DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2016

Automatic Creation of Mappings between Classification Systems for Bibliographic Data

Magnus Pfeffer
Stuttgart Media University, Germany
pfeffer@hdm-stuttgart.de

Abstract

In this paper, the implementation of an approach to automatically create mappings between classification systems is presented and results from a preliminary analysis are discussed. The approach is based in the idea of instance-based ontology matching and consists of three steps: First, bibliographic data from diverse sources that contain items classified by the required classification systems is aggregated in a single database. Next, an efficient clustering algorithm is used to group individual issues and editions of the same work. It works by matching names of authors and corporate bodies as well as title, subtitle and uniform title. Finally, the clusters containing information from both required systems are added up to create a co-occurrence table. This information is then used to generate candidates for a mapping between the individual classes of the two classification systems.

In an experiment, the implementation is utilized to generate mappings between two classification systems that are in use in Germany. The mappings are evaluated using existing partial mappings that have been manually created by domain experts as a gold standard for comparison. While the automatic mappings might be less accurate and exhaustive than manually created ones they are sufficient for retrieval and visualization purposes and could be further improved by refining the statistical analysis or including more datasets.

Keywords: library catalog, classification systems, instance-based ontology mapping

1. Introduction and Motivation

Classification systems are an important means to provide topic-based access to library collections. Depending on the collections at hand and the primary use cases, these classification systems can differ significantly in structure and organization. For example, systems used to arrange large collections on shelves need to be sophisticated and highly structured in order to keep the number of members of each class manageable. On the other hand, applications like topic-based facetted browsing in resource discovery systems or graphical representations of the contents of collections benefit from a simpler structure with fewer branches and depth to assure a clearly arranged presentation to the user. With the proliferation of more powerful search solutions in libraries, there is a renewed interest in using different classification systems for search or browsing.

As annotating a library collection using multiple classification systems would be prohibitive, using mappings to derive new annotations from existing data is a possible solution. The creation of such mappings can be an arduous process if done manually, but is still undertaken for applications in information retrieval systems or to assist library collection reorganization. Part of the ongoing projects of the Austrian National Library, as presented in Plößnig (2012) and Plößnig (2014) is the enrichment of the catalog data with annotations using multiple classification systems. For this purpose several partial mappings from the *Regensburger Verbundklassifikation* (RVK, engl.: Regensburg union classification system) to the the *Basisklassifikation* (BK, engl.: basic classification system) have been created manually and are already used to enrich catalog entries.

https://doi.org/10.23106/dcmi.952137572

In this paper, we propose an automated approach to automatically create mappings between classification systems used in libraries. It is based on the idea of instance-based ontology matching, which works on the annotated instances instead of comparing the labels of classes or the structure of the systems. The general applicability of this matching method to data from library catalogs has been shown in multiple projects in the past (Isaac et.al., 2007, Schopman, 2009 and Schopman et. al., 2012) and in own prior work, preliminary data generated from the implementation was used as input for the manual mapping project at the Austrian National Library with positive results (Aigner, 2005). The approach is tailored to library catalog data with its specific properties and its implementation prepared to scale up to very large datasets with more than 100 million entries.

To evaluate the results of the mappings process and to create a baseline for further experiments, a large dataset of catalog data containing entries annotated with RVK and BK classes has been collected and a full mapping was produced using the proposed approach. A relevant subsection of the existing manual mapping results from the projects of the Austrian national Library was selected to be used as a gold standard to evaluate the automated mapping.

The paper is structured as follows: First is a short review on the different methods of ontology matching and the related work on instance-based ontology matching in the library domain as well as prior work of the author that has influenced the development process. Next the implementation specifics of the approach, the design decisions and their inherent tradeoffs are discussed. The second half of the paper focusses on the evaluation: the used datasets and classification systems are introduced in detail and the resulting automated mapping is compared to the gold standard by calculating precision and recall for a range of parameters. The paper closes with a look at further possible enhancements to the approach itself and the current software implementation.

