Linking Library Data for Quality Improvement and Data Enrichment

Uldis Bojārs, Artūrs Žogla and Elita Eglīte
National Library of Latvia, Mūkusalas iela 3, Riga, Latvia

Keywords: Data Quality, Library Authority Files, Linked Data, Library of Congress Subject Headings, VIAF.

Abstract: Dataset interlinking holds the potential for data quality improvement and data enrichment as demonstrated by the Linked Open Data project. This paper explores the library domain characterized by carefully curated datasets that require high quality standards. It presents the results of an experiment in dataset quality improvement and data enrichment conducted by linking library datasets and analysing the results. The experiment was performed using subject authority files from the National Library of Latvia and the Library of Congress. The paper concludes by discussing how Linked Data can be used for data enrichment.

1 INTRODUCTION

The interlinking of machine-readable datasets creates the potential for rich reuse of information contained within these datasets. The increased availability of open data on the Web provides a large number of datasets that could be linked to one another. Connections between datasets can have multiple uses such as improving data quality by comparing information from multiple sources, enriching datasets with information from other linked datasets and facilitating the development of data-based applications that use these datasets.

Open datasets available on the Web create new opportunities for dataset linking and enrichment. Especially interesting in this context is Linked Open Data (LOD) that applies the principles of Linked Data to publishing datasets on the Web and linking them together, creating a web of interlinked, open datasets that is also known as the Linked Open Data cloud (Bizer et al., 2009). A significant portion of this cloud consists of library-related information such as the Virtual International Authority File (VIAF) and Linked Data sources provided by the Library of Congress (Summers et al., 2008; Hickey and Toves, 2014).

This paper presents work in progress on interlinking library data at the National Library of Latvia (NLL). In particular, it explores how interlinking of library authority records helps improving data quality by identifying and correcting errors, and how library datasets can be enriched using information from other related data sources.

2 METHODOLOGY FOR DATASET LINKING AND QUALITY IMPROVEMENT

Dataset linking activity consists of (1) the dataset *analysis phase* and (2) the *matching phase*. This study focuses on record-level linking of library datasets but the same approach can be applied to other types of data.

During the *analysis phase*, experts examine the datasets involved and determine which datasets to link and how to perform record matching. If both datasets to be linked are not provided in advance (e.g. an organization aims to link its data to the open data available on the Web), then the first analysis task is choosing a dataset to link to. The next task in the analysis phase is to define the matching function f(x,y) that indicates if the two records are equivalent. This task consists of identifying record fields from both datasets that will be used in the matching function and the transformations that may be necessary in order to make the records comparable.

During the *dataset matching phase* the matching function is applied to pairs of records from both datasets. Links are created between pairs of matching records. These links may be recorded in one or both datasets (thus enriching the datasets).

Bojārs U., Žogla A. and Eglīte E..
Linking Library Data for Quality Improvement and Data Enrichment.
DOI: 10.5220/0005554901840188
In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (DATA-2015), pages 184-188
ISBN: 978-989-758-103-8

The results of dataset linking consist of two types: links between records; and errors detected. Errors may be found during linking or while reviewing links for false positives and false negatives. Both the links and the errors may be used for improving these datasets.

We will distinguish between (a) errors that impact data linking (i.e. they influence whether or not the records involved will match) and (b) those that do not. In the latter case, dataset errors can be detected by comparing information in related records from both datasets (e.g. if the fields that should be the same differ between the datasets). If one of the datasets can be defined as the authoritative source then errors can be fixed automatically by using data from the authoritative dataset.

The former case – errors that affect record linking – is more challenging. In the case of *false negative* results the errors prevent us from finding matching records (i.e. the records that are related and should be linked together) and taking further steps for data improvement and enrichment. There may also be *false positive* results where records are mistakenly linked to records from the other dataset that they are not related to. These false matches may result in doing "quality improvement" over incorrect data, leading to further errors. The errors that affect record matching can be detected by domain experts reviewing the results of dataset linking (or a subset of the results) or by comparing the results a golden standard of known correct links. Consequently, the matching algorithm may be improved so that it works around these errors or alerts users about errors that should be fixed. As a result, dataset linking is an iterative process where results of the initial runs of the linking algorithm are used for improving the next runs and the overall quality of datasets involved.

In this paper we examine the process of linking the National Library of Latvia subject authority records where errors in record label fields directly affect the results of linking. Dataset quality is improved by (1) attempting to link the datasets, (2) examining the results for false positives and false negatives and (3) modifying the matching function to take into account the errors detected and to increase the matching precision and recall.

