Neubert, Joachim; Tochtermann, Klaus

Conference Paper

Linked Library Data: Offering a Backbone for the Semantic Web

Suggested Citation: Neubert, Joachim; Tochtermann, Klaus (2012) : Linked Library Data: Offering a Backbone for the Semantic Web, In: Dickson Lukose Abdul Rahim Ahmad Azizah Suliman (Ed.): Knowledge Technology: Third Knowledge Technology Week, KTW 2011, Kajang, Malaysia, July 18-22, 2011, Revised Selected Papers. Semantic Technology and Knowledge Engineering Conference (STAKE 2011), ISBN 978-3-642-32825-1, Springer, Berlin, pp. 37-45, http://dx.doi.org/10.1007/978-3-642-32826-8_4

Linked Library Data: Offering a Backbone for the Semantic Web

Joachim Neubert, Klaus Tochtermann

ZBW German National Library of Economics – Leibniz Centre for Economics
Neuer Jungfernstieg 21, 20354 Hamburg, Germany
j.neubert@zbw.eu, k.tochtermann@zbw.eu

Since the publication of Tim Berner-Lees "Linked Data – Design Issues" [1] in 2006, the number of linked datasets in the Semantic Web has exploded. However, coverage and quality of the datasets are seen as issues[1], since many of them are the outcome of time-limited academic projects and are not curated on a regular basis. Yet, in the domain of cultural heritage institutions, large and high quality datasets have been built over decades, not only about publications but also about the personal und corporate creators and about the subjects of publications. Libraries and information centres have started to publish such datasets as Linked Open Data (LOD). The paper will introduce types of such datasets (and services built on them), will present some examples and explore their possible role in the linked data universe.

Libraries and Information Centres gather and organize information, often for centuries. Their cataloging rules and data formats seem arcane to every outsider. Recently however, some of them opened up to the Semantic Web and especially to Linked (Open) Data – for their own purposes and for inter-library exchange and re-use of data, but also for making their results available to the general public.

In May 2010, within the World Wide Web Consortium (W3C) a "Library Linked Data Incubator Group"[2] was formed in order to better coordinate such efforts and suggest further action. An open wiki[3] allows access to the work of the group. The publicly accessible datasets were collected in the CKAN data registry[4]. Mainstream library organizations such as the International Federation of Library Associations and Institutions (IFLA) and the American Library Association (ALA) are currently forming special interest groups on Library Linked Data. Focussed conferences such as the LOD-LAM Summit[5] in San Francisco or the SWIB conference[6] in Germany organize a vivid exchange of ideas [2].

In the remainder of this paper, typical datasets from the library community will be presented. Authority files, as discussed in the first section, are used to disambiguate especially personal and corporate bodies names. Thesauri and classifications are used

[1] see e.g., thread starting at http://lists.w3.org/Archives/Public/public-lod/2011Apr/0096.html
[2] http://www.w3.org/2005/Incubator/lld/
[3] http://www.w3.org/2005/Incubator/lld/wiki/Main_Page
[4] http://ckan.net/group/lld
[5] Linked Open Data in Libraries Archives and Museums (2011), http://lod-lam.net/summit/
[6] Semantic Web in Libraries (since 2009), http://swib.org

to organize knowledge mostly in confined domains, as shown in the second section. As demonstrated in the third section, bibliographic datasets not only give us data about publications, but also link together data about people, organizations and subjects. Conclusions about the applicability of library linked data in the broader semantic web are drawn in the fourth, final section.

Agents Identified – Personal and Corporate Bodies Name Authority Files

Especially large scientific libraries have a constant need for identifying persons and – to a lesser extent – institutions which are creators, editors or subjects of works: They have to tell apart persons with the same name, or to track different names back to one single person. To this end, over decades so-called authority files have been built, typically under the curation of national libraries, often as a collaborative effort of many scientific libraries. Rules apply for additional properties which are necessary to identify a person, such as date of birth and/or death, profession or affiliation to an organsiation. Different spellings of the name – possibly in different scripts – are recorded, and publications by the person are referenced.

The rules for the individualization of entries differ from country to country. For the German "Personennormdatei" (PND, personal names authority file) for example, dates of birth and death are used, the field of activity, the title of a work by the person, the profession or occupation, the designation, country or location, relations to other persons or an affiliation to an organisation. Interpersonal relationships, such as child/parent/sibling/spouse of, are also denoted [3]. Normally, a minimum of two identifying properties (other than the name itself) is required for building a valid authority record. Since this information is carefully checked by professional staff, the resulting data is normally of high quality.