2. Preliminaries and Related Work

Ontology mapping is a vast and very active field of research with many applications in knowledge organization and knowledge representation, especially for the Semantic Web. While the ontologies discussed in this field are often very rich structures expressed in OWL or similar high level languages, there is also an interest in less rich ontologies that can be expressed in SKOS or data formats traditionally used in library information systems. The Ontology Alignment Evaluation Initiative[1] regularly invites participants to compare and benchmark their latest algorithms and includes a library track specifically for this kind of data since 2012 (Aguirre et. al., 2012).

Euzenat and Shvaiko (2007, p. 341) lists four automatic ontology matching methods: terminological, structure-based, semantic-based and instance-based. Terminological methods work on the lexical data contained in concept labels or descriptions and utilize it to find matches by string comparison. Structure-based methods use the relations between concepts to deduce possible matches. Semantic-based methods use generic or domain-specific rules or other background information outside the ontologies being matched. Instance-based then rely on the set of instances that are associated with a given concept. Depending on what type of instances are available, different methods can be applied: If instances exist that are annotated using two ontologies, one can directly analyze the co-occurrence of concepts; the idea being that two concepts are closer related, the more significant the overlap of common instances of two concepts is.

If no such dually annotated instances exist, it is possible to extend the concepts themselves using the contents of the annotated instances and compare these extended concepts. Alternatively one can try to match the instances themselves and create clusters of instances, which are then again the basis for a co-occurrence analysis.

[1] http://oaei.ontologymatching.org/

https://doi.org/10.23106/dcmi.952137572

Instance-based matching has advantageous properties: it is less affected by ambiguities like homonyms or synonyms that are inherent in limited lexical data like labels or short description. Also as the sets of instances are the result of practical application of the ontology on documents, they are a very precise representation of the concepts true meaning. Finally, the method can cope with small annotation errors or variances that are inherent to a manual annotation process that is done by several individuals. On the other hand, it is often difficult to find sufficient instances, i.e. annotated objects or documents.

Instance-based ontology matching has been successfully implemented and used with data from libraries in the past: Isaac et al. (2007) created a mapping between a classification system and a thesaurus based on data from the Dutch National Library and evaluated the result by comparing to an existing manual mapping. In Schopman (2009) this work was extended to include multilingual data from the European Library and the algorithms were further refined. Both reports showed very encouraging and positive results. Finally, in a paper by the same authors, the algorithm and application is further generalized and rigorously evaluated it using large multilingual data sets (Schopman et. al., 2012).

In the library domain, finding a large number of instances is less problematic, as most libraries seek to enable a topic-based search or access by using a classification system or thesaurus to annotate the catalog entries. Nonetheless, it is often not the case that catalog entries are uniformly and consistently annotated: the use of ontologies can change over time or resources may be insufficient to keep up with manual annotations. In Germany, there is an additional complication: due to historic developments, there are several large library unions, each with their own central cataloging database alongside the National Library with its own catalog. Data sharing between these entities has been limited and resulted in very heterogeneous data sets, especially in regards to annotations using classification systems. The author has applied different clustering methods on data sets from German library unions in order to enrich entries with annotations from other library unions and evaluated the results using existing manual annotations as gold standards (Pfeffer, 2009). One important result was that generic clustering methods like k-nearest neighbor based on string similarity tend to create inconsistent clusters, resulting in a low precision for the enrichments, while clustering based on exact matches of title and subtitle and author/corporate bodies resulted in very consistent clusters and very high precision for the enrichments, which was considered to be on par to most manual annotation by indexing experts (Pfeffer, 2013).

Data from these enrichment projects was used to evaluate the usefulness of co-occurrence analysis for the creation of mappings in theses by library science students: In Probstmeyer (2009), a mapping between the Schlagwortnormdatei (SWD, a subject headings authority file used by most German academic libraries) and the RVK was evaluated. Co-occurrence was calculated using the individual catalog entries, and the evaluation showed that the significance of the co-occurrence was not strongly correlated with the relation of the concepts. One reason was that in the catalog data, works with many different editions tended to have the same co-occurring annotations and overshadowed the co-occurances from works which only exist in a single edition. In Aigner (2015), the process of creating a manual mapping between the RVK and BK for the domain of geography is discussed. Here, co-occurrence was calculated using the consistent clusters and the resulting matches were used as one source of possible mappings (besides mostly manual lexical and structural analysis). The analysis showed that the significance of co-occurrence was correlated with the relation of the concepts and after choosing a suitable threshold almost all remaining mappings were deemed highly useful.

