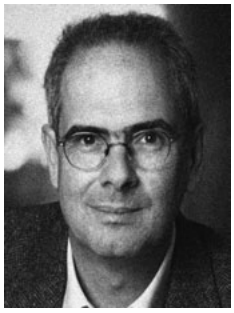
The Role of Thesauri in an Open Web: A Case Study of the STW Thesaurus for Economics

Andreas Oskar Kempf* and Joachim Neubert**

ZBW-Leibniz Information Centre for Economics,
Neuer Jungfernstieg 21, Hamburg, Germany, 20354
*<a.kempf@zbw.eu>, **< j.neubert@zbw.eu>

Andreas Oskar Kempf is a knowledge organization specialist at the ZBW-Leibniz Information Centre for Economics/German National Library of Economics. He holds a Ph.D. in sociology from Goethe University, Frankfurt am Main, and received a master's degree in library and information science from Humboldt-University, Berlin, Germany. His research and publications have largely focused on descriptive metadata and controlled vocabularies for the social and economic sciences. His research interests include the interoperability of knowledge organization systems, semi-automatic indexing, and knowledge representation of scientific literature and research data.



Joachim Neubert is a scientific software developer at the ZBW. He holds degrees in history and information science and has worked in information technology at a major press archives. At ZBW, he published several datasets as linked open data. In 2009, he started the SWIB-Semantic Web for Libraries conference and serves to date as co-chair of its program committee. As an invited expert, he took an active part in the Library Linked Data Incubator Group of the World Wide Web Consortium (W3C). His research interests include knowledge organization systems and authorities, linked data, and web-based information systems and applications.

Kempf, Andreas Oskar and Joachim Neubert. 2016. "The Role of Thesauri in an Open Web: A Case Study of the STW Thesaurus for Economics." *Knowledge Organization* 43 no. 3: 160-173. 21 references.

Abstract: This paper illustrates the changing role of thesauri interlinked with overall changes of modern information infrastructure services, referring to "STW Thesaurus for Economics" as a case study. It starts with an overview of the history and development of the STW and describes the far-reaching changes brought about by its publication on the Web, with regard to subject indexing, retrieval and new uses for Linked Open Data. It argues that only the most recent technological developments help thesauri to exploit their full potential which is why they more than ever have a place in current information retrieval and infrastructure.

Received 22 January 2016; Revised 25 February 2016; Accepted 5 February 2016

Keywords: STW Thesaurus for Economics, subject descriptors, indexing, data, search, thesauri, information

1.0 Introduction

The high costs of developing and maintaining thesauri give rise to the question of why institutions still maintain such a traditional instrument for subject indexing in times when other, much simpler and more cost-effective approaches for content access to library catalogs and databases exist. Have thesauri in general not lost their justification in the era of full-text search?

In this paper, we would like to point out that because of their strong adaptiveness and integrability into rapidly shifting information environments, thesauri still have a place, and indeed various places, in modern information retrieval and infrastructure. Here we refer to the STW Thesaurus for Economics[1] published by the ZBW-Leibniz Information Centre for Economics/German National Library of Economics. In the nearly two decades of its existence, it has not only served the traditional functions of a thesaurus, i.e. to control the huge variety (synonymy) and ambiguity (polysemy) of language to make relevant information retrievable, but first and foremost it has proven to be an instrument that continually adapts to new requirements arising from its integration in constantly changing information environments, thus steadily enlarging its field of application and the benefits it offers. We would like to prove our argument by presenting a case study of how the STW has evolved and of the diverse contexts of its use, which include more traditional and particularly innovative applications.

In section 2 we recap the history and evolution of the STW. Section 3 describes the changes that accompanied publication of the STW on the web with regard to format, perception, maintenance and interlinking of it with other data on the web. The next three sections describe its application in three different areas: subject indexing, retrieval and linked open data respectively. In each case, the section outlines the basics before describing more advanced capabilities enabled by the STW, some of them still under development. We conclude by summarizing the lessons that can be drawn from the integration of thesauri within the variety of new applications presented.

