Linking library metadata to the web: the German experience

Gabriele Meßmer

«What is the value of a catalogue of more than 23 million records?»was one of the questions we discussed when starting the linked open data project at the Bavarian State Library. Many generations of librarians have been doing a good job creating machine-readable catalogue records, nowadays called metadata, with the purpose of describing books, printed music, manuscripts and maps, of building up authority files, listing holdings and more. To increase the value of these expensive data, it is essential today not only to offer catalogue interfaces to retrieve information, but also to open the catalogue databases and to give free access to the records.

In Germany some libraries of the North-Rhine-Westphalian library network were the first to publish their records as linked open data (LOD) in March 2010. At the same time, the hbz created a LOD website. Since then more and more libraries have started to discuss the issue of open and linked open data as well as the question of making their data freely accessible. In Germany there are six library networks running five different union catalogs. Although the German cataloging networks and the libraries have worked closely together for many years, sharing common authority files as well as exchanging records for re-use, there are still in many cases different records for one and the same book in the various union

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5507

catalogs.

Because of the special situation of five different coexisting library network catalogs, there is no uniform identifier for catalogue records in Germany. Many records already have an OCLC number that serves almost as such an identifier, but this is by no means the case for all records. Sometimes records describing the same book may have two or even more different identifiers. After starting the LOD projects in Germany, it became immediately obvious that a common and persistent identifier was needed for every title record and ideally only one single identifier for different records describing one and the same resource. So Culturegraph[1] was born, a linked open data service with the aim to generate a specific identifier for all kinds of objects held by libraries in Germany. This identifier should be used to reference the description of various objects. It should have a defined syntax, must be unique and persistent.

In the first step, the German library networks provided records for monographs and multi-part works published after 1945 to be ingested into the Culturegraph database. Then the records were compared and bundles (clusters) were created with records which – although slightly differing – were supposed to belong together and described the same object. A number of identifiers served as match criteria such as ISBN, ISMN, OCLC number and others. By now a resolving and look-up service is available to retrieve these single titles or bundles in Culturegraph. The next step will be to establish a larger database with all records held in the German library networks, a task which is especially challenging when it comes to early printed books with long baroque titles.

The North-Rhine-Westphalian Library Service Center (hbz) commissioned the Berlin lawyer Dr. Till Kreutzer to analyse the legal aspects of open library data. In 2011 he published a guideline (*Open Data*

[1] http://www.culturegraph.org.

– *Freigabe von Daten aus Bibliothekskatalogen*) which is a good basis for all legal questions around open data. The guideline contains chapters about the protectability of data, especially of single data fields. It examines databases and collections of data such as catalogs respectively and it discusses the possibility to release catalogs and publish records as open data.

The first linked open data service of the German National Library was the publication of the authority file for personal names and the authority file for subject headings in 2010. Today, the German National Library makes available three data sets: title records of the main collection (without records for printed music), the German Union Catalogue of Serials (Zeitschriftendatenbank, ZDB) and the complete German integrated authority file (Gemeinsame Normdatei, GND). The data model for the bibliographic data is documented in a paper (*The Linked Data Service of the German National Library: Modelling of bibliographic data*), which is available on the web also in English. All data sets are published under a Creative Commons Zero (CC0) license.

The B3Kat project – open data

In 2010 the Cataloging and Metadata Commission of the Bavarian and the Berlin-Brandenburg library networks also started to discuss open and linked open data. The research libraries of Bavaria, Berlin and Brandenburg use a common catalogue database called B3Kat. This catalogue contains more than 23 million records of 180 member libraries. The records are held in MAB, the special German data exchange format, and are linked to records of the German authority file. A small working group was established in order to achieve quick results. At first the working group identified reasons for having open catalogue metadata.

- to make local and regional data visible worldwide, i.e. to have them no longer hidden in thedeep web;
- to complete and increase the value of already established web presences such as Wikipedia by giving links to authoritative resources;
- to provide data for newly developed web-based services;
- to integrate data into the semantic web with the possibility to re-use the completed and enriched data one's own catalogue environment;
- to contribute to and to actively promote the open access movement.