3 DATASET

The dataset used in this experiment is a random sample of NLL's subject authority records consisting of 1280 entries or ~3.5% of all NLL's subject authority records. The size of the dataset was chosen small enough to enable manual validation of record linkage, results of which are reported in Section 4.

Library authority files[1] are controlled vocabularies that provide standard names and identifiers for different types of entities – people, organisations, places and concepts (subjects) – that library catalogues may need to refer to in a unified way (Hickey and Toves, 2014). Authority records contain at least a preferred name of the concept and an identifier, and may also contain alternative labels (e.g. other spellings) and links to related records.

NLL's subject headings (NLL-SH) – a taxonomy of topical terms used by libraries in Latvia – was developed based on the Library of Congress Subject Headings (LCSH). Most of NLL-SH records were adapted from LCSH by translating preferred labels to Latvian while other records, specific to NLL and Latvia, were introduced without having a matching LCSH concept (Stūrmane et al., 2014). These two datasets – LCSH and NLL's subject headings – were selected to be linked as a part of this experiment because (1) it could be expected that a majority of NLL-SH records would be linked to LCSH; (2) library experts were available that could examine the results of linking the two datasets in order to identify false positive and false negative results.

NLL's name authority records were considered as another candidate for linking but they are already linked as a part of a large-scale Linked Data project VIAF[2], which interlinks authority records from libraries worldwide. Linking in VIAF is done using both the authority records and bibliographic records associated with them (Hickey and Toves, 2014).

3.1 Data Formats

The experiment involved linking library authority data represented in different formats. The LCSH dataset available online via the Library of Congress Linked Data Service[3] is represented in SKOS (Summers et al., 2008; Miles and Bechhofer, 2009). It consists of taxonomy concepts that each have a preferred label (*skos:prefLabel*) and may have a number of alternate labels (*skos:altLabel*). Concepts may have links to other concepts both inside the

[1] http://en.wikipedia.org/wiki/Authority_control
[2] http://viaf.org/
[3] The experiment used a SKOS version of the LCSH dataset, in N-Triples RDF serialization, published on 27-Oct-2014. Library of Congress Linked Data service is available online at http://id.loc.gov/download/.

taxonomy and outside it (e.g. links to related concepts in datasets from other libraries).

The NLL-SH dataset uses the MARC21 format for authority data[4]. Similar to LCSH, NLL-SH records have one preferred label and may have a number of alternate labels. These records consist of MARC fields whose type is identified by 3-digit numbers. In the case of subject heading records the preferred label is located in MARC field 150 while alternate labels use field 450. Other fields may contain additional information such as links to records that describe broader, narrower or related taxonomy concepts.

3.2 Matching Algorithm, Applied to Experimental Datasets

An NLL-SH record that has an equivalent LCSH record should contain the LCSH record's English label as one of its alternate labels. This should allow us to link both datasets (based on English language alternate labels of NLL-SH records) and to evaluate the quality of the NLL-SH dataset.

Authority records may either have simple labels or complex labels consisting of multiple components[5]. The way in which complex labels are represented differs between the two datasets: NLL-SH records use separate MARC subfields for the components of complex labels while the LCSH SKOS dataset concatenates label components using a "--" separator.

In order to make records from both datasets comparable, the matching function converted complex NLL-SH labels to the same format as used by the LCSH SKOS dataset.

The matching algorithm iterates through all pairs of NLL-SH and LCSH records. The matching function compares all alternate labels (MARC field 450) of a given NLL-SH record to the preferred label of the LCSH concept and returns True if any of them match.

4 RESULTS

This section describes the results of the dataset linking experiment. By using string equality we were able to link 82.7% of NLL-SH records (1058 out of 1280 records) to matching LCSH records. The linking algorithm identified matches for both simple and complex topics. Matching records in LCSH were identified for 88.3% of simple headings in NLL-SH (628 of 711 records) and 75.6% of complex headings (430 of 569 records).

4.1 Analysis of Dataset Linking Errors

The results were examined by library metadata experts in order to identify false positives and false negatives. All positively identified matches were valid (i.e. we did not find any false positives) but there were 32 false negatives (2.5% of NLL-SH records in the experimental dataset) that had relevant LCSH entries but were not matched to them. Table 1 lists the types and the number of linking errors encountered.

