# Linked Data as a Library Data Platform

## A Proof of Concept from National Technical Library of Czech Republic

Jindřich Mynarz
The Development of Electronic Services Department, National Technical Library
Technická 6, Praha 6, Czech Republic
jindrich.mynarz@techlib.cz

June 8, 2010

## Abstract

The Web has now become the dominant information environment. For libraries to come to be a part of this information ecosystem, their data must be transformed to fit within the capabilities and constraints of the Web architecture. In this paper we describe how the linked data publishing model can be used to transform library data.

Linked data offers a web-native publishing model built around the fundamental web technologies, which makes it particularly suited for library data on the Web. The transition from legacy library data to web-compatible linked data involves re-thinking focused on making the data more interconnected, interoperable, standardized, linkable, and ultimately more useful.

## 1 Introduction

Linked data is a publication model for publishing structured data on the web. It was specifically designed with the Web in mind as its target medium, so that it is considered the web-native data model. It serves as a means for achieving the semantic web vision. We argue that it can be an important part of the library vision as well.

This paper is structured as follows. In the next section we describe library data and their inherent problems. The following section introduces the linked data publishing model with a special focus on the ways in which it can help to overcome the issues of library data mentioned in the previous part. Fourth section delves into the process of converting library data to linked data and demonstrates this process on the data from the National Technical Library.[1] The final section concludes the paper with remarks about the possibilities of further convergence of library and web data models.

## 2 Library data

Usual library data are highly context-dependent. The context for which they are created are the libraries. They are tailored for distribution in a library-specific setting, such as online public access catalogues. Also, library data use the context of natural language for lexical identifiers of bibliographic entities (e.g. subject headings).

[1] http://www.techlib.cz/en/

This dependency entails that the data do not work if taken out of the designated context. If we put such data on the Web, outside of its original context, we would have a hard time understanding what it signifies. For example, if we take lexical identifiers outside of the scope of natural language when we try to share bibliographic records on international scale, we come to see their binding to entities they identify becomes somewhat fragile and not so obvious as we might have thought.

## 2.1 MARC

When we want to talk about library data it is almost impossible to leave out MARC. MARC stands for *Machine Readable Cataloging* and is a prototypical example of library data format. It is a niche data communication format approaching (or, exceeding) the end of its life cycle. MARC is library-specific and for that reason it works only in library-specific systems.

MARC segments data into records. The record's structure is made of fields and subfields. The values of MARC data elements are not restricted to a particular data type, so they should be treated rather as text. It is not really a format for data, but more likely a mark-up language for text.

The format is not self-describing as current semantic formats designed with a certain schema. Its semantics is implicit rather than explicit. To interpret the values found in MARC records an external source - the MARC specification - is needed. And this specification is not machine-readable so machines, in contrast to humans, in fact cannot interpret MARC format.

MARC structure has a relatively low data granularity. For example, there is not a single field for ISBN. There is a field 20 which can contain ISBN but also *something else* (strings such as "(pbk.)"). Therefore, almost every field value needs to be parsed, filtered and cleaned from superfluous punctuation.

Process of data extraction from MARC records can be quite complex. There is a lot of noise in MARC data and it is not always easy to determine what is the extracted value referring to. The meaning of MARC values can be derived from the position within the field or may be dependent on other values, such as field indicators.

## 2.2 Words as identifiers

When we export linked data we try to establish links. Fortunately, there already are links in MARC data. But the problem is that they use words as identifiers for the linked resources.

Words are only partial identifiers and, by implication, not universal, nor unique. They insufficiently identify resources as they are ambiguous and therefore in many cases need to be disambiguated. The ambiguity of words used as identifiers stems from the natural language phenomena including synonymy, antonymy or homonymy.

Words are context-dependent in that their resolution is bound to a natural language. URIs, on the other hand, work even if they are taken out of context.

As identifiers, they need to be easily recreated when accessing the resource they point to. In such cases, we stumble on problems such as mistakes in spelling or different lexicalizations to express the same concept.