2.0 The STW Thesaurus for Economics

The STW may serve in many respects as an example par excellence to illustrate how thesauri have typically advanced in the past and the changing roles of thesauri in the more recent networked information environment and the web of today. The main precursor of the STW dates back to the 1970s, when comparison with the situation in other countries convinced West German authorities at the federal level of the importance of information and documentation and triggered the launch of several large-scale funding programs to establish specialized information centers all over the country. Moving from the paper-based era to the age of the Internet, the further trajectory of the STW was marked by several transitions regarding its content as well as its format. This is why for some time now its reuse has gone beyond exclusive application in vast siloed collections towards integration into the web at large, spanning classical as well as innovative forms of application made possible largely due to technological developments of recent years.

The STW is a controlled, structured vocabulary serving the needs of information and documentation in the field of economics and business economics. It has existed under its present name and in its current form since 1998 and throughout this time has benefited from development in compliance with the latest, first German and later international, standard on thesauri. Ranging from concrete business practices to the global economy and the history of economic thought, it covers all economics-related topics, and, on a broader level, the most important related subject areas such as history and social sciences. Developed as a general indexing and retrieval tool for publications 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.

The STW was originally developed in a three-year cooperation between four leading public and private German providers of business and economics information services.[2] Partly funded by the German Federal Ministry of Economics, it aimed at a common standardized documentation language in the economic sciences. At the time, each of these institutions used their own indexing language ranging from a simple keyword list to an individually created cataloging system and a fully-fledged thesaurus, the so-called "Thesaurus Wirtschaft" "Thesaurus Wirtschaft" (Informationszentrum des HWWA 1992).All of these indexing languages reflected the specific collection focus of each of these participating institutions, which was on either empirically or theoretically based literature, on economics or business economics.

The advent of modern information technology in the mid 1990s made it possible to bring together the content of the different library catalogs and databases of the four information providers and to publish it on CD-ROM. However, users were disappointed to find that an integrated content search within this merged dataset was not supported. Each of the participating institutions still used its own indexing vocabulary and followed its own indexing rules. In order to ensure a greater demand for this jointly created information product, the development of a common indexing language to provide uniform content access to the published bibliographic references was seen as indispensable. This need for a common standardized indexing language marks the birth of the STW.

At the beginning, the standardization of content description in business and economics information was mainly addressed to a German-speaking scientific community. Therefore, the first version of the STW, in 1998, was available only in German. It was also with this essentially German-speaking user group in mind that the release of the STW claimed to serve as a standard, as reflected in the name of the thesaurus. However, similar to the situation in other disciplines, scientific discourses in many subdisciplines of the economic sciences became more and more international. The ZBW, which has taken over STW development, consequently directed its information services towards an international, English-speaking scientific community. As of version 8.02, released in 2007, all descriptors of the STW are bilingual, German and English.

Today, the STW is regularly updated and further developed by a group of ZBW's domain experts in economics and information science, following the latest international terminology usage in the field of economic sciences and the internal structural and formal principles of the thesaurus. A team of scientific software developers and engineers work continuously to integrate the thesaurus into existing and permanently changing information processes and to implement new applications for the thesaurus.

Starting with version 8.06 in 2010, the STW has been overhauled completely through version 9.0 in 2015. Subject area by subject area was reviewed and adapted to new developments in international economy and business

