Experience with transformation of bibliographic data into linked data

Jitka Hladka[1], Jindrich Mynarz[2], Vilem Sklenak[1]

[1]Department of Information and knowledge Engineering, University of Economics, Prague, Czech Republic

xhlaj00@isis.vse.cz, sklenak@vse.cz

[2]National Technical Library, Prague, Czech Republic

mynarzjindrich@gmail.com

Abstract: *Academic Bibliography Database of the University of Economics in Prague consists of bibliographic records of publications involving journal articles, conference papers, lecture notes, monographs and monograph chapters created by the academic staff of the university. We would like to discuss the experiences gained in the process of this dataset's transformation from its current data format to RDF-based data. During the course of conversion we will specify the entity types in our data and choose a way to model them. For data description we are going to use the most popular and widely implemented vocabularies and domain ontologies. Next, we will discuss the issues associated with interlinking the data with relevant and well-established datasets constituting a part of the web of data (e.g., DBLP Computer Science Bibliography).*

We would like to examine further possibilities of data re-use by making the data suitable for other applications such as citation managers (e.g., Zotero) to generate citations automatically. We are also interested in providing users with the information about availability of a resource as full text or as library holding, so we discuss cooperation with link resolvers (SFX, in our case) as well as other possible ways of linking. In this way, aligned with the concept of the semantic web, we maintain that the potential of the data would be maximized, as the information in bibliographic records would become easy to share, more visible due to incoming links, and more ready to be processed by web applications. We argue that academic bibliography data should be openly available, so it increases transparency of publishing activity of the university's academic staff. Beyond being re-usable, the data can be easily linked to so that they must not be duplicated at multiple locations. Likewise, these features together can provide a few benefits for bibliometric evaluation of science. The adoption of linked data publishing model for academic bibliography datasets involves moving to a more flexible data format for bibliographic data. While linked data is seen as a pragmatic implementation of the semantic web, we argue that it is not possible to implement the library data vision with the MARC, current standard for bibliographic data. Obviously, there is a need for more web-compatible and web-friendly data model. With this in mind, we suggest linked data as a part of the pragmatic implementation of the vision for bibliographic data on the Web.

Key words: bibliographic database, publication activities, semantic web, linked data, RDF

Introduction

Academic Bibliography Database stands for a very important resource of scientific research at the University of Economics, Prague. These bibliographic data are crucial for academic staff evaluation and when put on the Web they become important for the scientific community, they can be shared among the scientists and academic staff, the visibility of their works increases as well as the awareness of the University's scientific activities.

The current standard for bibliographic data, MARC is not suficient to express all the advantages of the data.[1] We adopt the linked data publishing model in order to make this bibliographic dataset openly available, more flexible, easy to share and visible, to increase its suitability for web applications.

In this paper we would like to discuss the experiences gained and issues encountered in the process of this dataset's transformation from its current data format to RDF-based data.

[1] *MARC is excellent for expression of "raw" bibliographic data, but academic bibliography database contains also information on the affiliation and on the relation to research activities.*

1. Current state of the Academic bibliography Database

Academic Bibliography Database of the University of Economics, Prague (so called database PCVSE) consists of bibliographic records of publications involving journal articles, conference papers, lecture notes, monographs and monograph chapters created by the academic staff of the university.

The database is updated continuously and it contains records available online with retrospective since 1982 (see fig. 1 and 2). Bibliographic records of works – results of long-term R&D intentions and projects supported by various state or public fundings are sent to the "*Information register of R&D results*", where the information on research and development in the Czech Republic is published.[2]