3. Data Sets and Implementation

The experiments presented in this paper are a direct result of the lessons learned in preparing the co-occurrence data that was used successfully in Aigner (2015). The implementation used was not running stable, used a lot of computing resources and did not scale well for larger datasets. Beside the performance issues, a new implementation should also be more flexible in

https://doi.org/10.23106/dcmi.952137572

regards to the data used as basis as well as the ontologies to be matched. To assess the properties of the new implementation, the full automatic process was run using several very large datasets mapping the classification systems that have been the focus at the Austrian National Library. This course of action ensured that enough information is available to evaluate the resulting mappings.

In this section, first the classification systems and the data sources used in the experiment are introduced. Next the clustering process and its implementation are presented and explained using a simplified example.

3.1 Classification systems

The RVK was developed in the 1960s as a local classification system for the library of the newly founded Regensburg University. Unlike most existing German university libraries, the collections in Regensburg were planned to be mostly openly accessible by users and this influenced the structure and design of the classification system. It is a monohierarchical universal classification system modelled on the Library of Congress classification (LCC) and consists of 33 domain-specific sections that mirror the structure of German university faculties. Granularity and hierarchies in these domain-specific sections vary to a certain extent, as well as the principles used to create further subdivisions. The RVK consists of about 80.000 classes in total. (Lorenz, 2008)

The RVK has seen continued adoption by other academic libraries and is now the most used universal library classification system in the German-speaking region, being in use at more than 140 libraries. .

All class notations have a common composition: Two uppercase roman letters are followed by a three to six digit number. For example, the notation "QF 100" is from the section "Q: Economics", subclass "QF: History of Economics" and represents "QF 100: History of Economics until 500 A.D.". See figure 1 for an excerpt of the class tree view[2].

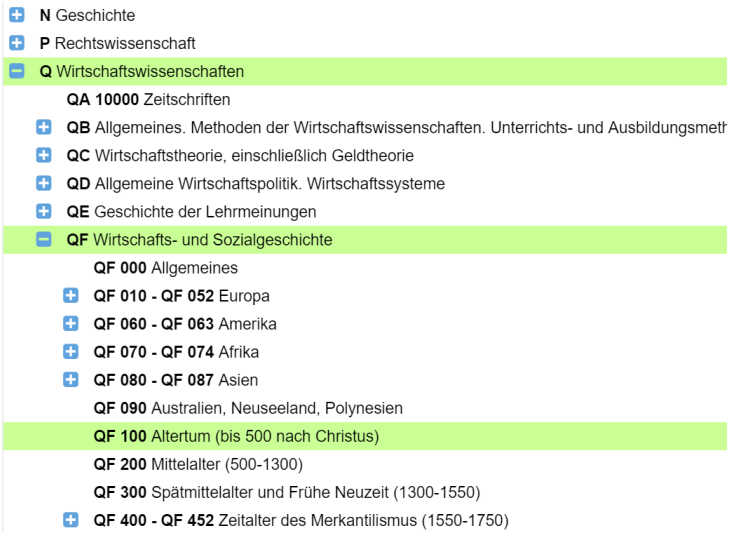

Figure 1: Excerpt from the tree view of the Regensburg union classification system

The BK was originally developed in the Netherlands by the PICA library foundation under the name "Nederlandse basisclassificatie", based on existing domain-specific classification systems used to index bibliographies. It was translated into German in 1992 and adapted by many libraries in the North German region. The BK is a monohierarchical universal classification system

[2] An online version of the full system is available at https://rvk.uni-regensburg.de/regensburger-verbundklassifikation-online

https://doi.org/10.23106/dcmi.952137572

consisting of about 2100 classes that are divided into 48 main divisions. The divisions are modelled after traditional domain structures in the sciences as well as certain interdisciplinary aspects. The classes within each main division are arranged mostly by topic, less often by region or historic timespan. BK was developed as a secondary annotation system that was to be used together with thesaurus-based indexing to provide multiple ways of topical access to collections. (Schulz, 1991)

Class notations are composed of a two-digit number, a dot as a separator and another two-digit number. The first number denotes the main division, the second one the class within that division. For example the notation "15.09: History of Economics" is part of "15: History". See figure 2 for an excerpt of the class view[3].

