Submitted on: 14.11.2017

2016 Satellite meeting - *Subject Access: Unlimited Opportunities*
11 – 12 August 2016
State Library of Ohio, Columbus, Ohio, USA

'Mixed Methods' Indexing: Building-Up a Multi-Level Infrastructure for Subject Indexing

Andreas Oskar Kempf
Academic Services, ZBW – German National Library of Economics – Leibniz Information Centre for Economics, Hamburg, Germany.
a.kempf@zbw.eu

Tobias Rebholz
Academic Services, ZBW – German National Library of Economics – Leibniz Information Centre for Economics, Kiel, Germany.
t.rebholz@zbw.eu

Abstract:

Increasing numbers of publications and decreasing personnel resources demand new methods besides intellectual subject indexing, to ensure a high quality of content-descriptive metadata even in the future. In addition, the starting situation for subject indexing has changed. A large proportion of publications already contain content-descriptive metadata from different sources.

Against this backdrop, we introduce a multi-level infrastructure for subject indexing which consists of four different indexing tiers. We distinguish between intellectual indexing, semi-automatic indexing approaches, and inter-vocabulary mapping approaches for third party transfer of content-descriptive metadata from controlled as well as from uncontrolled vocabularies. Finally, we switch to retrieval and explain how the power of a controlled vocabulary for indexing can be applied for advanced retrieval scenarios.

Keywords: subject indexing, semi-automatic indexing, inter-vocabulary mapping, query expansion.

1.0 INTRODUCTION

In recent years the indexing situation has changed tremendously. First, digitization, the advent of the Internet, and advances in text- and data-mining have led to new information environments and an unprecedented starting situation for subject indexing. Second, clearly defined collections less and less commonly stand alone, but become part of portals or discovery systems, which are usually characterized by a large heterogeneity of data sources and indexing

vocabularies, ranging from thesauri and classification systems to keywords (Kempf/Neubert 2016). For quite some time authors themselves have been encouraged or even obliged to assign keywords to their publications. This way content-descriptive metadata, usually generated outside of libraries, successively finds its way into bibliographic databases. Third, because of the exponential increase of publications and simultaneous resource constraints it becomes more and more obvious that it is impossible to index all the items of major collections intellectually. New methods of providing subject indexing, which take all these changes into account, seem inevitable.

At the same time, core principles and instruments of subject indexing seem to be largely unchallenged. The enrichment of documents' bibliographic records with subject headings, which represent their content in condensed form, facilitate retrieval, enable quick access and speed up relevance decision enormously. Thesauri and classifications help to control the huge variety and ambiguity of language and to specify the context of knowledge. Only a controlled vocabulary increases the indexing consistency and provides crucial added value for retrieval in cases where search terms are neither part of the title nor of the abstract or the full text. Thanks to lead-in entries for synonyms or quasi-synonyms of preferred terms relevant content can be found even in cases where search terms are not part of the text.

Against this background, we would like to illustrate in our paper how the process of subject indexing is changing more and more into a combination of various partly interwoven indexing levels, in which different indexing methods are applied. The different levels are not to be understood as part of a hierarchy. Referring to the indexing situation at the ZBW – German National Library of Economics – Leibniz Information Centre for Economics, which uses its own controlled vocabulary, the "STW Thesaurus for Economics", we exemplify this development on a conceptual level with regard to intellectual indexing, semi-automatic indexing approaches, and inter-vocabulary mapping for third-party indexing transfer. Finally, with regard to retrieval we would like to illustrate how controlled vocabularies as a core indexing instrument could provide added value for search in cases where there is limited or no content-descriptive metadata at all.

2.0 INSTITUTIONAL BACKGROUND

The ZBW is the world's largest information and research infrastructure for online as well as offline economic literature. As a member of the Leibniz Association and a foundation under public law it is financed jointly by the federal and state governments. The ZBW holds more than 4 million volumes and subscribes to 26,000 periodicals and journals. In its portal EconBiz (http://econbiz.de) it enables access to more than 10 million titles in economics. In addition, the ZBW provides one of the fastest-growing collections of Open Access documents on the Internet, EconStor (http://econstor.eu). The digital publication server currently gives free access to nearly 120,000 articles and working papers. As a high-tech information infrastructure the ZBW conducts research in computer science and related areas, particularly in Text and Data Mining, Semantic Web, and Science 2.0. In addition, it develops innovative technologies for operating the IT infrastructure for its own, world's largest, specialist library for economics.