In the project part A an OAI PMH repository was established in the Aleph500 environment of B3Kat. The title records held in the Aleph system and structured according to the German exchange format MAB were converted to MARCXML and then provided as open data. The particular challenge was on the one hand to map as many MAB fields as possible to MARCXML, in order to include a maximum of information, and on the other hand to include the basic information about the owning libraries. As the data pool contains more than 23 million records – a huge amount of data – two different methods are offered to pick up the records: there is the complete data set, split in three parts, frozen at a certain date, and an OAI repository comprising all continuously ongoing updates of records or newly created records.[2]It is possible to download the whole data set or to select only the records of a specific library, to select a single record, if the B3Kat ID number is known, or to obtain defined sets of records

[2] Information about the open data pool can be found here: https://opacplus.bib-bvb.de/TouchPoint_touchpoint/help.do?helpContext=opendata_en.

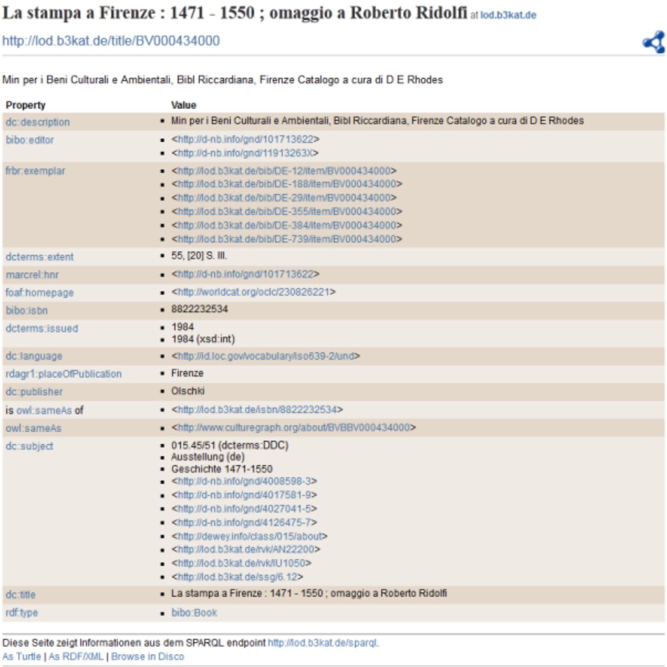
which can be identified because they include certain codes or fields. The complete data set will be published twice a year. As of March 2012 the complete set or parts of the set have been downloaded by this approach more than 400 times.

The B3Kat Project – linked open data

This MARCXML based open data pool serves as a basis for the linked open data pool (part B of the project). This was the easiest way for the next step, the transformation to RDF, because there are already tools for this process. Many fields provided in the MARCXML format had to be mapped to the RDF data model and published as RDF data. Wherever possible URIs are being used. Therefore for every title record a uniform resource identifier (URI) was created, based on the B3Kat identity number (starting with BV) and the name space reserved especially for B3Kat `lod.b3kat.de`. This name space was registered at DENIC, a registry for German domains under the top level domain `.de`.

To link the data to other data as well as possible many more links were implemented which were provided by the particular content of our records: for example links to the German integrated authority file, to WorldCat, to the language code ISO 639-2, to the Library of Congress subject headings and to the Dewey Decimal Classification. Currently, the RDF data pool consists of about 600 million RDF triples. As of March 2012 the pool has been downloaded more than 680 times and had more than 7600 visits. For the time being the linked open data set is published only bi-annually, while the updates for the open data are continuously provided. Information about the RDF data set, the data model, the used ontologies and the SPARQL endpoint can be found on the project webpage.[3]

[3] http://lod.b3kat.de. Currently this page is only in German.