15.08 Sozialgeschichte
Erl.: Zu verwenden für Sozialgeschichte allgemein ohne zeitliche, räumliche oder sachliche Einschränkung (z.B. Sozialgeschichte des Bürgertum). Darüber hinaus zu verwenden als Zweitnotation zu 15.25-15.96, sofern Darstellungen zu einzelnen Epochen oder Kulturräumen dezidiert sozialgeschichtliche Themen behandeln

15.09 Wirtschaftsgeschichte
Erl.: Zu verwenden für Wirtschaftsgeschichte allgemein ohne zeitliche, räumliche oder sachliche Einschränkung. Darüber hinaus zu verwenden als Zweitnotation zu 15.25-15.96, sofern Darstellungen zu einzelnen Epochen oder Kulturräumen dezidiert wirtschaftsgeschichtliche Themen behandeln

15.10 Historische Hilfswissenschaften
Hier: Historische Geographie <Geschichte>
Verw.: Archivkunde -> 06.90 (Archive, Archivkunde)
Geschichtsatlanten ohne räumliche und sachliche Einschränkung -> 15.20 (Allgemeine Weltgeschichte)
Handschriftenkunde -> 06.10-06.19 (Handschriftenkunde

Figure 2: Excerpt of the class view of the basic classification system

3.2 Data sources

Catalog title data from most German library unions is available as open data in the MARC21 format. For the project, the following catalogs were chosen:

- Gemeinsamer Bibliotheksverbund (GBV, engl.: Common Library Network). Spanning several states in northern Germany, it is the largest library union. Its catalog also includes the collection of the Berlin state library.
- Südwestverbund (SWB, engl.: Southwest German Library Union). Its member libraries are located in the states of Saarland, Baden-Württemberg and Saxony.
- Bibliotheksverbund Bayern (BVB, engl.: Bavarian Library Union). The union catalog contains the collections of libraries from the states of Bavaria, Berlin and Brandenburg.

Table 1 contains some statistics on contents and annotations of the three datasets. Non-monographic entries (like musical notes, dvds, maps, etc.) were filtered using information from the MARC21 field "Leader" and field 007. Annotations were taken from the main title data MARC21 field 084 (subfield 2 values "rvk", "bcl" or "bkl" respectively).

Table 1: Contents of the initial datasets

	All Entries	Monographic entries	Monographic with RVK	Monographic with BK
GBV	32,027,977	24,267,492	0	3,976,154
SWB	18,789,185	16,447,890	4,383,273	0
BVB	26,680,083	23,658,674	7,215,483	0

[3] An online version of the full system is available at https://www.gbv.de/bibliotheken/verbundbibliotheken/02Verbund/01Erschliessung/02Richtlinien/05Basisklassifikation/index

https://doi.org/10.23106/dcmi.952137572

The catalog of the Austrian National Library contains both RVK and BK annotations. As its entries have already been enriched extensively using the results from the manual mapping projects, it was considered to be unsuitable as a data source for this experiment.

3.3 Clustering process

The clustering process is implemented using the Perl scripting language. All data is stored in a NoSQL document database back end using only the very basic features of key-value storage and access. In the implementation for the evaluation experiment, MongoDB on a 16-core server with 16 GB of RAM is used to allow fast access even for large datasets.

In the first step, the original MARC21 data is transformed into a very simple JSON-like data format containing only the most important properties: id, title, subtitle, uniform title, author, corporate entity, publisher, year of publication and the annotations of RVK, BK and the dewey classification system (DDC) as well as index terms from the Gemeinsame Normdatei (GND, engl.: common authority file) used by most libraries in the German-speaking region. Properties that can contain more than one entry, like author or the annotations are stored as lists, all other properties as string literals. The original database ids are used as the access key ids for this *data* table.

