DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013

Integration of Research Data and Research Data Links into Library Catalogues

Dominique Ritze
Mannheim University Library, Germany
dominique.ritze@bib.unimannheim.de

Katarina Boland
GESIS - Leibniz Institute for the Social Sciences, Germany
katarina.boland@gesis.org

Abstract

Traditionally, research data and publications are held in separate systems. This results in a disadvantageous situation for researchers as they need to use a variety of different systems to find relevant information about a topic. We therefore face the challenge to overcome the boundaries between bibliographic records and research data by providing an integrated search environment for publications and research data. Because of the inherently different system structure and the diverse metadata for publications and datasets respectively, one type of data cannot easily be integrated into an information system designed for another data type. We present the challenges that arise when adapting a bibliographic library system to include the additional data and give recommendations for an efficient implementation. By presenting our enhanced prototype, we show the applicability and practicability of our proposed solutions. Since our library catalogue prototype features links between publications and underlying research datasets, we provide direct access to metadata of research data stored in remote research data repositories and thus connect both types of information systems.

Keywords: digital libraries; Primo enrichments; integrating research data into library catalogues; linking of library catalogues and research data repositories

1. Introduction

In the social sciences and other empirically oriented fields of research, primary data from surveys, interviews and other studies lay the basis for publications and the continuing research process. Due to the advances in technologies, research data can by now be published in electronic form. This significantly simplifies the access to the data and gives the opportunity for repeated processing by means of modern computer technology. Currently, primary data and publications are stored in separate and structurally distinct systems. Libraries mostly concentrate on publications while research institutions focus on research data.

Searching for connections between primary data and published findings can be very complicated and time-consuming. To find implicit or explicit references to research data, the researcher often has to read through the whole publications. Subsequently, he or she has to search in specific data repositories in order to find information on the datasets. This extensive process decreases the verifiability of research findings and makes data-reuse difficult. In the project *InFoLiS* (Integration of Research Data and Literature in the Social Sciences), GESIS, the Mannheim University Library and the Mannheim University face the challenge to overcome the boundaries between bibliographic records and research data by establishing links between the different data types and by integrating them into the different information systems.

In this paper, we focus on the integration of research data metadata and research data links into library catalogues. For this, we use the catalogue of the Mannheim University Library which is based on Primo, the resource discovery system by ExLibris[1]. Primo is used by a large number of

[1] http://www.exlibrisgroup.com/

https://doi.org/10.23106/dcmi.952136228

institutions (~1900[2]). We obtain the research data from da|ra[3], the registration agency for social and economic data. Since library catalogues are specialized in the presentation of publication metadata, several adaptations have to be made in order to integrate the metadata of research data:

1) the metadata of research data has to be acquired,

2) transformed into a compatible format and

3) loaded into the system together with the links between research data and publications.

In the following, we first discuss the mapping and transformation of the data (Section 2). We then explain the harvesting of the data and their integration into a library catalogue (Section 3). Afterwards, we introduce a method to create the links between research data and publications and show the integration (Section 4). Finally, we conclude with some recommendations on the technical implementation and describe the experience we gained (Section 5).

2. Mapping and Transformation

The very first step is the analysis of the research data metadata, more precisely, their schemata. To integrate data into a system, the data has to be represented in a compatible format. In Primo, metadata needs to be in a format compatible to PNX (Primo Normalized XML). Thus, we have to define some transformation/normalization rules and apply them on the research data metadata. At this point, we also have to decide which information we like to display in Primo. For example, the title of a study is surely relevant while a detailed description of all variables might not be important in this context. In TABLE 1, we list a few examples for the mapping of the current da|ra study metadata schema (Hausstein et al., 2013) to the bibliographic metadata schema in Primo. Several fields can be directly mapped, e.g. titleName to title. Others can only be mapped to similar concepts, e.g. principalInvestigator to creator. Finally, a few fields do not have any correspondence. For these fields, we have to use additional custom fields that have no predefined meaning like lds01 to indicate the dataCollector. Primo provides these fields to display any kind of further information not covered in the remaining fields. By providing a link to the research data itself, all additional information can be gathered. After exporting the data into the required format, it can be loaded into the system.