La stampa a Firenze : 1471 - 1550 ; omaggio a Roberto Ridolfi at lod.b3kat.de

http://lod.b3kat.de/title/BV000434000

Min per i Beni Culturali e Ambientali, Bibl Riccardiana, Firenze Catalogo a cura di D E Rhodes

Property	Value
dc:description	Min per i Beni Culturali e Ambientali, Bibl Riccardiana, Firenze Catalogo a cura di D E Rhodes
bibo:editor	<http://d-nb.info/gnd/101713622> <http://d-nb.info/gnd/11913263X>
frbr:exemplar	<http://lod.b3kat.de/bib/DE-12/item/BV000434000> <http://lod.b3kat.de/bib/DE-188/item/BV000434000> <http://lod.b3kat.de/bib/DE-29/item/BV000434000> <http://lod.b3kat.de/bib/DE-355/item/BV000434000> <http://lod.b3kat.de/bib/DE-384/item/BV000434000> <http://lod.b3kat.de/bib/DE-739/item/BV000434000>
dcterms:extent	55, [20] S. Ill.
marcrel:hnr	<http://d-nb.info/gnd/101713622>
foaf:homepage	<http://worldcat.org/oclc/230826221>
bibo:isbn	8822232534
dcterms:issued	1984 1984 (xsd:int)
dc:language	<http://id.loc.gov/vocabulary/iso639-2/und>
rdagr1:placeOfPublication	Firenze
dc:publisher	Olschki
is owl:sameAs of	<http://lod.b3kat.de/isbn/8822232534>
owl:sameAs	<http://www.culturegraph.org/about/BVBBV000434000>
dc:subject	015.45/51 (dcterms:DDC) Ausstellung (de) Geschichte 1471-1550 <http://d-nb.info/gnd/4008598-3> <http://d-nb.info/gnd/4017581-9> <http://d-nb.info/gnd/4027041-5> <http://d-nb.info/gnd/4126475-7> <http://dewey.info/class/015/about> <http://lod.b3kat.de/rvk/AN22200> <http://lod.b3kat.de/rvk/IU1050> <http://lod.b3kat.de/ssg/6.12>
dc:title	La stampa a Firenze : 1471 - 1550 ; omaggio a Roberto Ridolfi
rdf:type	bibo:Book

Diese Seite zeigt Informationen aus dem SPARQL endpoint http://lod.b3kat.de/sparql.
As Turtle | As RDF/XML | Browse in Disco

Figure 1: Esempio di un record in B3kat.

Both the open data and the linked open data pool of B3Kat went live in the first days of December 2011. At the moment it is the largest bibliographic record or title set available in Germany.

Legal aspects

An important topic in the discussion about open data is the legal aspect. The working group had to consider questions such as:

- Will service providers agree with their records being published?
- Will all libraries accept the publication of their records?
- Are there fields/tags that should not be published, e.g. subject headings or URLs?
- What about catalogue enrichment, e.g. abstracts or table of contents integrated in the records?
- Under which license should open data be provided?

Some German libraries decided not to publish the full records as open or linked open data, but to omit from the open data some fields such as URLs. Some librarians believe that particularly expensive parts of a record, like subject headings, should not be published for free. The Bavarian-Berlin-Brandenburg working group however recommended to publish the records as completely and fully as possible – in order to make them really meaningful for all interested parties and to make sure that this service is also beneficial for the general data exchange between libraries and networks. For the time being only the URLs linking to table of contents purchased from commercial service providers cannot be published for copyright reasons.

Before publishing metadata it is also necessary to consider and to define the legal conditions for their reuse. There are two models:

- to publish data under a special license or
- to waive any rights.

To really comply with the concept and prerequisites of Linked Open Data (LOD)it is necessary to provide these data without any restrictions under a completely free license. The B3Kat records are therefore published under CC0 Universal Public Domain Dedication, which is also used for the metadata in the Europeana context. This allows the maximum use of the records and the provider has no administrative overhead to control the licenses of the users.