In the second step, strings are generated for each entry of the *data* table by creating combinations of all author or corporate entity list entries with the title+subtitle and uniform title. These generated strings are used as access key ids for the *key* and *keyequiv* tables. In the *key* table the corresponding ids from the *data* table are stored as a list. In the *keyequiv* table, the other strings that were generated from the same data are stored in a set. Table 2 shows the resulting entries for a simplified example. Although the entries with 1 and 3 do not share an author, they should become part of the same cluster because they both share author and title with id 2.

To generate the clusters, in the third step the *key* table is traversed: The current id is stored in a set named "done" and all equivalent strings are retrieved from *keyequiv* and stored in a set "todo". As long a "todo" still contains entries, the first entry gets moved from "todo" to "done" and the equivalent strings for it are retrieved from *keyequiv* and stored in "todo" unless they are already contained in "done". Finally each entry of "done" is marked in *key* and the corresponding data ids from *key* are retrieved and stored in a temporary set, which is then saved as a new entry in the *cluster* table. The traversal continues with the next non-marked key in *keys*.

Table 2: Example tables illustrating the MongoDB implementation

data table	key table	keyequiv table
id: 1 author: [A, B] title: beer year: 1990	id: A|beer ref: [1]	id: A|beer eq: [B|beer]
id: 2 author: [B, C] title: beer year: 1995	Id: B|beer Ref: [1, 2]	id: B|beer eq: [A|beer, C|beer]
id: 3 author: [C] Title: beer Year: 1999	Id: C|beer Ref: [2,3]	id: C|beer eq: [B|beer]

The combination of fields to create keys in step 2 can be changed, thus influencing the resulting clusters. For this experiment, only authors, corporate bodies, uniform title and main title have been used. By ignoring the subtitles, the clustering is more aggressive and creates larger clusters, which can in theory lead to more inconsistent clusters. Earlier experiments had shown

https://doi.org/10.23106/dcmi.952137572

that this happens rarely in practice, as the combination of a short title consisting of a common word or phrase and two authors with the same name is highly unlikely.

Applying the clustering process to the data sources results in 21,653,606 clusters, of which 904,876 reference catalog entries that contain BK and RVK annotations. The co-occurrence data was then generated and for each pair of BK and RVK classes that occurred at least in one cluster, the final table containing the RVK class notation, the number of dually annotated clusters that were annotated with this RVK notation, the BK class notation that co-occurred, and the number of dually annotated clusters that were annotated with the exact pair. Co-occurrence data for 1,155,552 such pairs was found.

The whole process ran very stable and reliably, using only a small part of the server resources. The whole process, from importing the data sets to finished co-occurrence table took less than 3 days.

4. Evaluation

To assess the quality of the co-occurrence data and to determine possible thresholds to filter the data, an existing manual mapping from RVK to BK for the domain of economics was chosen for comparison. The mapping was provided by the Austrian National Library and was done by Andreas Waldhör, who had done a mapping for the domain of law as a Master's thesis (Waldhör, 2012). It contains 963 individual mappings from the "Q: Economics" division of the RVK to the BK; mapping each RVK class to exactly one BK class. The corresponding selection from the co-occurrence data contains 44710 pairs, with the strongest co-occurrence being 3195 clusters sharing a specific pair.

Of the 963 manual pairs, 808 were also found in the co-occurrence list, resulting in a maximum recall of 0.839 with a precision of 0.018. Of the missing 155 pairs, only 14 contained RVK classes that were completely missing in the co-occurrence data, while the RVK classes of the other 141 pairs appeared in co-occurrence, but with different BK classes.