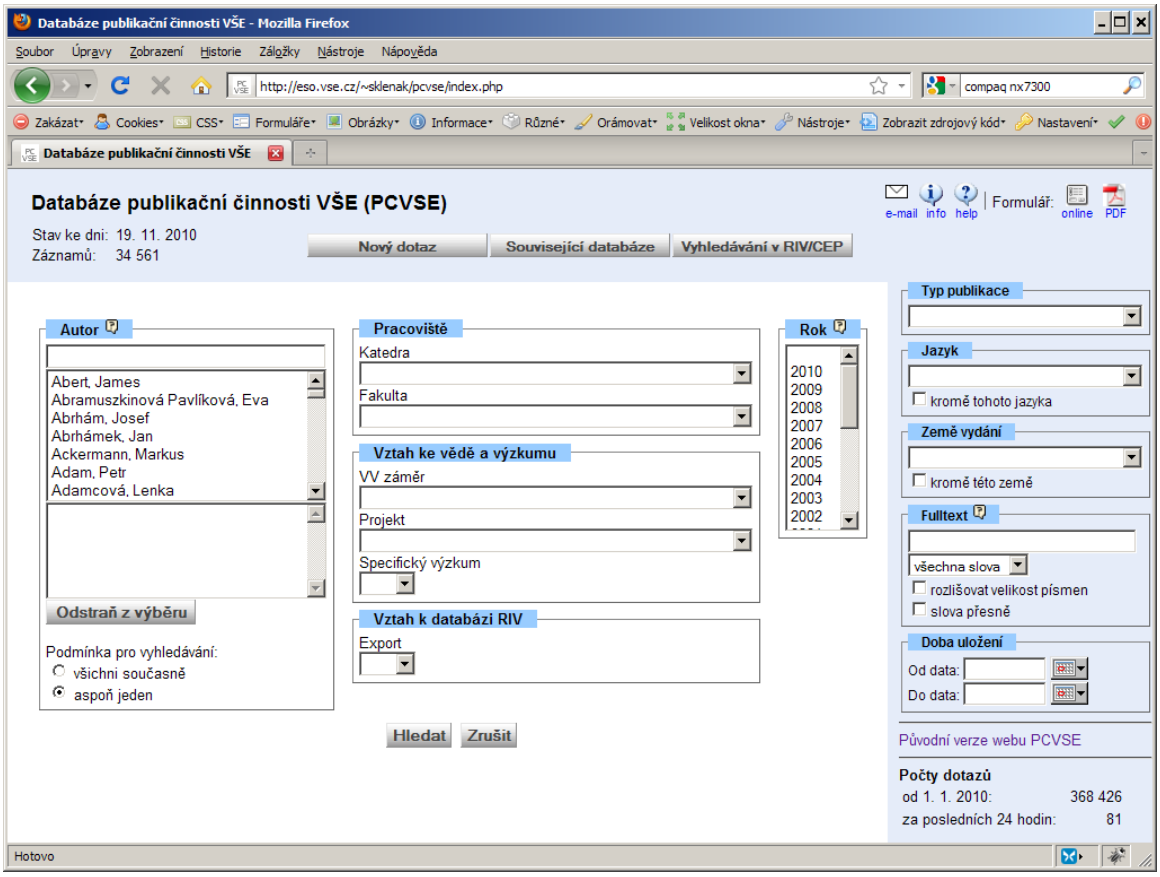

Fig. 1: PCVSE – current search form (only in Czech)

Academic publications represent a specific type of information resources. They may not be published in the formal way, but constitute rather a part of "*grey literature*" – the documents characteristic by their unoficial publication status. Such documents are often distributed in a small amount of copies, or posted to interested persons. Many of these releases are published in peer-reviewed journals, where the document quality is controlled by domain experts. Academic publishing was significantly affected by the increase of documents published in an electronic format. While in 1990s exclusive licensing for commercial publishers of electronic resources was preferred, current initiatives promote *open access* to scholarly journals [9]. Thus the content of these information resources is openly available via Internet for free or for subscription fees.

[2] Available at http://www.vyzkum.cz/FrontClanek.aspx?idsekce=1028.