TABLE 1: Mapping da|ra to Primo

Primo metadata field	da|ra metadata field
title	titleName
creator	principalInvestigator
lds01	dataCollector

3. Data Integration

We investigate two different ways of integrating metadata of research data into Primo: loading a transformed database dump and harvesting an OAI-PMH interface.

First, we use an XML-dump from the da|ra database containing the metadata in XML format. By applying the mapping described above, we are able to transform them into the compatible PNX format. Such a transformation cannot be performed in Primo itself; instead, an additional program has to be implemented and executed. Afterwards, the adapted dump can be loaded into Primo. This procedure has several advantages and disadvantages:

Advantages:
- This solution always works, even if no interface like OAI-PMH is available.
- The complete set with all metadata fields and any kind of available information can be used.

[2] According to ExLibris: http://www.exlibrisgroup.com/category/PrimoOverview, visited July, 2, 2013
[3] http://www.da-ra.de/

https://doi.org/10.23106/dcmi.952136228

- It is completely up to the library which information to use and how to represent them.

Disadvantages:
- A mechanism has to be devised to keep the metadata up-to-date and to add new data.
- A discipline- and even repository-specific mapping is needed for the data transformation.
- The whole transformation process has to be performed outside of Primo.

The second method to integrate the data into Primo is to use an OAI-PMH interface. The da|ra metadata is available via the DataCite[4] OAI-PMH interface as da|ra uploads the data according to the service level agreements with the respective data centers. It can be retrieved in the Dublin Core (DC) format. Because Primo supports OAI-PMH, the metadata can be directly harvested in DC format. Due to the organization of the records as sets within OAI-PMH interfaces, it is possible to only receive the records published by GESIS.

All advantages and disadvantages are contrary to the ones of the other approach. It is very easy to keep the data up-to-date since the OAI-PMH interface can be easily queried in certain intervals. Furthermore, the mapping of DC to the internal metadata schema is straightforward as both schemata describe bibliographic records. Of course, the mapping from the domain-specific schema to DC has to be performed, but in this case, it is the task of the repository provider, i.e. da|ra. This also has the advantage that changes of the repository metadata schema will not affect the harvested data. Since the OAI DC format only covers the most fundamental metadata fields, domain-specific information may get lost, i.e. the geographical coverage of a study cannot be directly matched to a DC field and is not included in the metadata harvested via OAI-PMH.

In conclusion, both ways of acquiring and integrating research data metadata are feasible but both have their advantages and disadvantages. Whenever an OAI-PMH interface is available and a potential information loss is acceptable, it is advantageous to use this interface since it can be completely handled by Primo. However, the presentation in Primo is the same for both methods.