Conclusions

The project has come to an end, but it is not finished. There are still things to do: both project webpages must also be published in English. The license information must be integrated into the records, the MARCXML records as well as in the RDF ones. Furthermore, an update process is needed to keep the linked open data up-to-date as well, which implies to handle corrected and deleted records. A request often articulated is to publish the complete set of 23 million records in the MARCXML based OAI PMH repository, not split it as it is in three parts. Up to now this could not be realized because of performance and hardware issues. Both parts of the project were successful. The implementation was realized quickly, the sets are frequently in demand and we learned a lot about publishing metadata in MARCXML (our experience so far has been predominantly with the German MAB format) and about doing this in an OAI repository. Technicians and catalogers had to work hand in hand to get the best out of the existing data. Publishing open data is no job to do on the fly, alongside the daily work. It needs time and money, because a sound calculation is necessary to account for staff cost, for the hard- and software needed and above all for the time required to hold the data up-to-date.

Providing open data may also mean a shift in data management inside the library community. Until now, delivery of data for different purposes was up to the database provider and a lot of work had to be done for different file definitions and transfers, always up to the sender, not the recipient. With Z 39.50 and even more with OAI it was still up to the data provider to define the method for the delivery of data and thus the structure and the fields to be supported. With open data and even more linked open data it is now up to the user to make the relevant choices and selections and even a re-modelling of records on their side. Nevertheless the learning curve in analyzing the data provided in order to make best use of them is still to be followed and a lot of standardization and harmonization of formats and contents is still to be done in order to make the use of open data a smooth method of library cooperation and record reuse.

Until late summer 2012 yet another OAI repository will be established in the context of the Europeana Libraries project in which the Bavarian State Library is one of the partners. One of the outcomes of the Europeana Libraries project is the *Report on the alignment of library metadata with the Europeana Data Model* (*EDM*). This repository will also use the open MARCXML data, but it will only contain metadata of digitized objects. These metadata will be enriched with links to thumbnails. Not least this repository – also in MARCXML – will serve as a data pool for German and European portals which present metadata for all kinds of digital materials. The expected advantage of EDM is to enrich the records in Europeana and thus make them interoperable and fit for the semantic web. With EDM for Europeana and CIDOC CRM, the common data format for the future German Digital Library we can clearly see which further requirements come up when dealing with digitized information. It is not only a thumbnail – it may be all images (surely not in the original high resolution, but in a lower one) and the structural metadata as transported in METS/MODS files and finally the full text information.

Linked open data will lead to new services which will be developed in the near future. They heavily rely on a lot of basic knowledge of librarians: metadata, structures, normalization, quality control, standard numbers etc. The librarian of today is no longer a cataloger, but a metadata specialist and a manager – working on how the rich information contained in library records can be most usefully exploited and integrated into the web.

References

Kreutzer, Till. *Open Data – Freigabe von Daten aus Bibliothekskatalogen*. Köln: Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen, 2011. http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/open-data-leitfaden.pdf. (Cit. on p. 392).

Report on the alignment of library metadata with the Europeana Data Model. 2011. http://www.europeana-libraries.eu/documents/868553/1eade085-34ac-487f-82af-d5cd2545e619. (Cit. on p. 399).

The Linked Data Service of the German National Library: Modelling of bibliographic data. Leipzig: Deutsche National Bibliothek, 2012. http://www.dnb.de/SharedDocs/Downloads/EN/DNB/service/linkedDataModellierungTiteldaten.pdf?__blob=publicationFile. (Cit. on p. 393).

Gabriele Messmer, Bayerische Staatsbibliothek.
messmer@bsb-muenchen.de

Meßmer, G. ”Linking library metadata to the web: the German experience”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5507. DOI: 10.4403/jlis.it-5507. Web.

Abstract: One of the major challenges of libraries today is to make metadata available for the usage and re-usage by researchers and the scientific community. Therefore it is necessary to open the cataloguing systems for non-restricted and completely free access. Libraries of Bavaria, Berlin and Brandenburg decided in 2011 to publish their shared network catalogue with nearly 23 million records as open data and as linked open data. In March 2012 this data pool won the second prize in the first German-wide programming competition ”Apps for Germany”. The paper presents the steps of the project and the versatile experiences in publishing the data of more than 170 libraries. In addition it will introduce the Europeana libraries project in that more than 5 million records among them 600.000 records of the Bavarian State Library will be ingested in Europeana and be published as linked open data.

Keywords: Library linked data; Semantic web; RDF; B3Kat Project

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15