Library data uses both string codes and formalized headings to link to external data. To maintain uniqueness of these identifiers they must be strictly formalized, and even then iden-

tifier clashes can arise. For example, to maintain uniqueness within *Polythematic Structured Subject Heading System*,[2] where each concept is identified by its preferred label, such as "anthropology", it must be supplemented with a two-character identifier indicating a top-level concept under which the concept is located.

These quirks make MARC difficult to process and contribute to the fact that MARC is not a web-compatible data carrier. Even though MARC somewhat works, it is a hindrance in further development. It increases the cost of *maintenance*, *development*, and *cooperation*, and for these reasons we should strongly consider replacing it with a format that does not suffer from these symptoms.

## 3 Linked data

Linked data knowledge technology enable straightforward data integration via globally unique identifiers and self-describing data formats. Linked data allows for a plethora of serializations suited for different purposes. The data are provided in a standardized manner through a single, easily accessible interface. The URI identification mechanism standardizes the manner in which the entities are referred and makes them addressable which in turn increases the possibility of their reuse.

Linked data are interoperable with other linked data on the Web. This enables easy communication within a range of applications. Linked data are open to be extended, re-used, or incorporated in applications.

Linked data publishing model builds a set of conventions for data behaviour through guidelines, principles and best practices. Adhering to such guidelines prescribing particular behaviour provides for straightforward development with linked data.

The main characteristics of linked data are prescribed by the *linked data principles*. Data conforming to these rules and conditions can be marked as linked data. There are four linked data principles (Berners-Lee, 2006):

1. *Use URIs as names for things.*
   Abiding with this principle, we have considered what resources we have in our data, and created consistent URI patterns for each resouce category. Also, we have minted different URIs for the resource and its representation so we can make assertions about them individually. In a couple of odd cases, we use *blank nodes* to name the resources figuring as subjects in just a few triples.

2. *Use HTTP URIs so that people can look up those names.*
   HTTP URIs we use for resource names are actually dereferenceable. Upon passing the URI reference to the user's HTTP agent a useful representation is returned in response.

3. *When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).*
   Consequently, we have implemented content negotiation according to the current best practice involving HTTP Accept headers to provide the most suitable representation (HTML, RDF/XML) of a resource upon HTTP request for its URI.

4. *Include links to other URIs, so that they can discover more things.*

[2] Subject headings system developed at the National Technical Library

The key concept of linked data is linking the described resources with one another resulting in a set of internal and external links. We link to resources we need in assertions about our data but for which we provide no description.

# 4 Converting library data to linked data

Our aim is to prove that libraries can serve an important role of data producers in the linked data ecosystem. We demonstrate the application of linked data as a library data platform on selected databases from the National Technical Library of Czech Republic.

This proof of concept should be realized by a fully functioning prototype that demonstrates feasibility of converting library data to linked data and verifies that the linked data publishing technology is capable of being used for library data.

Briefly mentioning the technical issues going with this process, we have programmed an abstraction for the API exposed by the integrated library system. In our case, that was Aleph X Services which serves MARC records in XML and provides basic search functionality. Accordingly, we have created an abstraction for dealing with MARC records providing useful methods while encapsulating their complexity. For every selected database we have instatiated a crawler that got passed in a base-specific callback function that did the extraction, mapping, and outputted the triples yielded to our RDF store.

To indicate the overall direction of the process of conversion we have formalized it into several phases, starting with choosing the datasets and the way of description of the obtained data, followed by data extraction and linking, and concluding with data distribution.

## 4.1 Choosing datasets

We begin with the process of choosing a set of accessible and machine-readable datasets consisting of the types commonly found in library setting, such as bibliographic and authority data.

We have selected datasets that are available, structured, and potentially interesting outside of the National Technical Library. We have ended up with a selection including the database of all Czech periodicals with assigned ISSN, Polythematic Structured Subject Heading System, and the database of bibliographic records.