Table 1: Types of dataset linking errors.

#	Error type	Errors
1	Different apostrophe characters used in NLL-SH and LCSH	18
2	Shortcomings of matching algorithm	2
3	Other errors (spelling mistakes, missing or incorrect MARC fields, etc.)	12
	Total:	32

The most common were errors caused by differences in the apostrophe symbols used in these NLL-SH records and matching LCSH records. Since NLL's dataset records must have the same English labels as in the LCSH dataset this is considered an error. This is easy to fix by using the same apostrophe symbols as in LCSH.

The second type of error is where NLL-SH records did not contain MARC fields 450 with English labels because they were identical to the Latvian labels (MARC field 150). Consequently, the matching algorithm was improved to include NLL-SH preferred labels (field 150) in searching for records to interlink.

The remaining errors were spelling mistakes (e.g. "ltierature" instead of "literature"), use of singular instead of plural and other differences between labels in two datasets, as well as missing or incorrect MARC fields (e.g. the English label was added to field 430 instead of 450).

Especially interesting was one record with the apostrophe error because it also had a *semantic error* where a record of different meaning was identified by the algorithm as a match. The NLL heading for

[4] The details of MARC21 data formats are beyond the scope of this paper but we provide some information that is necessary for understanding MARC record structure. http://www.loc.gov/marc/authority/

[5] For example, “History” and “Latvia” are simple labels while “Latvia--History” is a complex label, combining both topics.

this record was "Men's magazines" ("Periodiskie izdevumi vīriešiem" in Latvian) but its English label was incorrect and pointed to another LCSH record: "Women's periodicals". Had it not been for the apostrophe error the matching algorithm would have missed the more serious error that was detected by metadata experts when reviewing matching results. An important task for future study is how to detect such semantic errors and attempt to correct them.

The next section examines how the results of this dataset linking experiment can be used for dataset quality improvement.

4.2 Data Quality Improvement

Once the information about the most common errors detected is available it can be used to improve the quality of data. This paper examines how data quality can be improved by linking datasets to one another. As discussed in Section 2, data quality issues can be discovered: (a) by analyzing the errors found while linking datasets; or (b) by comparing linked records from both datasets.

In the case of linking NLL-SH and LCSH data there are no other fields that should be the same except for English labels used in linking. Therefore for data quality improvements we concentrated on fixing the errors that affected the linking process.

The matching algorithm was improved, taking into account the errors discussed in Section 4.1: (1) by adding the use of preferred labels of NLL-SH records to the matching function; and (2) by introducing fuzzy record matching using the Damerau–Levenshtein edit distance metric that takes into account character transposition.

The improved matching algorithm uses fuzzy matching with edit distance 1 (detecting errors where labels differ by no more than 1 edit operation) on all NLL-SH records that were not matched using string equality. This approach detected most of the errors identified when linking datasets including apostrophe errors, extra dots at the end of labels and other spelling errors.

The second iteration of the matching algorithm was not aimed at detecting spelling errors that had edit distance larger than one. It could be modified to allow for larger edit distances however this is not advisable because even at distance 1 there were false positives (e.g. "19th century" instead of "18th century"; "SETL" instead of "SEAL") that would end up introducing errors in data if not spotted during review.

The six remaining errors of type 3 cannot be detected just by fuzzy matching. Four of these cases were errors in MARC fields, for example, English labels not found in field 450 (sometimes misplaced in other MARC fields). In the remaining two cases: (a) the record's English label was different from LCSH; (b) a component of a complex label was not translated to English.

Fuzzy matching is useful for identifying records that are similar (e.g., it helped us to find two spelling errors that were not found by metadata experts when reviewing the results of the initial matching run). However, in order to further improve data quality, errors need to be classified based on how certain we can be that fixing them leads to a valid match and not a false positive.

Based on data quality requirements "harmless" errors (e.g. an extra dot at the end of the label) can be fixed automatically or suggested to the editor as likely fixes while more serious cases that may lead to false positives (see above) need to be handled with extra care. By examining the false positives we may identify a set of conditions that can help to determine which cases need an extra review (e.g. to warn about fuzzy matches in numbers or abbreviations).

5 DATA ENRICHMENT

Links between datasets provide an opportunity for enriching the datasets involved. Data enrichment can take place at the time of linking or on the fly, when requesting information from datasets.