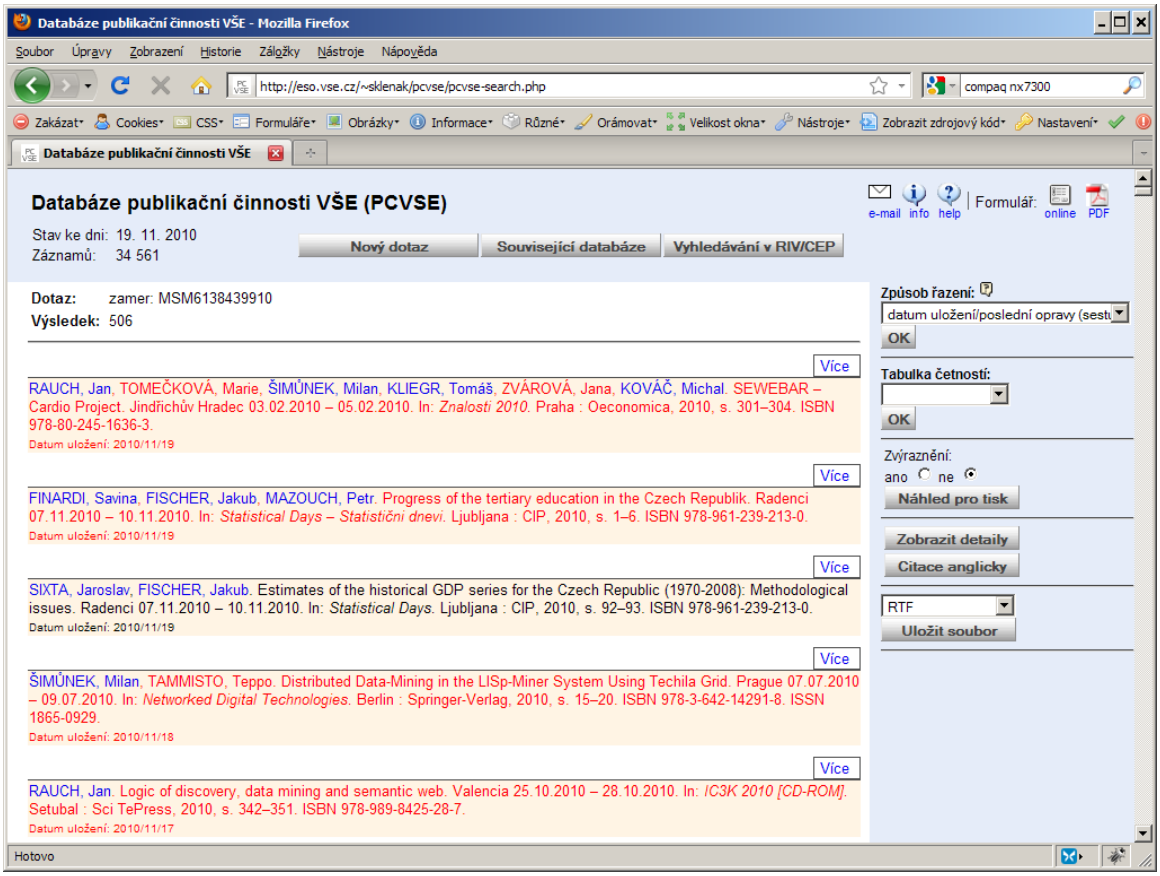

Fig. 2: PCVSE – list of query results

Current initiatives in various fields of scientific research, industry, or business involve taking advantages of the web technology to enable easier sharing and processing of the knowledge. Academic bibliography as an important resource of information about research and development activities of University of Economics in Prague can follow this trend. This is why we would like this stand-alone database of bibliographic records to adopt a more web-compatible and web-friendly data model.

The information about publishing activities of the members of academia has been stored in the bibliographic format MARC, which has remained the same till the present time. In addition to the bibliographic data, records are enriched by information about the project or R&D intentions related to the described document.

MARC is the standard format for representation and communication of bibliographic and related information. Machine-readable cataloging initiative came up in the 1960s led by the *Library of Congress* when the library processes automation were enhanced by the use of computers. Adoption of the MARC standard enabled more effective cataloguing, as well as sharing bibliographic information among institutions [8].

However, development in ICT and web technologies significantly affects and changes the bibliographic information processing as well as the format of documents in library collections. Discussions have risen up, whether or not is the MARC format still suitable for the environment of the Web: is it a mystifying antiquated data structure or a good metadata standard [10]?

While the changes to the content of bibliographic data need wide array of changes in the nature of creating metadata we decide to go for an easier and more immediately accesible goal that is to increase the potential of bibliographic data on the Web by the transformation of its *data format*.

2. Linked Bibliographic Data

The bibliographic data have been produced mainly by libraries, the standards for their creating, processing and exchange have been developed over the history of librarianship or information profession. As the Web becomes a common place to publish information and an important research

tool, there have been developed new formats of the "data about data" (i.e., metadata), produced by the *non-librarian* communities.

Bibliographic information has also become available online, but, for the most part, it is not yet an integral part of the Web. Suitability of the traditional and established standards for the communication of this information[3] in the Web environment is disputable [5].

Nowadays, there are efforts to combine the traditional and established library principles and standards with the new standards based on the common practises emerging on the Web that are aimed primarily at the pragmatic use of data in this environment [1]. The iniatives that are trying to bridge between those two sets of practice and to achieve a greater interoperability between library systems and networked information available on-line while encouraging libraries to re-orient their native approaches towards the Web (e.g. *W3C incubator group for Library Linked Data,* http://www.w3.org/2005/Incubator/lld/).

