wikidata as a linking hub“ (Neubert 2017) oder „Wikidata as a universal identifier“ (van Veen 2019) bereits vorgeschlagen. Wikidatas reiches Angebot von sowohl Entitäten als auch Eigenschaften ermöglicht es, Wikidata und die nahestehende Datenbank DBpedia[9] als „Backup-Vokabular“ bzw. „Linked Data hub“ zu verwenden. Genau dieser Weg wird fortan vom Institut für Zeitgeschichte München - Berlin mit seiner Forschungsdatenbank von personenbezogenen Daten getestet.

Derselbe Salz wird dank XSL-Mapping einmal laut des fachspezifischen (archivalischen) Standards RiC (Records in Contexts Ontology)

```
http://example.com/33
http://purl.org/ica/ric#RiC-hadDeathDate
"04.01.1988"^^https://schema.org/Date .
```

[4]Ich nehme als Beispiel die Datei https://d-nb.info/129307343/about/lds.rdf.
[5]GND Ontology: https://d-nb.info/standards/elementset/gnd.
[6]Friend of a Friend (FOAF): http://www.foaf-project.org/.
[7]Dublin Core Metadata Initiative: DCMI Metadata Terms: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
[8]Gemeinsame Normdatei (GND): https://www.dnb.de/gnd.
[9]DBpedia: https://wiki.dbpedia.org/.

 ISSN: 1860-7950

und einmal laut der fachübergreifenden Datenbank Wikidata erscheinen.

```
http://example.com/33
https://www.wikidata.org/wiki/Property:P570
"04.01.1988"^^https://www.wikidata.org/wiki/Q205892 .
```

Wäre es für tausende Einträge redundant oder komplementär? Übermäßig oder vorsichtig? Arbeitsintensiv oder nachhaltig? Hinweise und Kommentare sind herzlich willkommen.

Bibliographie

Francesco Gelati. (2019). Archival Authority Records as Linked Data thanks to Wikidata, schema.org and the Records in Context Ontology. ICARUS (International Centre for Archival Research) Convention „Archives and Archival Research in the Digital Environment“, Belgrad, Serbien, 2019-09-23 bis 2019-09-25. https://doi.org/10.5281/zenodo.3465304

Fabiana Guernaccini, Silvia Mazzini, Giovanni Bruno. (2019). LOD publication in the archival domain: methods and practices. In *Open Data and Ontologies for Cultural Heritage. Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage co-located with the 31st International Conference on Advanced Information Systems Engineering (CAiSE 2019)*, hrsg. von Antonella Poggi: 15–26. http://ceur-ws.org/Vol-2375

Joachim Neubert. (2017). Wikidata as a linking hub for knowledge organization systems? Integrating an authority mapping into Wikidata and learning lessons for KOS mappings. In *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017)*, hrsg. von Philipp Mayr, Douglas Tudhope, Koraljka Golub, Christian Wartena and Ernesto William De Luca: 14–25. http://ceur-ws.org/Vol-1937

Theo van Veen. (2019). Wikidata: from „an“ Identifier to „the“ Identifier. In *Information Technology and Libraries*, 38(2), 72–81. https://doi.org/10.6017/ital.v38i2.10886

Jakob Voß. (2017). Normdaten-Mappings in Wikidata. Subject Indexing & Information Technology Workshop (SIIT), Göttingen, Germany, 2017-05-11. https://doi.org/10.5281/zenodo.574452

Websites

https://www.w3.org/TR/n-triples/

https://www.wikidata.org/

https://lov.linkeddata.es/

https://wiki.dbpedia.org/

https://d-nb.info/standards/elementset/gnd

 ISSN: 1860-7950

http://xmlns.com/foaf/spec/

https://dublincore.org/specifications/dublin-core/dcmi-terms/

http://purl.org/ica/ric

https://schema.org/

Francesco Gelati (1987) erwarb die Master-Abschlüsse Linguistik an der Universität Ca' Foscari Venedig und Geschichte an der Universität Straßburg. Auch besuchte er die Schule für Archivwissenschaft des Staatsarchivs zu Venedig. Von 2017 bis 2019 arbeitete er beim Belgischen Staatsarchiv als Datenimportmanager und ist seit 2019 Archivar am Institut für Zeitgeschichte München - Berlin. Er interessiert sich für digitale archivalische Standards, Linked Data und Forschungsdatenmanagement. ORCID ID: https://orcid.org/0000-0002-6066-1308

 ISSN: 1860-7950