## 4.2 Choosing the way of description

The following step lies in choosing the way of describing the datasets with domain ontologies and creating mapping from the selected datasets.

Before mapping any of the MARC data elements, we have harvested the data about MARC tag usage in every of the chosen databases, so we would not deal with fields and subfields that are never or rarely used. We began with the most frequently used MARC fields and tried to assign a corresponding description by the properties and classes taken from the available domain ontologies.

For the description of the data we have used domain ontologies that are preferably already established in the field of libraries (such as Dublin Core), or we have chosen to use the ontologies popular on the Web. We have inferred the popularity of an ontology from the statistical data for namespaces provided by the *Ping the Semantic Web*.[3] A great help in finding the appropriate

[3] `http://pingthesemanticweb.com/stats/`

concepts from ontologies was the Falcons semantic web search engine's *concept search* feature.[4]

While mapping MARC tags we were looking up into three standards representing the three flavours of MARC 21 format present in our data: *MARC 21 Format for Bibliographic Data*[5], *MARC 21 Format for Authority Data*[6], and *ISSN MARC 21*.[7]

The semantics of the newly created data can be considered more explicit when it is clearly marked-up by domain ontologies. In this way, there is a one well-defined interpretation of the data. Instead of deducing the meaning from the context of use or consulting manuals, it is simply read from the annotation by ontologies.

## 4.3 Extraction

From the data we have selected and described we need to extract values to upload to RDF store. First, we identified the data values we wanted and then extracted them from MARC, cleaned them for a reasonable degree by removing most of the extraneous characters, including significant punctuation specific to MARC 21. In this way, we tried to normalize the data to a well-defined form with each value provided with its description.

## 4.4 Linking

The essence of linked data is in links it contains that are referring to both internal and external data. It is quite the opposite approach to data harvesting, based on referencing instead of duplicating. Links enhance and enrich the data and allow for precise references.

Linking constitutes an effective way to refer to other datasets and re-use the resources they publish. It turns out that we can use linking to "outsource" authority data using non-library datasets (e.g. DBPedia or Geonames) or to bind together the records representing the same resource (as in union catalogues).

To find links suitable for the data processed a few methods can be employed. The most obvious approach is simple string matching. However, it suffers from the drawbacks of words as identifiers, because even if these strings are strictly formalized, in most cases they constitute only a partial identifier for an entity. For example, in most cases you cannot disambiguate geographic entity just by its name. Try searching for "Prague" in Geonames database[8] and you will find multiple places with exactly this name situated in different countries and continents.

So, we need to build a complete, unique identifiers (such as URIs) based on partial identifiers (such as lexical labels). The gist of the process is to come to a single, shared concept the analysed data are pointing to. Disambiguation involves creating boundaries around data entities that separate them from the others.

Some data values in library data can be used relatively successfully as common keys. We are talking about identifiers such as ISBN or DOI that can be designated as *inverse-functional properties* of entities we model. If two entities share the same value for one inverse-functional property, we can treat them with a reasonable degree of certainty as equivalent. Having these properties unambiguously identifying data enti-

namespaces.php

[4] http://iws.seu.edu.cn/services/falcons/objectsearch/index.jsp

[5] http://www.loc.gov/marc/bibliographic/

[6] http://www.loc.gov/marc/authority/

[7] http://www.issn.org/2-22682-ISSN-MARC-21.php

[8] http://www.geonames.org/

ties helps immensely to link them to other data.

Of course, there are many other matching algorithms we have not covered, such as property-based matching or hierarchical matching, but not all of them may be applicable to library data.

## 4.5 Distribution

Linked data is above all a data publishing model for distributing structured data. It encompasses a set of recommendations and best practices for exposing data on the Web. The access interface to linked data is defined in linked data principles. The Web itself, *"is increasingly a set of interfaces to datasets"* (Whitelaw, 2008).

