Data Streams in *linked.swissbib.ch*

The Swiss Metacatalog in the Linked Open Data Cloud

Nicolas Prongué, René Schneider

University of Applied Sciences HEG Genève, Switzerland
{nicolas.prongue, rene.schneider}@hesge.ch

Keywords: bibliographic data; daily updates; data processing; data stream; interlinking; library metacatalog; linked data; swissbib

The project linked.swissbib.ch aims to integrate the metacatalog swissbib into Linked (Open) Data, by transforming, interlinking and enriching the data. This implies the setting up of an infrastructure providing on the one hand a data service for other applications via a specific interface and on the other hand an improved interface for the end user (e.g. a searcher). Expected benefits of that approach include better data interoperability, an easier data reuse and a more enriching user experience. The project is jointly carried out by the University of Applied Sciences HEG Genève, the University of Applied Sciences HTW Chur and the Basel University Library.

The aim of this poster is to describe the whole system infrastructure of linked.swissbib.ch, and particularly to reveal the challenges related to metadata operations at the level of a metacatalog, namely at a level where data is not produced but only harvested at a daily frequency. The originality of this poster lies in its representation of the data environment, using the metaphor of the data lake and data stream, introduced by Redman (2008).

Some fifteen Swiss library networks are providing bibliographic metadata to swissbib, which processes, deduplicates, and then transforms it. For these operations, various data formats are managed, among others the traditional library format MARC/XML, as well as more recent RDF based formats like

In: M. Gäde/V. Trkulja/V. Petras (Eds.): Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017), Berlin, 13th–15th March 2017. Glückstadt: Verlag Werner Hülsbusch, pp. 359–361.

RDF/XML, JSON-LD or NT. For the transformation process, the software Metafacture[1] is mainly used. As the data store of linked.swissbib.ch is actualised daily, it consists of a continually changing mass causing further challenges for the data management. On the one hand, the attribution of unique and permanent identifiers in the form of URIs – being the first of the four Linked Data grounding principles of Tim Berners-Lee (2010) – is problematic, because the records vary every day in function of the deduplication operation of swissbib. On the other hand, the interlinking with external datasets like VIAF or DBpedia – being the fourth of these principles – becomes tricky as it relies on the existing URIs. These data processing operations are made even more complex due to the large amount of data (about 21 million records, the equivalent of ca. 39 GB) composing the metacatalog. Once the data deduplicated, transformed and interlinked, it is made available for computer clients and human users. Firstly, massive data reuse is possible for machines through a RESTful API. Since the library networks providing swissbib have various terms of use, this could lead to difficulty of attribution for the re-user. To address this issue, a mechanism filters the data in such a way that only CC0-compatible records are made accessible via the API. Secondly, an experimental interface is being developed for the end user, whose goal is to offer an improved search and exploration experience based upon the new interconnected data.

Linked Open Data is often said to be a very promising technology. Nevertheless, its implementation into a concrete and sustainable application reveals extremely complex data processing operations. This is notably the case for a library metacatalog treating exclusively secondary data. This poster highlights the key challenges of such an approach and illustrates the solution found for the specific case of linked.swissbib.ch. The result will be of interest for all researchers who face similar problems in other metacatalogs.

Project website:

http://www.swissbib.org/wiki/index.php?title=Linked_swissbib

[1] Software website: https://github.com/culturegraph/metafacture-core <28.12.2016>

References

Berners-Lee, T. (2010): Linked Data. http://www.w3.org/DesignIssues/LinkedData.html

Redman, T. C. (2008): *Data driven: profiting from your most important business asset*. Boston, Mass.: Harvard Business Press.