Two parameters were selected for filtering the raw co-occurrence data: first, the ratio of the number of clusters with a given pair to the number of pairs containing the same RVK class and second, the absolute number of clusters with a given pair. The first is a Jaccard-like measure with a maximum of 1, when all clusters that contain the RVK class from a given pair also contain the BK class. The ratio is smaller, the more clusters with the same RVK class but different BK classes exist. It was preferred over the classic Jaccard measure, i.e. the ratio of the number of clusters with a given pair to the number of pairs containing the RVK class *or* BK class of the pair, because of the imbalanced size and structure of the two classification systems being mapped: As RVK contains far more classes, any BK class is expected to be correctly mapped to a high number of RVK classes. Including the number of pairs with the BK class as well would have led to significantly higher numbers, which would in turn result in very small ratios that are harder to compare. With the goal of a mapping from RVK to BK (and not vice versa) in mind, the chosen ratio was considered to be far superior.

The second parameter can be used to filter pairs that only occur in few clusters. Tables 3 and 4 contain the precision and recall results for a range of values for both parameters. The results are decent, but not overly impressive. It is interesting to see that increasing the required number of clusters results in a significant increase in precision while the recall is not affected very much. The ratio on the other hand affects both precision and recall, with a quickly decreasing gain on precision for ratios of 0.6 and more.

https://doi.org/10.23106/dcmi.952137572

Table 3: Precision results. Values >0.5 are highlighted

	ratio ≥ 0	ratio ≥ 0.1	ratio ≥ 0.2	ratio ≥ 0.3	ratio ≥ 0.4	ratio ≥ 0.5	ratio ≥ 0.6	ratio ≥ 0.7	ratio ≥ 0.8
num ≥ 0	0.0181	0.1639	0.2177	0.2410	0.2383	0.2015	0.1759	0.1100	0.0553
num ≥ 2	0.0183	0.1979	0.2979	0.3769	0.4436	0.4236	0.6218	0.6319	0.5333
num ≥ 4	0.0179	0.2129	0.3499	0.4989	0.5918	0.6269	0.7067	0.7288	0.7143
num ≥ 6	0.0173	0.2222	0.3954	0.5177	0.6353	0.6724	0.7525	0.7714	0.7561
num ≥ 8	0.0171	0.2308	0.4053	0.5280	0.6529	0.6951	0.7814	**0.8125**	**0.8056**
num ≥ 10	0.0167	0.2386	0.4089	0.4206	0.6603	0.7066	**0.7877**	**0.8261**	**0.7941**

Table 4: Recall results. Top 5 values are highlighted

	ratio ≥ 0	ratio ≥ 0.1	ratio ≥ 0.2	ratio ≥ 0.3	ratio ≥ 0.4	ratio ≥ 0.5	ratio ≥ 0.6	ratio ≥ 0.7	ratio ≥ 0.8
num ≥ 0	**0.8390**	0.6947	0.5940	0.4922	0.3801	0.2835	0.1817	0.0987	0.0457
num ≥ 2	**0.8349**	0.6906	0.5898	0.4881	0.3759	0.2793	0.1776	0.0945	0.0415
num ≥ 4	**0.8089**	0.6646	0.5639	0.4621	0.3583	0.2617	0.1651	0.0893	0.0363
num ≥ 6	**0.7809**	0.6366	0.5358	0.4403	0.3364	0.2451	0.1547	0.0841	0.0322
num ≥ 8	**0.7653**	0.6210	0.5265	0.4309	0.3281	0.2368	0.1485	0.0810	0.0301
num ≥ 10	0.7487	0.6044	0.5130	0.5315	0.3229	0.2326	0.1464	0.0789	0.0280

In order to get threshold values that balance precision and recall, f-measures were calculated. Table 5 contains the results for the f-measure, with double weighted precision. The higher weight for precision was chosen with the intended use cases in mind: using the mapping for enrichment in catalogs or as a basis for creating manual mappings would be significantly negatively affected by low precision results, and less by low recall results.