For indexing the ZBW uses its own thesaurus, the STW. Having been originally developed in the late 1990s in a three-year project, which was partly funded by the German Federal Ministry of Economics, it is a controlled, structured vocabulary, which serves the needs of information and documentation in the field of economics and business studies. As a bilingual thesaurus it contains about 6,000 descriptors in German and in English, and nearly 20,000 non-descriptors

as lead-in entries, covering all economics-related subject areas, and, on a broader level, the most important related subject areas. Furthermore, it comprises an elaborate systematic structure in form of domain-specific subject categories on up to four different hierarchical levels. As a navigation tree on the web page and in addition to alphabetical search it allows STW users to browse descriptors thematically.

The responsibility for the continuous maintenance of the STW lies with the ZBW. It has undergone a complete overhaul in 2010-2015 and is regularly updated and further developed in compliance with the latest international standard on thesauri (ISO 2013) by a regular editorial team, which consists of a group of domain experts in economics and information science from the ZBW. Following the latest international terminology usage in the field of economic sciences and taking into account the internal structural and formal principles of the thesaurus, it verifies and decides on changes and updates concerning vocabulary and thesaurus structure. Besides the current professional discussion, key sources for changes to be made are in-house proposals by subject indexers and proposals from external institutions which reuse the STW.

The STW has proven to be an instrument that has continually adapted to new requirements arising from its integration in permanently changing information environments (Kempf/Neubert 2016). In 2009, the STW was published on the web (http://zbw.eu/stw) and was made available as Linked Open Data (Borst/Neubert 2009). With the aim of facilitating broader reuse and supported by a liberal license, browsing and download of the complete thesaurus and all mapping files were made freely available. Developed as a general indexing and retrieval tool for publications and research data in the economic sciences, the STW can be used by different kinds of organizations, such as documentation centers and database producers, for indexing their own materials. This implies an increased responsibility for transparency and sustainability in the release of new versions of the vocabulary. In total, more than 1,000 individuals and institutions from 80 countries have downloaded the STW between 2009 and 2015. To find out how the STW is reused exactly, the establishment of a user community is intended.

3.0 'MIXED METHODS' INDEXING IN A MULTI-LEVEL INDEXING INFRASTRUCTURE

The starting situation for subject indexing at the ZBW is complex and varies depending on the type of the documents and their form of publication. Journal articles, articles in collective works, and working papers are often already indexed with classification codes and author keywords, electronically available within the cataloguing system. More than 60 percent of the journals indexed on article level contain keywords assigned by authors or journals. Since the year 2013, existing keywords have been included in the bibliographic records during the descriptive cataloguing process. These records are not indexed intellectually any more.

Monographs are often already indexed with subject headings taken from controlled vocabularies. Mostly they come from the Integrated Authority File (GND) of the German National Library (DNB), or from the Library of Congress Subject Headings (LCSH), which can be reused for third-party metadata transfer.

To achieve a more homogeneous high-quality specialized indexing in the future, regardless of the amount of publications, various indexing methods which are already applied or under active development can be distinguished. Apart from intellectual indexing using a controlled

vocabulary (3.1), these approaches include three complementary indexing methods (3.2), namely semi-automatic indexing approaches (3.2.1) and approaches of inter-vocabulary mapping, including mappings to other controlled vocabularies (3.2.2) and uncontrolled vocabulary (3.2.3).

3.1 INTELLECTUAL SUBJECT INDEXING

On the first level, the method of conventional intellectual subject indexing is applied. According to ZBW's indexing rules this means that the main content should be represented in full, and that the most precise subject headings should be assigned. Indexing refers to the content of a text – not necessarily to the words written in the text – and in cases of spatial reference geographical subject headings should be assigned.

The aim of intellectual indexing is twofold. The primary goal of intellectual indexing is to make the relevant economics literature, which lacks content-descriptive metadata, discoverable. Therefore, intellectual indexing is limited to a subset of incoming documents. Taking existing search tools and search engine technologies, which could lead to quite satisfactory search results, into account, intellectual indexing is limited to documents for which subject metadata is not yet available.