Data enrichment is a complex task and details of how it can be performed depend on the datasets involved. For example, selected data record fields may be copied from one dataset to the other, converting and merging data as necessary. In the case of taxonomies, such as library authority data, linked records from both taxonomies may contain labels in different languages and these records can be enriched by copying labels across datasets, facilitating creation of multilingual taxonomies.

Authority data records may contain links, both internal and external, that can be a valuable resource for data enrichment. Once a link between NLL-SH and LCSH records is established, NLL-SH records can be enriched with links to authority records from the National Library of France and the German National Library that are included in the LCSH dataset. Links from NLL-SH to other open datasets that link to LCSH, for example, to authority records from the National Diet Library, Japan, may also be established. The resulting network of authority data

would be a useful tool for facilitating multilingual discovery of cultural heritage information.

The fact that a link between records has been discovered is valuable information by itself and this linkage may be recorded in one or both datasets. These external links may later be used for enriching datasets "on the fly" or for monitoring changes to the linked dataset. An example of this approach is the *datos.bne.es* service from the National Library of Spain which uses the already established links to VIAF in order to enrich their records with links to the authority records of other national libraries (Vila-Suero et al., 2013).

Linked Data is a technique for publishing data on the Web in a way that facilitates object interlinking and data access "on the fly" (Berners-Lee, 2006; Bizer et al., 2009). It publishes data so that data identifiers (URIs) can be *dereferenced* (i.e. users can access structured information about these objects online, by making HTTP requests) and provides a way for including URIs of linked objects in the data published.

Information published as Linked Data (e.g. LCSH dataset used in the experiment) is well-suited for data enrichment: (1) data is published on the Web, making it possible for users to find it, reuse it and link to it; (2) the Linked Data model makes it easy to enrich records with new information; and (3) these records have web-accessible URI identifiers for accessing up-to-date information about them.

The National Library of Latvia is in the process of publishing NLL's authority data as Linked Data. Once this dataset is published it will enable the benefits listed above such as the opportunity for other users to explore and link to NLL's authority data. The data published by NLL's linked data service will be enriched with additional information including Linked Data from other data sources.

6 CONCLUSIONS

Dataset interlinking creates new opportunities for data quality improvement and data enrichment.

This paper discussed principles for dataset linking and improvement, and presented results of an experiment for linking and enriching library authority data.

The experiment was conducted using the National Library of Latvia authority file and Linked Data from the Library of Congress. The experiment helped us identify and fix data quality issues in the NLL-SH dataset, and to enrich it using information from matching LCSH records. Links between taxonomy records from the two datasets may be used for multilingual discovery of bibliographic data.

Datasets that are published as Linked Data are especially useful for data enrichment as their records are available "on the fly" and may include links to other related datasets. The National Library of Latvia is in the process of publishing its authority file as Linked Data, making it possible for user worldwide to reuse it and to interlink it with other datasets.

ACKNOWLEDGEMENTS

This research is a part of the project "Competence Centre of Information and Communication Technologies" by IT Competence Centre, contract No.
L-KC-11-0003, co-financed by European Regional Development Fund, Research No. 1.18 “Data array quality analysis and enhancement technologies”. More information: http://www.itkc.lv/.

REFERENCES

Berners-Lee, T. (2006). *Linked Data – Design Issues*. W3C [online, accessed: 2015-03-31]. Available from: http://www.w3.org/DesignIssues/LinkedData.

Bizer, C., Heath, T. and Berners-Lee, T. (2009). *Linked Data – The Story So Far*. International Journal on Semantic Web and Information Systems, *5*(3), 1-22.

Hickey, T. B. and Toves, J. (2014). *Managing Ambiguity in VIAF*. D-Lib Magazine, 20(7), 3.

Miles, A. and Bechhofer, S. (eds.) (2009). *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation. Available from: http://www.w3.org/TR/skos-reference.

Stūrmane, A., Eglīte, E. and Jankevica-Balode, M. (2014). *Subject Metadata Development for Digital Resources in Latvia*. Cataloging & Classification Quarterly, 52(1), 20-31.

Summers, E., Isaac, A., Redding, C. and Krech, D. (2008). *LCSH, SKOS and Linked Data*. In Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications (DC-2008), pp. 25-33. Dublin Core Metadata Initiative.

Vila-Suero, D., Villazón-Terrazas, B. and Gómez-Pérez, A. (2013). *datos.bne.es: a Library Linked Dataset*. Semantic Web, 4(3), 307-313.