In the final stage of the conversion, we have stored the data in an RDF triple store and have built a lightweight API for accessing the data.

Linked data are accessed by using the URIs. The design of URIs should preferably adhere to the REST architecture. This is represented in URI design patterns by the recommendation to mint URIs that are simple, readable, and hierarchical.

The basic operation of linked data API is to return a resource description upon HTTP request for its URI. This constitutes a single interface that is very much in line with the concept of *information commons*. An integrated interface for data provides back-end for integrated services opened up to the end users. This simple way of data publishing opens the wide spectrum of possible uses and distribution through multiple channels on the Web. There are many options for user interfaces that can be built upon linked data. There is a great potential for various navigation interfaces, such as faceted navigation.

We must notice that linked data is very much a back-end technology, so for library users to appreciate it, we must first build applications that harness the possibilities stemming from it. This can be seen as an extension of the third *linked data principle* which recommends to provide useful information upon requesting resource's URI.

### 4.5.1 Open data

The Web in its nature is an open system so the web data should be available preferably under an open licence.

In some countries there is a dispute whether data can be licenced as other works under copyright protection. This is important because licences such as Creative Commons take their obligatory nature from the copyright law. That is why they can be applied only on works protected by copyright.

In Czech Republic we consider the library records suitable for licencing under Creative Commons licences. The Creative Commons' licence codes were translated and adapted for Czech legislation in the year 2009. For the National Technical Library's data we will be using the *Attribution-Noncommercial-Share Alike 3.0 Czech Republic* licence.[9]

### 4.5.2 National Technical Library's prototype

The prototype of the application built upon the National Technical Library's data is for the time being an incomplete beta version. Not all the data we intended to cover are present and in some cases it may contain inconsistencies.

This incoherency stems from imperfect data extraction mechanisms which we still tune to produce the best results possible. Provisional na-

[9] `http://creativecommons.org/licenses/by-nc-sa/3.0/cz/`

ture of the current prototype implies some pieces of data may not be interpreted as expected.

Even after the public release of the prototype we plan to maintain it in parallel coexistence with the original way of data distribution. We take it as a small-scale experiment that will eventually point out the usefulness of linked data approach in building end-user applications.

## 5 Conclusion

Since the library's presence is increasingly shifting towards the Web, libraries have to adapt to fit in this broader context. In embracing linked data publishing model the libraries will benefit themselves in many respects.

The Web has now become the dominant information environment and the primary way users interact with information. Library users are more often than not online and expecting the fully-fledged library services being online as well.

Libraries must not forget they are creating data primarily for machines. Users get the data only through applications. And that is why we can say there are at least three categories of library data users: library patrons, librarians, and *machines*.

The linked data publishing process implies that the data are capable of being linked to, which in turn increases their visibility on search engines. The data are crawlable by search engines which may by implication increase the number of incoming links to the library.

Publishing data on the Web detaches library data from the tight relationship with intergrated library systems. The clear and explicit description that is provided by the use of domain ontologies allows library data to be taken out of its primary context and still retain enough meaning. It means freeing library data from the boundaries of library catalogs and library-specific access protocols.

The question that comes to mind is whether libraries have something to give in return to the linked data community. Which are the areas of this data ecosystem that could benefit from the expertise of the library community?

We suggest that libraries do have a vast amount of experience in managing document collections and providing information services, which they might "translate" to the new environments. In this way, the library's brand of quality and trustworthiness may be associated with the linked data they produce and libraries can become major providers of reliable, authoritative datasets.

The reorientation towards linked data introduces a fundamental change to library data. From the usual view of library data divided into records, we see new *"metadata models for libraries, which are now evolving towards the Web"* (Baker, 2010).

Library data can no longer be only a *set of records*, rather it should constitute a *web of data*. The information that was formerly hidden in *text* is in this form encoded as *data*. From the original *context-dependent* data we instead made data that are more *context-independent*. This is possibly largely because the previously implicit meaning was made much more *explicit* so that it can be read and interpreted by machines without the need to know its original context. The data that was originally heavily *library-specific* can be treated as *environment-agnostic*.