In the linked data communities the goals are similar. These initiatives encourage institutions and other data publishers to enrich their data by linking it with other datasets across the Web. Such links may be based on the typed relationships between "*pieces*" of data (entities). Data structure, relationships and the semantics of the data can be described in a format suitable for automated machine processing in applications and ultimately even in some artificial intelligence procedures, such as inferencing. The desired output of this undertaking is to make the human-readable knowledge available on the Web as machine processable.

The adoption of the linked data publishing model by the library community can be seen as being beneficial both for libraries and the wider linked data community. In this way, bibliographic data can be available in a data format that is familiar to the experts outside of libraries: the RDF. As it is often mentioned when discussing the role of libraries on the Web, the data produced by libraries, archives, digital repositories or academic institutions tend to be maintained by trained professionals and the data outputs are therefore of high quality. These data then have a potential a much-needed *backbone of trust* for the growth of the so-called semantic web [5].

One of the basic building blocks of the semantic web we have already mentioned is the *Resource Definition Framework* (RDF) (http://www.w3.org/TR/2 004/REC-rdf-concepts-20040210/) that serves for knowledge representation and storage of the represented data. While HTML documents contain information available for human to understand, with the use of RDF the structure of information can be parsed, re-constructed and processed also by web applications in a more straight-forward manner [4]. To illustrate this point, making the bibliographic data suitable for applications such as *citation managers* would enable them to consume the data and in this respect increase its further re-use.

RDF data format provides a simple way to connect information resources located in separated datasets by the interlinking of related pieces of information across the Web, from one dataset to another. In the work we describe in this paper, we consider the interlinking of our data with relevant and well-established datasets as its key feature that constitutes a fundamental component for the vision towards making the dataset a part of the web of data.

In the following sections, we will discuss the process of transforming the legacy data to fully-fledged linked data that is ready to be published on the Web.

3. Transformation

In the course of the transformation process, including data preparation and conceptual data modelling, we have encountered several problems and challenges. The most significant experiences and remarks will be presented in next section.

3.1 Data preparation

The process of the dataset's transformation from its original data format to RDF-based data was initiated by several data preparations.

Even though the original data were available in a standard format for description of bibliographic information (MARC) the bibliographic records in the database of publication activity were created by the members of academic staff without any cataloguing rules, standards, or controlled vocabularies applied. In fact, these data may be called *user-generated content*.

[3] *Such as the library-specific protocols of Z39.50 or OAI-PMH.*

There were no controlled lists used for maintaing the consistency of name references in the dataset, so, for example, subject classiication of the information resources was provided by keywords only, and it was not mapped to any established classification scheme.

Also the names of publishers and institutions were used in different variations and forms, including abbreviations and their translations to other languages. We decided to merge these variation names together by creating *synonym* sets ("synsets") that are common in language thesauri (e.g. *Wordnet* lexical database, http://wordnet.princeton.edu/wordnet/). Each group of the names belonging to a single entity was given a preferred name form. The selection of the most appropriate variation was not random, we referred to the *Publishers' Directory* authority file maintained by the *National Library of the Czech Republic*.[4]

Although the MARC format, in which we have obtained the data, captures a wide range of bibliographic and related data, we have decided not to transform records in their entirety including all the fields and subfields contained in them due to the complexity of this task. We argue that it is not in fact necessary to keep the full bibliographic record intact and provide a lossless transformation into RDF. The reason for this argument is that some of the information that is contained in MARC records was designed for the *library-specific* environments, for example local identifiers to use in an integrated library system, and therefore this kind of information is not that useful when put on the Web.

With these considerations in mind we have analysed the *frequency* with which each MARC data element occurs in our dataset. Based of the MARC fields and subfields occurencies we were thus able to determine which data elements are the most used and the most important to model correctly. The list of the most frequent MARC fields then points out to the essential parts of bibliographic data that we would like to involve in further data modelling.

3.2 Data modelling

At the stage of data modelling our aim was to preserve the original semantics of the processed MARC data elements by capturing it via RDF vocabularies and domain ontologies. The process of data modelling consisted of specifying entity types and the choice of the most appropriate classes and concepts for data description [11]. When we were choosing among the available means of expression for such a task that are currently available for re-use in RDF vocabularies we had the following criteria in mind:

- How widely is the vocabulary (ontology) used? We prefer the established and popular vocabularies.
- The description of the concept should match the MARC data element, as it is defined is MARC specification, as closely as possible. We were comparing the ways how different MARC elements are used and the ways RDF vocabularies are used in a search to find good-enough match.