Table 5: f-measure, with double weighted precision. Top 5 values are highlighted

	ratio ≥ 0	ratio ≥ 0.1	ratio ≥ 0.2	ratio ≥ 0.3	ratio ≥ 0.4	ratio ≥ 0.5	ratio ≥ 0.6	ratio ≥ 0.7	ratio ≥ 0.8
num ≥ 0	0.0270	0.2322	0.2991	0.3221	0.3090	0.2566	0.2124	0.1290	0.0637
num ≥ 2	0.0273	0.2770	0.3968	0.4739	0.5138	0.4607	0.4974	0.3548	0.1899
num ≥ 4	0.0267	0.2957	0.4544	0.5893	**0.6283**	0.5881	0.5121	0.3596	0.1810
num ≥ 6	0.0258	0.3066	0.5007	0.6001	**0.6473**	0.5983	0.5093	0.3514	0.1651
num ≥ 8	0.0255	0.3168	0.5098	**0.6063**	**0.6540**	0.6014	0.5062	0.3474	0.1571
num ≥ 10	0.0249	0.3258	0.5114	0.5267	**0.6554**	0.6024	0.5038	0.3425	0.1472

https://doi.org/10.23106/dcmi.952137572

One question remained: What kind of mappings have a highly significant co-occurrence yet are not part of the manual mappings? In an additional analysis step, the co-occurrence data was filtered by rather high thresholds of a ratio larger or equal than 0.6 and a number of clusters larger or equal than 6 and again compared to the manual gold standard. The 49 mapping pairs that were not contained in the manual list were individually assessed using the class descriptions and classification system structure.

Of the 49 mapping pairs, 31 were considered to be correct, 12 partially correct, 1 false and 5 contained RVK classes that are no longer in active use. In this sample, most of the "correct" mappings were for RVK classes for the history of economics of specific countries, which were mapped to the BK classes representing the history of those countries. In the manual mapping, there was only a descriptive note for these classes, but not an exhaustive mapping for each country. This is a clear shortcoming of the manual gold standard, that was not obvious in the beginning of the analysis. Another example is the RVK class "QP 624: product and product range selection" (a subclass of "QP 620 - QP 624: demand management instruments" being mapped to BK class "85.40: marketing" instead of the manual choice of "85.15 research and development (economics)". The manual choice was probably caused by a misunderstanding of the German labels "Produktgestaltung" vs. "Produktentwicklung" (product design and product development). The structural analysis indicates that this topic belongs to the field of marketing, so the automatic mapping can be considered the superior match.

This preliminary first analysis shows that the approach has a high potential to further improve and augment the existing manual mappings as well as create automatic mappings that can be used to improve the retrieval in resource discovery systems or be used as a first draft for manual mapping projects.

5. Discussion and Future Work

The current analysis is limited and will need to be significantly extended to more closely follow the work of the other research groups, especially in regards to the effect of different statistical measures used to select the co-occurrences. Nonetheless, several important goals for the current project have been accomplished: the implementation is fast, very robust and can handle large datasets with ease. The evaluation of the approach against the manual mapping gave decent results for precision and recall, and the in-depth analysis showed that many of the automatic mappings "false positive" pairs were actually correct and can be used to significantly improve the existing mapping.

On a more practical side, work is ongoing to document the data management pipeline and switch it over to a more maintainable and user-friendly solution based on the Knime.org framework as well as implementing the statistical analysis directly on top of the data in the back-end database. Also, the manual mappings are currently only provided on request by the Austrian National Library and are contained in Excel files with a varying layout and degree of mapping granularity. The author intends to convert them into a single, well documented format and work together with the original authors to publish them in an open data repository. The same format can then be used to publish the full automatically generated mappings from RVK to BK, so that libraries interested in enriching their catalogs can easily access and use them.

The chosen approach to simply aggregate all classes from RVK and BK from the entries of a given cluster could also be questioned: In clusters with a large number of entries, some classes will be likely to appear more often than others, and this information is lost in the aggregation process. Future experiments should test if preserving the relative frequency of the found classes can help to improve the final mapping.

It is also planned to include additional open data sets from other libraries as sources. Dutch sources would offer the possibility of more data containing BK annotations, while other international sources could add enough DDC or LCC annotations to generate mappings between these classification systems and the RVK.

https://doi.org/10.23106/dcmi.952137572

References

Aguirre, José Luis, Kai Eckert, Jérôme Euzenat and others (2012). Results of the Ontology Alignment Evaluation Initiative 2012. In Proceedings of the 7th International Workshop on Ontology Matching (OM-2012) collocated with the 11th International Semantic Web Conference (ISWC-2012), Boston, MA, USA, November 11, 2012

Aigner, Sebastian (2015). Das Informationspotential geographischer Metadaten im Kontext von Bibliotheken und webbasierten Diensten : Catalogue enrichment durch Abgleich von Metadaten am Beispiel einer Konkordanz für den Fachbereich Geographie nach Basisklassifikation (BK) und Regensburger Verbundklassifikation (RVK). Universität Wien, Master thesis.