The second objective is to use intellectual indexing as the essential basis for further development of the thesaurus. Only a considerable amount of intellectual subject indexing guarantees that the indexers stay in touch with the scientific field for further development and constant re-adjustment of a thesaurus according to the professional discussion and the latest developments within a discipline. This usually requires a team of domain experts and information scientists working closely together. It is the intellectual indexing by information professionals which builds an indispensable feedback loop about the relevance and discriminatory power of thesaurus concepts in the field (Kempf/Neubert 2016).

3.2 COMPLEMENTARY INDEXING METHODS

The following three levels apply indexing methods complementary to intellectual subject indexing. Partly they include machine-processed approaches. This is the case for semi-automatic indexing using machine learning algorithms on the basis of text- and data-mining and (semi-)automatic approaches of inter-vocabulary mapping.

3.2.1 SEMI-AUTOMATIC SUBJECT INDEXING

On the second level, a semi-automatic indexing approach is under active development. The aim is to generate automatically a sufficiently high-quality indexing using the STW for electronically available information resources regardless of the amount of incoming publications. Strategically, providing a standardized description of the content of digital information resources is a crucial success factor for the ZBW in competition with other providers of economic information. To achieve this goal and to maximize the exchange of knowledge, there is a close in-house cooperation between the department responsible for subject indexing and the research department in media computer science at the ZBW. Computer and information scientists collaborate closely with domain experts to develop an indexing assistant which automatically suggests descriptors from the STW and which could be used in production for online information resources.

An integral part in the development process of an automatic suggest service is to establish a quality management system that takes both the quality of the automatic indexing as well as the quality of the intellectual indexing into account. As a first step, the quality of the in-house intellectual subject indexing was determined with regard to indexing consistency. As a second step, based on text- and data-mining approaches, machine learning algorithms have been developed, which try to model human indexing behavior after being trained on a dataset of STW-indexed open access publications. These algorithms are evaluated by domain experts after being applied on a test dataset of electronically available full texts which were analyzed and automatically annotated by using STW descriptors. So far, research is ongoing on the Maui-indexer (github.com/zelandiya/maui), on the basis of a novel combination of graph-based concept activation methods with the k-Nearest Neighbors algorithm as a concept selection method (Große-Bölting et al. 2015), and on an two-tier multi-labeling classification based on logistic regression and decision trees, using a training dataset of STW-indexed open access publications.

3.2.2 INTER-VOCABULARY MAPPING TO CONTROLLED VOCABULARIES

The third level applies methods of inter-vocabulary mapping to other controlled vocabularies to achieve a truly collaboratively organized subject indexing beyond library boundaries. In the light of new information environments as depicted above, there is an increased demand for thesauri to be interoperable with other controlled vocabularies (ISO 2013) in order to exploit subject information taken from other thesauri to a much greater extent than ever before. This allows an integrated search across various databases indexed with different controlled vocabularies.

In the past, these mappings, so-called cross-concordances, were established in numerous ways. At the beginning, these mappings were built up exclusively intellectually. Bilateral intellectual mappings were established between the STW and the Thesaurus for the Social Sciences (TSS) and the German Integrated Authority File (GND) published by the DNB. If a title in EconBiz has descriptors from the GND, the EconBiz search index is expanded by descriptors from the STW, for which equivalence relations in the mapping exist, and by their English and German synonyms. The index is expanded further by equivalent descriptors from other vocabularies and their synonyms. In addition, an automatic matching procedure between bibliographic records from the Union Catalogue, indexed with the GND, and ZBW titles is intended. For matching records GND subject headings are translated into STW descriptors. It is for these practical reuse scenarios that inter-vocabulary mapping marks an important step towards truly collaboratively organized subject indexing between specialized libraries on the one hand and the national library on the other hand. To underline this, the ZBW and the DNB have entered into an agreement to maintain and further develop the cross-concordances cooperatively.

Meanwhile, various automatic and semi-automatic mapping procedures on the basis of the web-based formats SKOS and OWL have been built up. A completely automatic mapping has been created to DBpedia, a LOD representation of Wikipedia content. STW-descriptors and non-descriptors in German and English have been looked up in DBpedia. A semi-automatic inter-vocabulary mapping approach has been used for a mapping to the AGROVOC thesaurus of the Food and Agriculture Organization of the United Nations (FAO) and for the updates of the mappings to the TSS. For the former, mapping results of an automatic string matching process have been verified intellectually by a domain expert. The latter has been annually updated as part of the library track of the Ontology Alignment Evaluation Initiative (OAEI) (Kempf et al. 2014).