Because of the way linked library data are exposed, they promote sharing and linking. Links stand for connections to the outer Web, while the incoming links to the library data bring new

users to the library, be it humans or machines.

From the viewpoint of the Web as the new dominant information ecosystem, the libraries should reassess the role they want to play and consider the Web as the new context for data delivery. We would very much like to see libraries as a part of the data supply chain, integrated by a common set of standards with other data producers and consumers, such as publishers, book-sellers, and other cultural institutions (museums, archives).

We propose to adopt linked data principles as *library data principles*. Then, library data could become a *"part of the WWW, the biggest interoperability framework the world has ever seen"* (Gradmann, 2010). The suggested outcome implies a change in the infrastructure in which libraries are more comfortable with relying on the work of others and are more open to share their work.

Towards achieving this goal, we suggest that the best strategy to drive the change towards linked data is to begin with small-scale experiments that can prove the advantages of linked data approach mainly from the user's perspective. While continuing with such experiments, we think that it is still unlikely that linked data would become the *primary* way to store and access library data in the near term future.

## References


- BERNERS-LEE, Tim. *Linked data : design issues* [online]. Published 2006-07-27. Last change 2009-06-18 [cit. 2010-06-01]. Available from WWW: <http://www.w3.org/DesignIssue/LinkedData.html>.
- GRADMANN, Stefan. *Knowledge = information in context : on the importance of semantic contextualisation in Europeana* [online]. Europeana White Paper 1. EDL, 2010 [cit. 2010-06-01], 19 p. Available from WWW: <http://version1.europeana.eu/web/europeana-project/whitepapers>.
- HOGAN, Aidan [et al.]. Weaving the Pedantic Web. In *Proceedings of the* $3^{rd}$ *International Workshop on Linked Data on the Web (LDOW2010), in conjunction with* $19^{th}$ *International World Wide Web Conference.* Raleigh (NC) : CEUR, 2010. Also available from WWW: <http://events.linkeddata.org/ldow2010/papers/ldow2010_paper04.pdf>.
- Library of Congress Working Group on the Future of Bibliographic Control. *On the record : report of the Library of Congress Working Group on the Future of Bibliographic Control* [online]. Washington (DC) : Library of Congress, 2008 [cit. 2010-06-01]. 44 p. Available from WWW: <http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf>.
- BAKER, Tom; BERMES, Emmanuelle; ISAAC, Antoine. *Library Linked Data Incubator Group charter* [online]. Last modified

2010-05-20. W3C, c2010 [cit. 2010-05-30]. Available from WWW: <http://www.w3.org/2005/Incubator/lld/charter>.

- SINI, Margherita [et al.]. Ontology-based navigation of bibliographic metadata. In *Proceedings of the International Conference on the Semantic Web and Digital Libraries.* 2007. Available from WWW: <https://drtc.isibang.ac.in/bitstream/handle/1849/320/007_p10_sini_formatted.pdf>.

- SMITH-YOSHIMURA, Karen [et al.]. *Implications of MARC tag usage on library metadata practices.* Dublin (OH) : OCLC, 2010. Report produced by OCLC Research in support of the RLG Partnership. Available from WWW: <www.oclc.org/research/publications/library/2010/2010-06.pdf>. ISBN 978-1-55653-378-5.

- WEINSTEIN, Peter C. Ontology-based metadata : transforming the MARC legacy. In *Proceedings of the 3rd ACM International Conference on Digital Libraries.* New York (NY) : ACM, 1998, p. 254–263. ISBN 0-89791-965-3.

- WHITELAW, Mitchell. Art against information : case studies in data practice. *Fibreculture Journal.* 2008, iss. 11. <http://journal.fibreculture.org/issue11/issue11_whitelaw.html>. ISSN 1449-1443.