The German National Library (DNB) has published this data in its "Linked Data Service"[4]. And like many other national institutions it feeds its personal and corporate bodies name authority files into the Virtual International Authority File (VIAF)[7][5]. National libraries or union catalogs of Australia, the Czech Republic, France, Hungary, Israel, Italy, Portugal, Spain, Sweden, Switzerland, the Library of Congress, the Vatican Library, the Bibliotheca Alexandrina and the Getty Research Institute add their data on a regular basis; Canada, Poland and Russia do this with a test status. VIAF merges the records from different sources to clusters under own URIs, which "expands the concept of universal bibliographic control by (1) allowing national and regional variations in authorized form to coexist; and (2) supporting needs for variations in preferred language, script and spelling"[8]. VIAF offers access to RDF representations and an autosuggest lookup service for the clusters. An alternate, more artifical-intelligence-inspired approach is taken by the (currently not publicly accessible) ONKI People service [6].

[7] http://viaf.org/
[8] http://www.oclc.org/research/activities/viaf/

An example for a VIAF personal name authority entity is given in Fig. 1. It shows how a person (the "real world entity" Anton Chekhov) is described by VIAF itself and by the aggregated authority files (e.g., of the German and Russian National Library). The collected foaf:name entries sum up to 165 (!) variants in different spellings and scripts. As may be noticed, VIAF itself takes no choice of a preferred name. The different libraries however do so.

```
<http://viaf.org/viaf/95216565>
  rdaGr2:dateOfBirth "1860" ;
  rdaGr2:dateOfDeath "1904" ;
  a rdaEnt:Person, foaf:Person ;
  owl:sameAs <http://d-nb.info/gnd/118638289>, <http://dbpedia.org/resource/Anton_Chekhov>,
             <http://libris.kb.se/resource/auth/201439> ;
  foaf:name "Bogemskiĭ, A., 1860-1904", "Cecov, A. 1860-1904", "Cecov, Anton, 1860-1904", "Cekhava, Eṇtana, 1860-
             1904", "Cekhovha, Ayāṇtan, 1860-1904", "Cekoff, Antonio 1860-1904", "Chechov, Anton 1860-1904",
             ... .
<http://viaf.org/viaf/sourceID/DNB%7C118638289#skos:Concept>
    a skos:Concept ;
    skos:altLabel "Bogemskiĭ, A., 1860-1904", "Cekhava, Eṇtana, 1860-1904", "Chehov, Anton, 1860-1904", ...  ;
    skos:inScheme <http://viaf.org/authorityScheme/DNB> ;
    skos:prefLabel "Čechov, Anton P. 1860-1904" ;
    foaf:focus <http://viaf.org/viaf/95216565> .

<http://viaf.org/viaf/sourceID/RSL%7Cnafpn-000082167#skos:Concept>
    a skos:Concept ;
    skos:altLabel "Chekhov, Anton, 1860-1904", "Chéjov, Antón 1860-1904", "Csehov 1860-1904", ... ;
    skos:inScheme <http://viaf.org/authorityScheme/RSL> ;
    skos:prefLabel "Чехов, Антон Павлович, 1860-1904" ;
    foaf:focus <http://viaf.org/viaf/95216565> .
```

Fig. 1. VIAF data about the Russian writer Anton Chekhov (heavily shortened)

Together, libraries' authority files form a large fund of interlinked identities for persons and organisations. As table 1 shows, it outnumbers in this respect by far the most used dataset on the Semantic Web, DBpedia: 364,000 persons in DBpedia as compared to 10 million in VIAF. The VIAF entries cover persons (and in parallel organisations) who created works, but also persons – living as well as historical– who are or were subjects of works. In conjunction with different national cultures, this leads to a very broad coverage of "publicly known" persons and organisations.

The authorities become even more valuable when interlinked with already existing Linked Data hubs such as DBpedia/Wikipedia. To this aim, a project of DNB and the German Wikipedia initiated the crowdsourced enrichment of Wikipedia pages with PND ids [7], which resulted in up to now 55,000 DBpedia–PND links. Projects such as "Linked History"[9] (University Leipzig) already make use of these links.