Most recent mapping initiatives include the use of an interactive alignment tool based on the SKOS format. Using the web-based Amsterdam Alignment GenerAtion MEtatool (AMALGAME), the mapping procedure has turned into an iterative mapping process. After being applied for a mapping to the Statistical Data and Metadata eXchange (SDMX), an international initiative that aims at standardizing and modernizing the mechanisms and processes for the exchange of statistical data and metadata among international organizations and their member countries, it has also been tested to a mapping to the Journal of Economics Literature Classification system (JEL), which is the internationally established standard classification for scholarly literature in the economic sciences.

In a new way, this mapping combines inter-vocabulary mapping to another controlled vocabulary with user-generated subject indexing. In economics researchers are already accustomed to assign JEL classes to their publications, for example during the self-upload of their publications on ZBW´s open access repository EconStor. Based on these JEL classes and their mapping to STW subject categories, an application can suggest a selection of STW descriptors to the researchers. In this way, authors themselves can be encouraged to provide a more fine-grained content description of their publications by using the STW as a standardized controlled vocabulary (Kempf et al. 2015).

Finally, new candidates for further development of the thesaurus can be identified during the mapping process.

3.2.3 INTER-VOCABULARY MAPPING TO CONTROLLED VOCABULARIES

On the fourth level, methods of semi-automatic mapping to uncontrolled vocabulary are applied. Uncontrolled vocabulary usually stems from free keywords assigned by authors themselves.

Automatic mapping to uncontrolled vocabulary can be done by converting author keywords into STW descriptors through string match. Comparing keywords with descriptors from the STW based on similarity metrics enable mappings between the STW and author keywords. Apart from exact matches, close matches could be used for vocabulary enrichment. However, this approach might be limited by the ambiguity of natural language as encountered in author keywords.

Author keywords which could not be converted into STW descriptors are a source of potential new candidates for STW descriptors or non-descriptors. These candidates may be additionally compared with log files taken from search queries to get an indication of whether a term is relevant in the professional community.

Author keywords which do not explicitly find their way into the thesaurus may instead be stored in the background and assigned to descriptors as hidden synonyms specifically used for automatic indexing (for example spelling variants or frequent typos). This could lead to an increased homogeneity of content-descriptive metadata in the database. First results of a mapping between author keywords and the STW are already available for a limited document corpus (Bahls/Rebholz 2015).

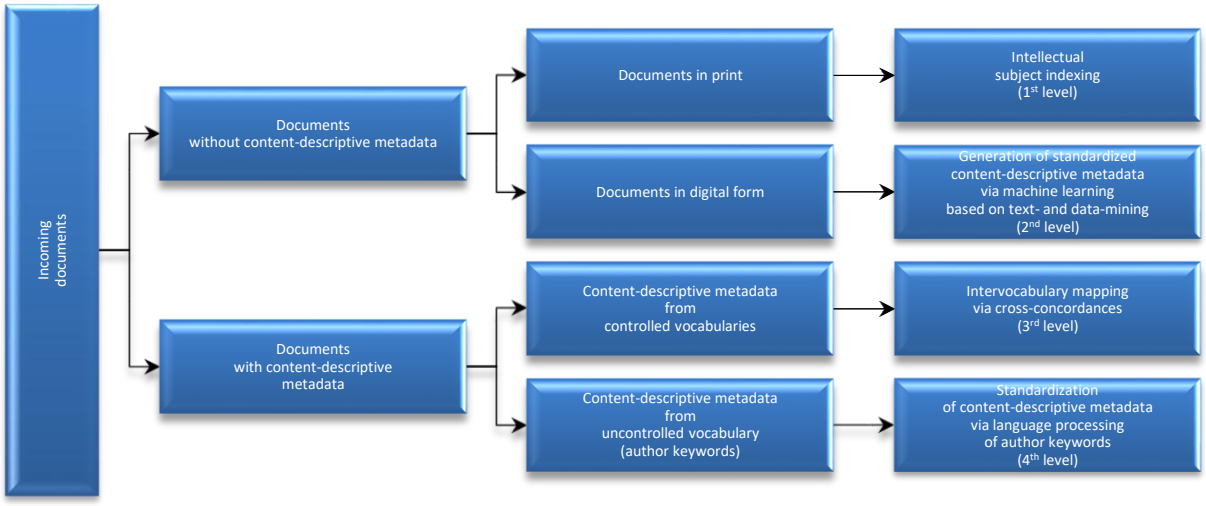

Figure 1: Conceptual model of a multi-level infrastructure for subject indexing

4.0 NEW BROWSING AND ADVANCED RETRIEVAL SCENARIOS

Apart from explicitly generating index terms through the methods described above, the power of a thesaurus can be applied dynamically during retrieval. That includes the broadening of search results and the support of associative browsing.