In this step we examined the use of various vocabularies not only targeted at library community. First, we have examined the RDF vocabularies designed to represent bibliographic information, such as *Dublin Core* (http://dublincore.org/documents/dcmi-terms/), *Biblioontology* (http://bibliontology.com/) or *MarcOnt* (http://marcont.corrib.org/). Then, we have looked at the vocabularies that are not library-oriented but widely re-used in the wider web environment.

Taking into account the specifics of our dataset and reflecting the types of information contained in it, we have considered the use of vocabularies for structured description of scientific research, academic environment and research projects. We felt free to combine the various vocabularies we have selected to find the most appropriate descriptions for our data.

The vocabularies and domain ontologies we have finally decided to adopt include:

- **Bibliographic Ontology** (bibo). <http://purl.org/ontology/bibo/>
- **Dublin Core** (dc). <http://purl.org/dc/elements/1.1/>
- **DC Terms** (dcterms). <http://purl.org/dc/terms/>
- **Friend of a Friend** (foaf). <http://xmlns.com/foaf/0.1/>
- **Academic Research Project Funding Ontology** (arpfo). <http://vocab.ouls.ox.ac.uk/projectfunding#>

[4] *http://aleph.nkp.cz/F/?func=file&file name=find-b&local base=NAK*

- **Identity of Resources on the Web Ontology** (irw). <http://www.ontologydesignpatterns.org/ont/web/irw.owl#>
- **Semantic Web for Research Communities** (swrc). http://ontoware.org/swrc/swrc/SWRCOWL/swrc updated v0.7.1.owl

We have not started with the data modelling and mapping from MARC to RDF from scratch. There were previous approaches suggested that we have taken into account and used to bootstrap our modelling. Several works on mapping MARC format to RDF have been already done:

- **MARC to Dublin Core Crosswalk** – http://www.loc.gov/marc/marc2dc-2001.html,
- **MARC to RDF Mapping** – http://marc-must-die.info/index.php/MARC_to_RDF_mapping,
- **ConverterToRDF** – http://www.w3.org/wiki/ConverterToRdf (tool marcmods2rdf transforms MARC records from Z39.2 format into MODS and then from MODS to an RDF representation of MODS),

so we had a very useful starting point for our further work.

To find an appropriate concept for each MARC field or subfield was not always a straight-forward task without difficulties. We learnt about some potentially useful vocabularies that they are currently not available due to reconstruction, are newly established, or dataset-specific with no re-use happening outside the originating dataset. In order to follow the criteria stated above, in some cases (such as the expression of the *place of publication*) we had to choose an option with more general meaning than the original MARC element had from the well-established vocabularies (such as dc:coverage or dc:spatial from *Dublin Core*).

3.3 Data specifics

Some of the issues we have encountered in the process of its conversion to RDF were closely linked to the specifics of our dataset. We will now mention some of these issues that were caused by some of the dataset-specific characteristics of the information contained in the records of *Academic bibliography database*.

3.3.1 Modelling of the information about research projects

Academic bibliography of the University of Economics collects bibliographic data enriched with the information about the project or R&D intention related to the described document. This required us to do data modelling based on the examination of the available ontologies for the domain of research communities.

When considering the range of vocabularies that in some way relate to the domain of academic research we have decided to adopt ARPFO (*Academic Research Project Funding Ontology*, http://vocab.ox.ac.uk/projectfunding/schema), describing the project funding structure of academic research[5], and SWRC (*Semantic Web for Research Communities*, http://ontoware.org/swrc/), modelling the entities that can be found in research communities.

3.3.2 Aligning bibliographic data with the FRBR model

Following the trend in librarianship we have examined the implementation of the *Functional Requirements for Bibliographic Records* (FRBR) abstract model[6],a framework for relating data in bibliographic records and defining the basic level of functionality for the records created by national bibliographic agencies [6]. There are a few formalizations of this model available in RDF and we have chosen to use the basic concepts and relations that are defined in the Expression of Core FRBR Concepts in RDF vocabulary (http://vocab.org/frbr/core.html).