Euzenat, Jérôme and Pavel Shvaiko (2007). Ontology matching. Heidelberg: Springer.

Isaac, Antoine, Lourens van der Meij, Stefan Schlobach, and Shenghui Wang (2007). An empirical study of instance-based ontology matching. In Karl Aberer (Editor), The Semantic Web. 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, (Lecture Notes in Computer Science, 4825). Berlin: Springer. 253–266.

Pfeffer, Magnus (2009). Automatische Vergabe von RVK-Notationen mittels fallbasiertem Schließen. In Ulrich Hohoff und Per Knudsen (Eds.), 97. Deutscher Bibliothekartag in Mannheim 2008 - Wissen bewegen, Bibliotheken in der Informationsgesellschaft. Frankfurt: Klostermann. 245–254.

Pfeffer, Magnus (2013). Using clustering across union catalogues to enrich entries with indexing information. In Myra Spiliopoulou, Lars Schmidt-Thieme, Ruth Janning (Eds.), Data analysis, machine learning and knowledge discovery. (Studies in Classification, Data Analysis, and Knowledge Organization). Berlin: Springer. 437–445

Plößnig, Veronika (2012). Konkordanzen und Kataloganreicherung in Form von Klassifikationen im Österreichischen Bibliothekenverbund (ÖBV) – ein Werkstattbericht. Retrieved September 12th, 2016, from http://epub.uni-regensburg.de/34089/1/plnig%20rvk-anwendertreffen_2012.pdf

Plößnig, Veronika; Christoph Steiner (2014). Klassifikationen: Konkordanzen, Anreicherungsprojekte und RVK - Datenkorrekturen im Österreichischen Bibliothekenverbund. Ein Update. Retrieved September 12th, 2016, from http://epub.uni-regensburg.de/34088/1/ppt%20plnig_steiner-rvk-bk-12-11-2014.pdf

Probstmeyer, Judith (2009). Analyse von maschinell generierten Korrelationen zwischen der Regensburger Verbundklassifikation (RVK) und der Schlagwortnormdatei (SWD). Hochschule der Medien Stuttgart, Bachelor thesis. Retrieved September 12th, 2016, from http://opus.bsz-bw.de/hdms/volltexte/2009/667

Lorenz, Bernd (2008). Handbuch zur Regensburger Verbundklassifikation. Materialien zur Einführung. (Beiträge zum Buch- und Bibliothekswesen, 55). Wiesbaden: Harrassowitz.

Schopman, Balthasar (2009). Instance-Based Ontology Matching by Instance Enrichment. Vrije Universiteit Amsterdam, Master thesis. Retrieved September 12th, 2016, from https://sites.google.com/site/bschopman/master-thesis

Schopman, Balthasar, Shenghui Wang, Antoine Isaac, Stefan Schlobach (2012). Instance-Based Ontology Matching by Instance Enrichment. Journal on Data Semantics, 1(4), 219–236. Retrieved September 12th, 2016, from http://link.springer.com/article/10.1007/s13740-012-0011-z

Schulz, Ursula (1991). Die niederländische Basisklassifikation: eine Alternative für die "Sachgruppen" im Fremddatenangebot der Deutschen Bibliothek. Bibliotheksdienst, 25(8), 1196–1219. Retrieved September 12th, 2016, from http://www2.bui.haw-hamburg.de/pers/ursula.schulz/publikationen/nl_bk.pdf

Waldhör, Andreas (2012). Erstellung einer Konkordanz zwischen Basisklassifikation (BK) und Regensburger Verbundklassifikation (RVK) für den Fachbereich Recht. Universität Wien, Master thesis.

https://doi.org/10.23106/dcmi.952137572