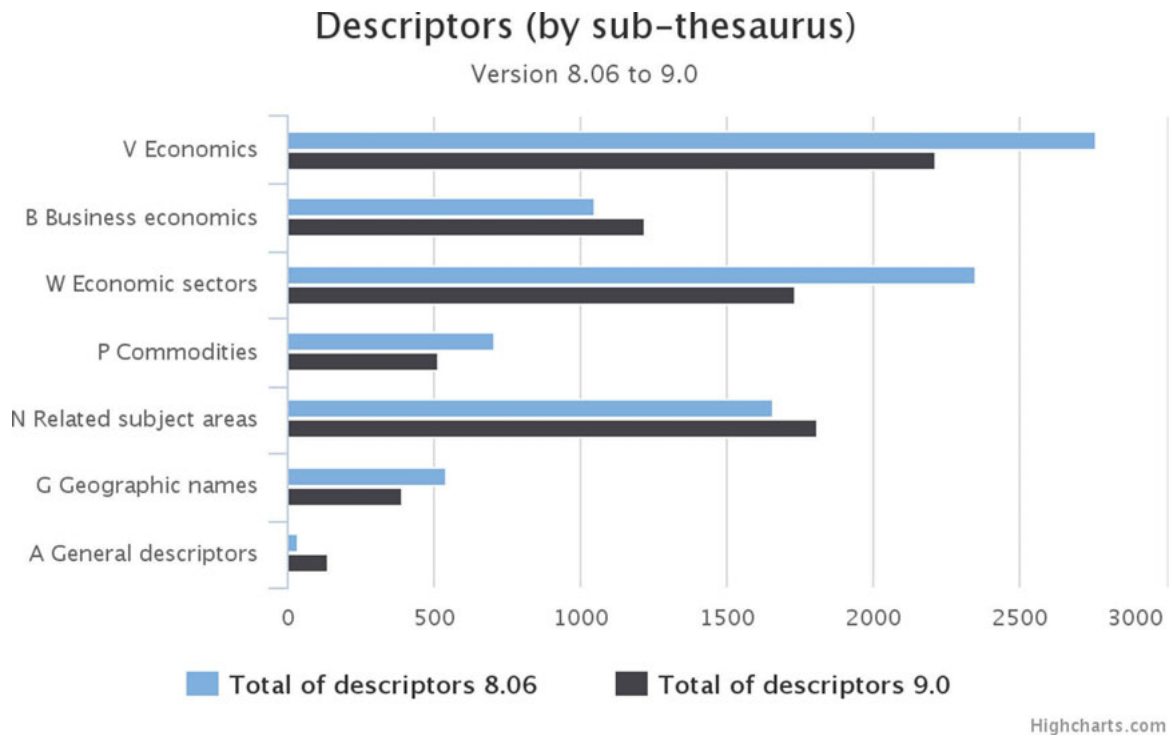


Figure 1. Distribution of the STW descriptors by sub-thesaurus.

economics. More than 750 descriptors have been added. The deprecation of more than 1,000 descriptors was part of a well-considered re-balancing and re-focusing, accompanied by substantial changes in the subject category system of the STW. In the process, the international orientation of the thesaurus was strengthened, and many English terms were added. Figure 1 shows the broad domains of the STW and gives a high level overview of the changes, which are described in detail at http://zbw.eu/stw.relaunch.

3.0 "The Web Changes Everything"

Libraries have a long tradition of data sharing and cooperative work. Computers and computer networks have long been the environment of this work—yet, it was an environment shared with other libraries, while the emerging World Wide Web was seen as secondary. The STW has, from its beginning, been built and maintained electronically, using a customized thesaurus management system. However, for the first ten years it was published only in print. Data files in custom formats were imported into the library union catalog and were made available via bilateral cooperation, usually requiring a written agreement.

In 2009, the STW was published on the web (http://zbw.eu/stw). Browsing and download of the complete thesaurus and all mapping files were made freely available with the aim of facilitating broader reuse and supported by a liberal license (CC-by-nc, since 2014 ODbL). As facilitating easy reuse was the main objective, the choice fell on the SKOS (Simple Knowledge Organization System) format recommended by the W3C and covering all requirements for easy and complete data publication and exchange (Neubert 2009). For each concept, an HTTP web address (URI) as immutable identifier was published. When looked up, this address provided information about the concept and links to others. The data was embedded in the HTML web pages themselves via RDFa (Resource Description Framework in attributes), which made them readable for humans as well as for machines, implementing the principles of linked open data (LOD) as defined by Tim Berners-Lee (2006).

From the beginning, the STW concepts were linked to other data on the web. That started with DBpedia, an LOD representation of Wikipedia content. Notably, the link target was not authority data in any library-specific format, but a well-known, publicly available web source, which was already the most important hub in the emerging LOD cloud. Other parties, in turn, have later on linked to the persistent addresses of STW concepts, thus making, in turn, their concepts part of a globally interlinked web of data.

With publication on the web, the perception of the STW within the institution changed from just an internal working tool for indexers to an active asset of the ZBW which is recognized world-wide (in total, more than 1,000 individuals and institutions from 80 countries have downloaded the STW between 2009 and 2015), which con-

tributes to its reputation as the world's largest library in the economic sciences. As a significant investment into the future of the vocabulary, it was subjected to the complete overhaul described above—which was unprecedented for STW, and presumably for most other thesauri too.