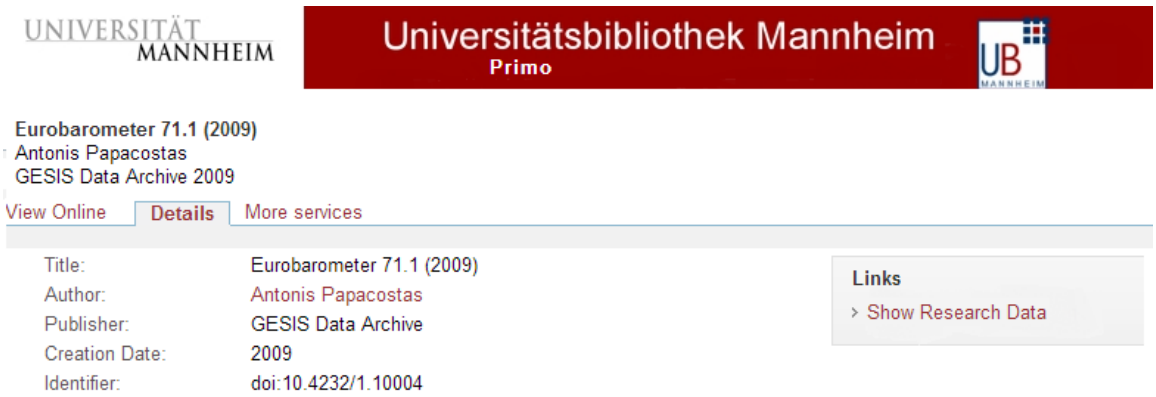

FIG. 1. Integration of research data metadata in Primo

FIG. 1 shows an example (excerpt) of a research data metadata record integrated in Primo. Beside the information about title, author etc., also a link to the resource itself is provided. By clicking on the "Show Research Data" link, the DOI is resolved and the user is forwarded.

4. Links between Research Data and Publications

Since links between research data and publications are rarely included in the metadata of publications or research data, the first step is to identify such links. For this, we employ methods developed in the InFoLiS project. In our prototype, we link publications of the Social Science

[4] http://datacite.org/

https://doi.org/10.23106/dcmi.952136228

Open Access Repository SSOAR[5] to research data of da|ra and vice versa. We chose SSOAR because it contains a large amount of freely available full texts, but our methods are not restricted to this specific data. Since the Mannheim Library catalogue mostly represents monographs and collections, we created records for all SSOAR publications on the article-level. In principle, links between research data and collections can be established as well, but we consider more fine-grained links to be more helpful for the users. Once detected, the links can be integrated into the systems which can be achieved in different ways.

4.1. Generation of Links between Research Data and Publications

Links between publications and research datasets are generated automatically by employing an iterative bootstrapping approach on publication full texts (Boland et al., 2012). It starts with an arbitrary study name as seed, searches for this name and extracts the contexts, e.g. the words and characters surrounding the name. Afterwards, it continues with the search of these contexts to find further study names. The algorithm terminates if neither new contexts nor new study names can be found. As a result, a list of referenced study names (strings) for each publication is returned. In summary, the algorithm learns typical patterns for references to research data which can then be used to identify research data references in unseen texts of the same domain.

After applying the algorithm, the detected study names need to be mapped to entities in a dataset repository. Otherwise, the links only refer to a dataset name but not to the dataset itself. While the algorithm identifies the name of a study, year, version or number specifications have to be extracted from the reference string in a separate step. Together with the study title, this information is used to search for corresponding studies and their DOIs in the da|ra repository.

For publications, the URN is extracted from the SSOAR metadata to serve as an identifier for this data type. For our prototype presented here, we use a manually checked subset of generated links. These are stored in a simple CSV file for further processing.

4.2. Integration of Links between Research Data and Publications

When integrating the links into information systems like Primo, at least two alternatives exist: loading the links directly into the system (server-side enrichment) or merely displaying the links without saving them (client-side enrichment).

Server-side Enrichment: Depending on the technique used to integrate the metadata of research data, several ways are possible to load the links into the system. If the metadata is integrated using a dump, the links can simply be added before loading the records into Primo. This way, the links are just another piece of information in the metadata.

When the integration via OAI-PMH is applied, we cannot directly add further information. However, we can use Primo enrichments (ExLibris, 2013) to include the links. Primo enrichments are Java applications that receive a PNX record. This record can be modified, e.g. fields can be added/removed/changed, before the record is saved in the database. All modifications can be applied by using simple XML APIs like SAX Parser. To enrich the metadata with links, the enrichment application checks whether the according DOI/URN has an entry in our link file. If so, the link is inserted into the record and written to the database. During our project, we implemented such functionalities to enrich the harvested records.

Client-side Enrichment: Even if no additional data shall be loaded into an information system, the links can be displayed to the user. In this case, a minimal-invasive solution can be applied, e.g. by using the Primo Plugin API developed at the Mannheim University Library (Ritze et al., 2012). With this API, it is possible to superimpose additional information like the link to related research data via JavaScript. Therefore, the link file needs to be located at a server such that the information are accessible during the runtime.