First, we have considered the suitability of the FRBR framework for our data. If implemented in a classical library catalogue, the adoption of the FRBR concepts can have significant benefits for the improvement of the retrieval and display of bibliographic information due to the clearly defined, structured framework it provides for relating data in bibliographic records; including multiple variations, versions, formats or editions of a single work that can be collocated and linked together.

---

[5] *Another vocabulary suitable for project description is .e.g. DOAP (Description of a Project, https://github.com/edumbill/doap/wiki), but DOAP is something more specific than ARPFO – DOAP is vocabulary to describe software projects.*

[6] *http://www.ifla.org/en/publications/functional-requirements-for-bibliograph ic-records*

However, the data of the *Academic Bibliography Database* are not classical library data, although they share the same data format. Multiple versions, formats or editions of works are occurring only rarely in it. Thus we did proceed to do the "*FRBRization*" of our dataset because we think doing it would not bring as many benefits as it can do for traditional library data.

4. Interlinking

After the data modelling stage of the conversion we have done the interlinking so that our dataset is joined with others in the web of data.

It was the adoption of linked data principles and best practices that has lead to the extension of the Web to a global data space connecting data from diverse domains of knowledge. URIs (Uniform Resource Identifiers), along with HTTP (Hypertext Transfer Protocol), a retrieval mechanism for the URIs of resources, provides the technology background for this Web of data.

First, we have already described a kind of interlinking in the previous section on data modelling – this is the interlinking that happens on the *schema level* by re-using common RDF vocabularies by referring to the URIs of resources in them. We acknowledged the need for using widely used and shared vocabularies in our restriction on the popularity of a given vocabulary and we then preferred using the most common vocabularies. It turns out that this practice of re-purposing well-known vocabularies makes it easier for the client applications to process linked data of this kind [3].

Besides using the most common vocabularies and joining our dataset with the web of data on the level of data schema, we also consider interlinking our data with related well-established datasets, and thus provide links on the *instance level*. If the information published as linked data referrs to external resources, it is possible to navigate between disparate resources through the interlinked data elements provided in RDF descriptions.

Using URIs minted in an own namespace in combination with other namespaces' URIs is suggested and encouraged practice [2] so that sometimes we just re-purposed a URI for an entity from another dataset and attached some assertions directly to it rather than providing the resource with yet another URI and then using it as a proxy via the equivalence links to the external resource. When a data source is enriched by the links to related entities in the other data sources, it enables a light-weight *integration* of such data source into the Web of data [3] and makes the datasets easy to merge.

It is common that a several datasets provide URIs for more re-usable common concepts which are often used in many specialized datasets, such as *DBPedia*[7] or *Geonames*[8] . These form the so-called linking hubs that serve as sources of authority data for re-usable content and are therefore frequently referenced [3].

We have decided to provide our data with interlinks to the following datasets, with which we could build useful and interesting connections:

- **Geonames** at <www.geonames.org> – geographic location of the place of publication
- **VIAF** (Virtual International Authority File) at <www.viaf.org> – name of publication's author in national libraries' authority files

This list is not exhaustive, we have alse considered other datasets to include, for example the *DBLP Computer science bibliography*.[9]

According to the common practice, we have used an *automated approach* to generate links [7]. This process could be based upon the shared vocabulary schema, or the similarity of entities within the both datasets.

We have encountered several difficulties during this process. At first, we had to deal with the *ambiguity of the names*. Geographic names were not uniquely specified so we had to choose from several different locations identified by the same name. We were able to select the appropriate geographic name by assigning publisher's name to the location.

Manual revision of the links was often necessary: several authors had the same name as others, so the selection was made upon the specialization of their work.

Obviously, the interlinking is not by any means finished. We may discover existing or newly created datasets worth linking to. What we have presented here is mainly the best practices and methods for

[7] *http://dbpedia.org*

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

[9] *http://www.informatik.uni-trier.de/~ley/db/*

doing the interlinking, and even if the actual list of the datasets we have linked to provides a little guidance, we see the general practices suggesting how to the linking more important.

5. Conclusion

Now that we have described in detail our project we will provide an overview of the main points we think were the most interesting, for the most part the practical experiences we have gained in this project.

As in any data conversion undertaking there were issues related to the format of the legacy data that we have converted. Issues with data quality lead us to employ a number of data cleaning techniques, including data reconciliation or the creation of the synonym sets.