[9] http://aksw.org/Projects/LinkedHistory?v=oi6

	Persons	Organisations / Corporate Bodies
DBpedia[10]	364,000	148,000
Library of Congress Authorities[11]	3,800,000	900,000
German National Library Authority Files[12]	1,797,911	1,262,404
VIAF[13]	10 million	3.25 million

Table 1. Numbers of individual persons and corporate bodies identified by selected sources

The potential of interlinking through authorities as shared identities for web resources is demonstrated by the PND-BEACON project. Though not in any Semantic Web format, it connects more than 50 datasets[14] by use of identifiers from the personal name authority file as a common key.

Structured Knowledge Organisation – Thesauri and Classifications

In contrast to the comparatively simple and well definable entities of persons and organisations, things get much more differentiated and fuzzy when it comes to general subjects. Libraries deal with the thematic scope of works by assigning subject headings or, more sophisticated, descriptors from a thesaurus or classes from a classification. These instruments form knowledge organization systems in that their concepts normally are well defined and separated against each other. Classifications form a strict mono-hierarchy of concepts about the world as a whole (Universal or Dewey Decimal Classification, UDC and DDC respectively) or about a domain of specific knowledge (e.g., the classification of the Journal of Economic Literature, JEL). Thesauri include much richer properties and relationships, especially preferred and alternate labels, possibly multilingual, editorial and scope notes, and (poly-) hierarchical as well as associative relations.

The SKOS standard[15], developed by W3C, was “designed to provide a low-cost migration path for porting existing [knowledge] organization systems to the Semantic Web” and to provide “a lightweight, intuitive conceptual modeling language for

[10] http://wiki.dbpedia.org/Ontology (as of 2011-01-11)

[11] http://authorities.loc.gov/help/contents.htm. The LoC name authority is currently available in RDF only through VIAF.

[12] http://www.slideshare.net/ah641054/linkedrdadata-in-der-praxis (as of 2010-11-30). 58,307 entities, which represent geographical entities, have been subtracted (Email by Alexander Haffner (DNB), 2011-03-22)

[13] Clusters of merged personal and corporate name records from 18 participating libraries, http://outgoing.typepad.com/outgoing/2010/11/corporate-viaf.html (as of 2010-11-23) and email by Jeff Young (OCLC), 2011-03-21

[14] http://ckan.net/package/pndbeacon

[15] http://www.w3.org/TR/skos-reference/

developing and sharing new KOSs. … SKOS can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools, as exemplified by social tagging applications." [8]

An example of a SKOS concept, as a human-readable XHTML page with embedded RDFa data prepared for use in the Semantic Web, and in its Turtle representation, is given in Fig. 1 and 2 (taken from STW Thesaurus for Economics[16]).

Fig. 2. STW Thesaurus for Economics concept, XHTML+RDFa representation

[16] http://zbw.eu/stw

```
<stw/descriptor/12094-5>
   gbv:gvkppn "091386640"^^xsd:string ;
   a skos:Concept, zbwext:Descriptor ;
   rdfs:isDefinedBy <stw/descriptor/12094-5/about> ;
   rdfs:seeAlso <econis/search/descriptor/Corporate%20restructuring>, <econis/search/descriptor/Reorganisation> ;
   skos:altLabel "Business redesign"@en, "Business reengineering"@en, "Reengineering"@en,
            "Reorganisationsprozess"@de, "Reorganisationsprozeß"@de, "Unternehmensrestrukturierung"@de ;
   skos:broader <stw/descriptor/12105-5>, <stw/thsys/70562> ;
   skos:closeMatch <http://dbpedia.org/resource/Restructuring> ;
   skos:inScheme <stw> ;
   skos:narrower <stw/descriptor/24661-1> ;
   skos:prefLabel "Corporate restructuring"@en, "Reorganisation"@de ;
   skos:related <stw/descriptor/12196-4>, <stw/descriptor/12697-3>, <stw/descriptor/19254-2> ;
   skos:scopeNote "Für betriebliche Anpassungsmaßnahmen der Organisationsstruktur."@de ;
   zbwext:indexedItem <econis/search/descriptor/Corporate%20restructuring>,
            <econis/search/descriptor/Reorganisation> .
```

Fig. 3. STW Thesaurus for Economics concept, Turtle representation (extract)

More and more thesauri are published using SKOS, in the fields of public administration[17], social sciences[18], environmental information[19], medical subjects[20], agriculture and food[21], astronomy[22] and many more. The Library of Congress Subject Headings[23] try to cover the complete domain of human knowledge, similar to the German (SWD)[24] and French (RAMEAU)[25] subject heading files. The UDC is working on a "SKOSification", for the DDC [9] the first levels of the classification are already published[26].