[5] http://ssoar.info

https://doi.org/10.23106/dcmi.952136228

Similar to the integration of the research data metadata itself, a server-side enrichment needs to implement some proper update-mechanisms while a client-side enrichment always displays the most up-to-date links as they are integrated at runtime. Since we want to load the research data metadata into our system in order to make it searchable through our catalogue, we choose server-side enrichment for our purpose. Again, the presentation of the links looks exactly the same. An example is depicted in FIG. 2 where an additional link to related research data is shown.

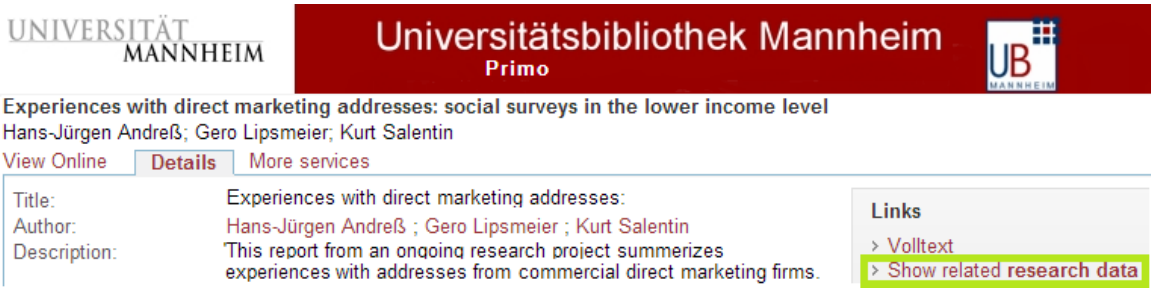

FIG. 2: Integration of linked research data in Primo

5. Conclusion

We presented our approach to integrate metadata of research data and links between research data and publications into library catalogues, e.g. Primo. We described the challenges arising when adapting bibliographic information systems to incorporate this different kind of data. One task is to map the metadata vocabulary of the research data to enable the compatibility. If additional metadata that is relevant for describing research data but not for describing publications shall be included, this may require the definition of custom metadata fields or a usage of existing metadata fields that diverts from their intended use. After the transformation, the metadata can be integrated into the library catalogue. However, the mapping is very domain- and repository- specific and it might not be possible to reuse it when integrating other data sources. Further, some kind of update mechanism is required to keep the data up-to-date. Instead of integrating metadata of research data using a dump, it is also possible to harvest the data via an OAI-PMH interface, if available. Thus, they are present in a standardized format like DC and the whole integration can be performed within the information system itself. Finally, we gave insights into the creation and integration process of links between research data and publications. We implemented a server-side enrichment to integrate the links into our system by making use of Primo enrichments. Altogether, we are able to present a working prototype based on our recommendations for integration of metadata of research data and links into library catalogues.

As part of our future work, we will conduct a user-study to determine the usefulness of our enrichments for finding relevant literature and datasets for research tasks in the social sciences.

Acknowledgements

This work is funded by the DFG as part of the InFoLiS project (SU 647/2-1). We thank Bettina Kaldenberg, Benjamin Zapilko and the da|ra team, especially Brigitte Hausstein, Dimitar Dimitrov and Nicole Quitzsch, for their support.

References

Boland, Katarina, Dominique Ritze, Kai Eckert, and Brigitte Mathiak. (2012). Identifying References to Datasets in Publications. Proceedings of the Second International Conference on Theory and Practice of Digital Libraries, 2012, 150-161.

ExLibris. (2013). Back Office Guide Version 4.x. Retrieved March, 28, 2013.

https://doi.org/10.23106/dcmi.952136228

Hausstein, Brigitte, Nicole Quitzsch, Kirsten Jeude, Natalija Schleinstein, and Wolfgang Zenk-Möltgen. (2013). da|ra Metadaten Schema Version 2.2.1. GESIS Technical Reports.

Ritze, Dominique, and Kai Eckert. (2012). Data Enrichment in Discovery Systems using Linked Data. Proceedings of the 36th Annual Conference of the German Classification Society, 2012, to be published.

https://doi.org/10.23106/dcmi.952136228