Titles, abstracts and full texts of books or articles may reference concepts which are not caught in the descriptive metadata. The ZBW has developed terminological web services for economics which return all synonym terms, optionally extended by mappings to other vocabularies, for the construction of retrieval queries (Neubert 2012). An application which is based on these web services can include all available alternative search terms to expand a query before submitting it to a backend.

This mechanism for expanding queries overlaps with the automated indexing approaches described earlier. The advantage of the latter lies in the applicability of much more advanced and resource-intensive algorithms while generating the index terms. If supplemented by intellectual control, falsely generated stray terms can be eliminated. However, every time that the vocabulary is further developed and updated a re-indexing of the whole content might be too expensive. In contrast, dynamically applied query expansion can be adapted easily to new versions of a vocabulary. Furthermore, it can be applied to retrieval systems where the indexing process is completely beyond the control of the searching party.

Another function of the developed web services provides all concepts referenced in a query string, and narrower and related concepts to these. ZBW's digital repository EconStor[1] (see

[1] At the moment, EconStor is in a relaunch process which has not yet been completed. This is how the web service looked like in the previous version of EconStor.

Figure 2) uses a combination of both functionalities and presents alternative search proposals to the user.

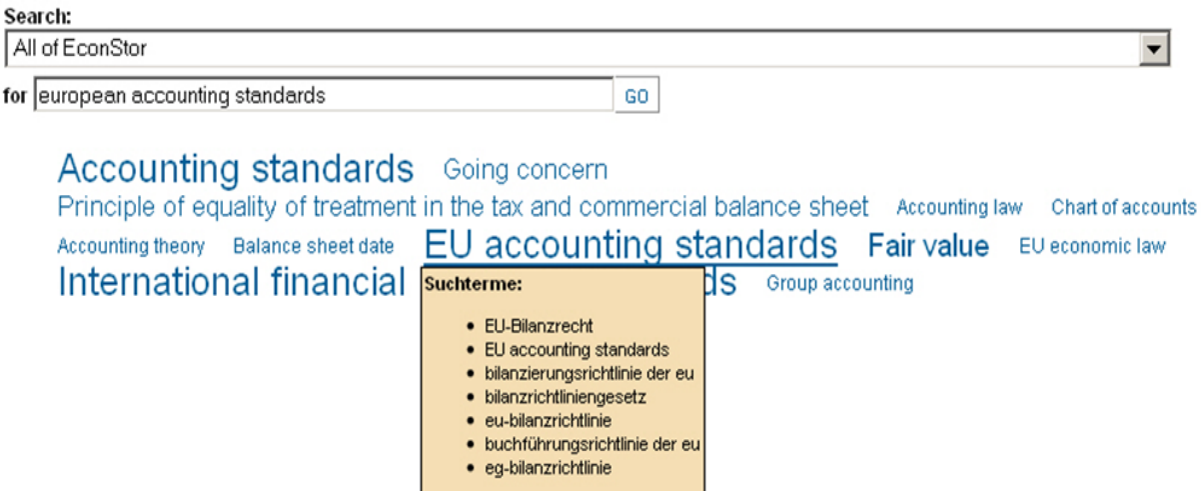

Figure 2: Web service in EconStor, which provides all concepts referenced in a query string.

This allows for associative browsing to possibly also relevant concepts, which never would have been included in any form of intellectual or automatic indexing, because their relevance solely results from the search intent of the current user. Anyway – when the user clicks on one of the alternate search proposals, a new query can be sent to the database, expanded by all synonyms of that concept.

5.0 CONCLUSION

As we have seen by referring to the indexing situation at the ZBW, the process of subject indexing and the use of controlled indexing vocabulary for retrieval are changing. For a long time, subject indexing was synonymous with intellectual indexing. Today, due to content-descriptive metadata already available and advances in text and data mining, the process of subject indexing spreads out into multiple levels with different methods of subject indexing applied. Apart from traditional intellectual indexing, they include inter-vocabulary mapping with other controlled as well as uncontrolled vocabulary for third-party indexing transfer and semi-automatic indexing approaches.