These thesauri and classifications are a valuable source for well defined concepts. Although it is not straightforward, as Hyvönen et al. [10] assess, to convert such vocabularies to ontologies – since their understanding often implies implicit human knowledge – they can inform work on general or specific ontologies. A rough outline for a "thesaurus-to-ontology transformation" is given in the cited work.

In the field of subject thesauri, as well as in that of authority files, more and more mappings are published. This includes the results of experiments in multi-lingual matching, e.g., between LCSH, Rameau and SWD concepts [11], as well as domain centric mappings. The Agrovoc thesaurus of FAO for example was mapped to

[17] Eurovoc (the EU's multilingual thesaurus), http://eurovoc.europa.eu/, and UK ESD standards, http://standards.esd.org.uk/

[18] Thesaurus for the Social Sciences (TheSoz), http://www.gesis.org/en/services/tools-standards/social-science-thesaurus/

[19] General Multilingual Environmental Thesaurus, http://eionet.europa.eu/gemet

[20] Medical subject Headings (MeSH), http://neurocommons.org/page/Bundles/mesh/mesh-skos

[21] Agrovoc, http://aims.fao.org/website/Download/sub; NALT, http://agclass.nal.usda.gov/

[22] International Virtual Observatory Alliance astronomy vocabularies http://www.ivoa.net/Documents/latest/Vocabularies.html

[23] http://id.loc.gov/

[24] http://www.d-nb.de/eng/hilfe/service/linked_data_service.htm

[25] http://stitch.cs.vu.nl/rameau

[26] http://dewey.info/

Eurovoc, NALT, GEMET, LCSH, and STW.[27] This adds inter-scheme relationships for concepts to the inner-scheme relationships described above, and adds value through often multi-lingual labels. The resulting concept network can be exploited for retrieval applications as well as for ontology building.

Tied together – Publications

Publications – working papers, journal articles, books – and their archiving and provision are the libraries' main business. The collected metadata about publications is essential for the administration of their holdings, for their internal workflows and for providing access to their patrons. Seen from outside, publications can be viewed as small nodes or linking hubs, structuring the landscape of science. Publications tie together subjects (ideally represented as links to formally expressed concepts), people (as authors or editors), organizations (involved as publishers or via affiliation of authors), both ideally represented as links to authorities, and – more and more frequently – machine readable contents, which are open to text analysis via natural language processing (NLP) tools. This offers opportunities for interrelating data, for example about co-authorship networks or about main foci of research for different countries.

Different libraries have started to release data about publications in Linked Data formats: The National Library of Sweden [12] published the complete catalog as RDF (dcterms, skos, foaf, bibo ontology), including links to DBpedia and the Library of Congress Subject Headings. The Hungarian National Library [13] did the same, and the French National Library presented plans to follow [14]. The British Library released parts of its catalog as open data and also prepares for LOD [15]. In Germany, the Library Service Center of North-Rhine Westphalia [16] and some university libraries published their catalogs, too. The "Open Library", as an "open, editable library catalog, building towards a web page for every book ever published"[28], makes about 20 million book description in RDF available, more than a million with searchable full text. It combines different editions of a work and links out to OCLC's WorldCat and the Library of Congress.

Since all of this happened recently, few of the opportunities offered by this large amount of RDF structured bibliographic data have been exploited yet. Issues remain: For example, merging all the data will not be an easy task, because common identifiers are lacking (ISBN numbers apply only to a fraction of the data, and they are not reliably unique) and the mappings to RDF structures differ widely. However, we expect to see much more use in the future.

[27] http://www.taxobank.org/content/agrovoc-thesaurus
[28] http://openlibrary.org/

Conclusions

The linked open datasets published by libraries and information centres create opportunities for the Semantic Web as a whole. As Hannemann and Kett [17] put it: “Library data tends to be of very high quality, being collected, revised and maintained by trained professionals. As such, it has the potential to become a much-needed backbone of trust for the growing semantic web.” The curation of large datasets and terminologies is expensive, especially if strict and transparent policies are to be applied, because this means personnel expenditures over a long time. Therefore, it is difficult to achieve by academic projects. Publicly funded cultural heritage institutions on the other hand have long experience in data curation. They have a high score in long-term stability, reliability and independence from commercial interests. The Linked Data community could benefit from the offering of their autoritative and trusted data sets, using it for linking hubs in the web of data.