One of the most re-usable parts of the process we have described is the data modelling. Because we have done an RDF data model for a standard format - MARC - there is a chance that other projects related to RDF conversions of library data may re-use the mappings we have specified to get RDF out of MARC data. Data modelling is still a difficult process and it can be subject to personal opinions as well. Thus, instead of providing the data model itself in this paper, we have mentioned the best practices and techniques how to go about this task. The re-use of the common RDF vocabularies is one of them which have its use in creating linked data that is interoperable.

The focus of the second main part of the paper was the question of interlinking with external datasets. We have described our approach that harnesses both deterministic and probabilistic matching to produce links between datasets. The effective and reliable linking is still an open research question but we have shown that there are some established methods and software that make the task easier.

The particular application of these methods and practices to the case of the *Academic bibliography database* we have demonstrated can serve as a prototype implementation that may be an inspiration for the future projects similar to ours and constitute a step in the evolution towards library linked data.

Acknowledgements

The work of Jindrich Mynarz and Vilem Sklenak is partially supported by CSF grant no. P202/10/0761, Web Semantization.

References

[1] BAKER, T., BERMES, E., ISAAC, A. *Library Linked Data Incubator Group Charter* [online]. W3C, 2010/06/01. Available from WWW: <http://www.w3.org/2005/Incubator/lld/charter>

[2] BIZER, Ch., CYGANIAC, R., HEATH, T.. *How to Publish Linked Data on the Web* [online]. 2008-07-17. Available from WWW: <http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/#whichvocabs>

[3] BIZER, Ch., HEATH, T., BERNERS-LEE, T. Linked Data – The Story So Far. In *Special Issue on Linked Data, International Journal on Semantic Web and Information Systems* [online]. Last modified 23rd July 2009. Available from WWW: <http://tomheath.com/papers/bizer -heath-berners-lee-ijswis-linked-data.pdf>

[4] CHO, A. *How RDF can use MARC in the semantic web world : using existing library cataloguing methods in organizing the Web* [online]. June 2009. Available from WWW: <http://www.suite101.com/content/how-marc-realizes-the-rdf -in-the-semantic-web-a122881#ixzz15ODoyD4C>

[5] HANNEMANN, J., KETT, J. Linked Data for Libraries. In *World Library and Information Congress: 76th General Conference and Assembly, Meeting 149. Information Technology, Cataloguing, Classification and Indexing with Knowledge Management, 10. – 15. August 2010, Gothenburg, Sweden* [online]. IFLA, date submitted: 22/06/2010. Available from WWW: http://www.ifla.org/files/hq/papers/ifla76/149-hannema nn-en.pdf

[6] FLA. *Functional Requirements for Bibliographic Records - Final Report 1998* [online]. International Federation of Library Associations and Institutions, latest revision 11 April 2010. Available from WWW: <http://archive.ifla.org/VII/s13/frbr/frbr1.htm>

[7] JENTZSCH, A.; ISELE, R.; BIZER, Ch. *Silk - Generating RDF Links while publishing or consuming Linked Data* [online]. Poster at the International Semantic Web Conference (ISWC2010), Shanghai, China, November 2010. Available from WWW: <http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/JentzschIseleBizer-Silk-Poster-ISWC2010.pdf>.

[8] Library of Congress. *What is a MARC record, and why is it important* [online]? Library of Congres, 10/27/2009. Available from WWW: <http://www.loc.gov/marc/umb/um01to06.html>.

[9] Promoting Open Access. In SWAN, A., CHAN, L. *Open Access Scholarly Information Sourcebook* [online]. Last Updated on Sunday, 22 August 2010. Available from WWW: <http://www.openoasis.org/index .php?option=com content&view=article&id=133&Itemid=257>.

[10] SCHWARTZ, Ch. Future of MARC and the semantic web. In *Cataloging futures : a work in progress* [online]. Tuesday, October 02, 2007. Available from WWW: <http://www.catalogingfutures.com/catalogin gfutures/2007/10/future-of-marc-.html>.

[11] W3C. *Ontology Dowsing* [online]. ESW Wiki, last modified on 30 November 2010. Available from WWW: <http://esw.w3.org/Ontology Dowsing>.

This article should be cited as:

HLADKA, J. & MYNARZ, J. & SKLENAK, V., 2012. Experience with transformation of bibliographic data into linked data, *Journal of Systems Integration* 3(1), pp. 54 - 62. [Online] Available at: http://www.si-journal.org. ISSN: 1804-2724