It is in particular with regard to vocabulary development that the different indexing levels intertwine. On the one hand, all the complementary indexing levels depend on a high-quality controlled vocabulary which is up to date with regard to the professional discussion and the latest changes in the terminology of a discipline. The first level of intellectual subject indexing is the backbone for further development of a controlled vocabulary. Contrary to the expected decline in intellectual indexing, its importance for the development of a high-quality controlled vocabulary and all emerging indexing services is likely to increase. While the quantity of intellectual indexing may decline in the future, its importance as a core source of high-quality metadata forming the basis for all further indexing services and applications is likely to increase. A considerable number of subject indexers in close contact with the scientific field is required to ensure that the further development and constant re-adjustment of an indexing language properly reflects the latest developments within the discipline (Kempf/Neubert 2016).

It is in particular the indexing assistant on level 2 which includes machine learning approaches, that is extremely dependent on a high-quality vocabulary and a core high-quality training dataset.

On the other hand, vocabulary development and its automatic processing could profit from an extension of proposals for new vocabulary candidates. For one thing, the indexing levels 3 and 4 could lead to suggestions for new candidate vocabulary and entry terms. For another thing, descriptors from other controlled vocabularies (3rd level) as well as keywords from uncontrolled vocabulary (4th level) could be stored in the background to support automatic indexing.

Finally, controlled vocabulary can be applied for query expansion during retrieval. Terminological web services could expand search queries by synonym terms, optionally extended by mappings to other vocabularies, as well as by all concepts referenced in a query string, and concepts related to these. Dynamically applied query expansion could be adapted easily to new versions of a vocabulary and allows for associative browsing.

As we have seen, an effective interplay between the different intellectual as well as automatic indexing methods on the one hand and retrieval-oriented approaches on the other hand is crucial for the future.

Acknowledgments

We would like to thank Joachim Neubert, who contributed crucially with his profound expertise to the description of new browsing and advanced retrieval scenarios.

References

Bahls, Daniel; Rebholz, Tobias. 2015. “Evidenzbasierte Begriffs- und Synonymerweiterung des STW. ” In *Proceedings of the 104th Bibliothekartag*, Nuremberg, Germany https://opus4.kobv.de/opus4-bib-info/files/2498/Praesentation_Bibliothekartag_2015_Rebholz_Bahls.pdf

Borst, Timo; Neubert, Joachim. 2009. “Case Study: Publishing STW Thesaurus for Economics as Linked Open Data.” *W3C Semantic Web Use Cases and Case Studies*, 2009, https://www.w3.org/2001/sw/sweo/public/UseCases/ZBW/

Große-Bölting, Gregor; Nishioka, Chifumi; Scherp, Ansgar. 2015. “A Comparison of Different Strategies for Automated Semantic Document Annotation.” In *Proceedings of the 8th International Conference on Knowledge Capture*, Palisades, NY, USA, 2015, http://dx.doi.org/10.1145/2815833.2815838

International Organization for Standardization (ed.). 2013. *Information and documentation - Thesauri and interoperability with other vocabularies*. Part 2: Interoperability with other vocabularies. (ISO 25964-2).

Kempf, Andreas Oskar; Neubert, Joachim. 2016. “The Role of Thesauri in an Open Web. A Case Study of the STW Thesaurus for Economics.” In: *Knowledge Organization* 43, 3, 160-173.

Kempf, Andreas Oskar; Neubert, Joachim; Faden, Manfred. 2015. “The Missing Link – A Vocabulary Mapping Effort in Economics.” *14th European Networked Knowledge Organization Systems (NKOS) Workshop*, Poznan, 2015, https://at-web1.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2015/content/NKOS2015-presentation-kempf.pdf

Kempf, Andreas Oskar; Ritze, Dominique; Eckert, Kai; Zapilko, Benjamin. 2014. “New ways of mapping knowledge organization systems: using a semi-automatic matching procedure for building up vocabulary crosswalks.” In: *Knowledge Organization* 41 (1): 66-75.

Neubert, Joachim. 2012. “Linked data based library web services for economics.” In: *Proceedings of the 2012 International Conference on Dublin Core and Metadata Applications*, 12-22.