References

1. Berners-Lee, T.: Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html, (2006).
2. Borst, T., Fingerle, B., Neubert, J., Seiler, A.: How do Libraries Find their Way onto the Semantic Web?, http://liber.library.uu.nl/publish/articles/000482/index.html, (2010).
3. Deutsche Nationalbibliothek ed: PND-Praxisregel zu RAK-WB § 311. Individualisierung von Personennamen beim Katalogisieren mit der Personennamendatei (PND), http://www.dnb.de/standardisierung/pdf/praxisregel_individualisierung_311.pdf, (2010).
4. German National Library: The Linked Data Service of the German National Library. Version 3.0, http://files.d-nb.de/pdf/linked_data_e.pdf, (2011).
5. Tillett, B.B., Harper, C.: Library of Congress controlled vocabularies, the Virtual International Authority File, and their application to the Semantic Web. Presented at the World Library and Information Congress: 73rd IFLA General Conference and Council , Durban, South Africa May 22 (2007).
6. Kurki, J., Hyvönen, E.: Authority Control of People and Organizations on the Semantic Web. Proceedings of the International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009). , Trento, Italy (2009).
7. Danowski, P.: Library 2.0 and User-Generated Content. What can the users do for us? Presented at the World Library and Information Congress: 73rd IFLA General Conference and Council , Durban, South Africa August (2007).
8. Isaac, A., Summers, E.: SKOS Simple Knowledge Organization System Primer. W3C Working Group Note, http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/, (2009).

9. Panzer, M., Zeng, M.L.: Modeling classification systems in SKOS: some challenges and best-practice recommendations. Proceedings of the 2009 International Conference on Dublin Core and Metadata Applications. pp. 3–14 Dublin Core Metadata Initiative, Seoul, Korea (2009).
10. Hyvönen, E., Viljanen, K., Mäkelä, E., Kauppinen, T., Ruotsalo, T., Valkeapää, O., Seppälä, K., Suominen, O., Alm, O., Lindroos, R., Känsälä, T., Henriksson, R., Frosterus, M., Tuominen, J., Sinkkilä, R., Kurki, J.: Elements of a National Semantic Web Infrastructure - Case Study Finland on the Semantic Web (Invited paper). Proceedings of the First International Semantic Computing Conference (IEEE ICSC 2007), Irvine, California. (2007).
11. Wang, S., Isaac, A., Schopman, B., Schlobach, S., Van Der Meij, L.: Matching multi-lingual subject vocabularies. Proceedings of the 13th European conference on Research and advanced technology for digital libraries. pp. 125–137 Springer-Verlag, Berlin, Heidelberg (2009).
12. Malmsten, M.: Making a Library Catalogue Part of the Semantic Web. Proc. Int'l Conf. on Dublin Core and Metadata Applications. (2008).
13. Horváth, Á.: Linked data at the National Széchényi Library - road to the publication. SWIB10. , Köln (2010).
14. Wenz, R.: data.bnf.fr: describing library resources through an information hub. SWIB10. , Köln (2010).
15. Wilson, N.: Linked Open Data Prototyping at The British Library. UK Library Metadata Services: Future Directions. , London (2010).
16. Ostrowski, F., Pohl, A.: ‚Linked Data` - und warum wir uns im hbz-Verbund damit beschäftigen. BIT Online. 13, 259-268 (2010).
17. Hannemann, J., Kett, J.: Linked Data for Libraries. Presented at the World Library and Information Congress: 76th IFLA General Conference and Assembly , Gothenburg, Sweden August 15 (2010).

Third Semantic Technology and Knowledge Engineering (STAKE 2011) conference, Kajang, Malaysia, July 18-22, 2011. Preprint of

Neubert, Joachim, and Klaus Tochtermann. “Linked Library Data: Offering a Backbone for the Semantic Web.” In *Knowledge Technology*, edited by Dickson Lukose, Abdul Rahim Ahmad, and Azizah Suliman, 37–45. Communications in Computer and Information Science 295. Springer Berlin Heidelberg, 2012.