STW is not alone in this type of development. Large investments have also taken place in the GACS (Global Agricultural Concept Scheme) thesaurus project, and which aims to unify the three major agricultural thesauri: the Agrovoc (Food and Agriculture Organization), the thesaurus of the U.S. National Agricultural Libary and the CAB thesaurus of the Centre for Agriculture and Biosciences International (Baker and Suominen 2015). The Getty institute streamlined its highly structured vocabularies for linked data use, made it available under an open license (Baca and Gill 2015), and are currently developing a completely new "Cultural Objects Name Authority" structured vocabulary targeting the same goals. The Finto national ontology project resulted in the interlinking of a number of general or specialized thesauri with an upper-level ontology (Suominen et al. 2014), undertaken by the Finnish National Library. Other institutions, such as the European Union's EuroVoc, the International Virtual Observatory Alliance or the U.S. National Aeronatuics and Space Administration have likewise invested in the publication of their vocabularies on the web and as SKOS downloadable files.

For STW, the widespread acceptance of the download facilities offered has had important consequences for the maintenance of the vocabulary. Prior to web publication, all STW users were well known. Now—due to its open availability—the STW may be in use in places about which the ZBW as its publisher has no knowledge at all. To keep users informed about new versions of STW or other changes, and to get an impression of who is using STW regularly, we established an "stw-announce" mailing list for new version announcements and other changes. Since subscription to the list is optional, we can however not be sure that the approximately 140 subscribers cover all STW users. Another mailing list is intended to foster exchange between "stw-users." These steps aim at building an STW community. Being open to the public use has already led to a higher level of participation in thesaurus development. Institutions, which use the STW for indexing or as a reference tool, for example, suggest new descriptors or alternative translations.

Serving a partly unknown user community brings a new level of responsibility for the vocabulary publisher. To give an example, when obsolete descriptors were removed in the past, index entries in the respective publications could be reassigned to other descriptors in continued use in the affected library systems, and the descriptors could be safely deleted. Now, with unknown users trusting in the persistence of the published URIs, such concepts, flagged as "deprecated" and often accompanied by a "replaced-by" link, exist alongside as the thesaurus itself. Right from the beginning, the STW was clearly versioned. All published versions are still available on the web. With the latest versions, the ZBW started to provide extended and detailed change reports which made changes of the STW traceable for known and unknown re-users, and applications working on it, as well as for ZBW's own indexers (Neubert 2015).

4.0 Subject indexing

One of the core application fields of thesauri still valid today is subject indexing—for various reasons. Due to the fact that each descriptor, in principle, can be combined with any other descriptor, thesauri allow for a polydimensional description of a given document. With a rather restricted amount of descriptors they can express an unlimited number of different subjects. For this reason, thesauri are regarded as the most economical, flexible and expressive of the conventional documentation languages (Bertram 2005).

Additionally, given the huge variety and ambiguity of the economic terminology, which is comparable to other disciplines of the social sciences, the need for a standardized documentation language to achieve a high level of consistency in the description of information resources in economics is undeniable. The field of economic sciences is characterized by a juxtaposition and sometimes opposition of different schools and theoretical models. One example of this are the divergent and still persisting assumptions between "Western" and "Socialist" economics (Gastmeyer 2000). Furthermore, huge differences may occur in the meaning of one and the same term between the different sub-disciplines within the economic sciences, as for example between economics and business economics, illustrating the need to clarify the context in which a term is used. In a thesaurus, this could be done on a linguistic level by using a disambiguated preferred label or by the assignment of a term to different subject categories. Considering the subject of "investment theory," the STW descriptor "Corporate investment theory" has been coined and assigned to the subject category "Capital budgeting" within the field of business economics, while the descriptor "Theory of aggregate investment" was assigned to the subject category "Capital and investment" as a branch of economics. Moreover, the terminology of business practice in itself is a highly ambiguous field of economics. Vast interdisciplinary differences also exist in the use of a term in the social sciences, of which economic sciences only represent one. A prime example is the term "society," which on the one hand could be used in a sociological context and on the

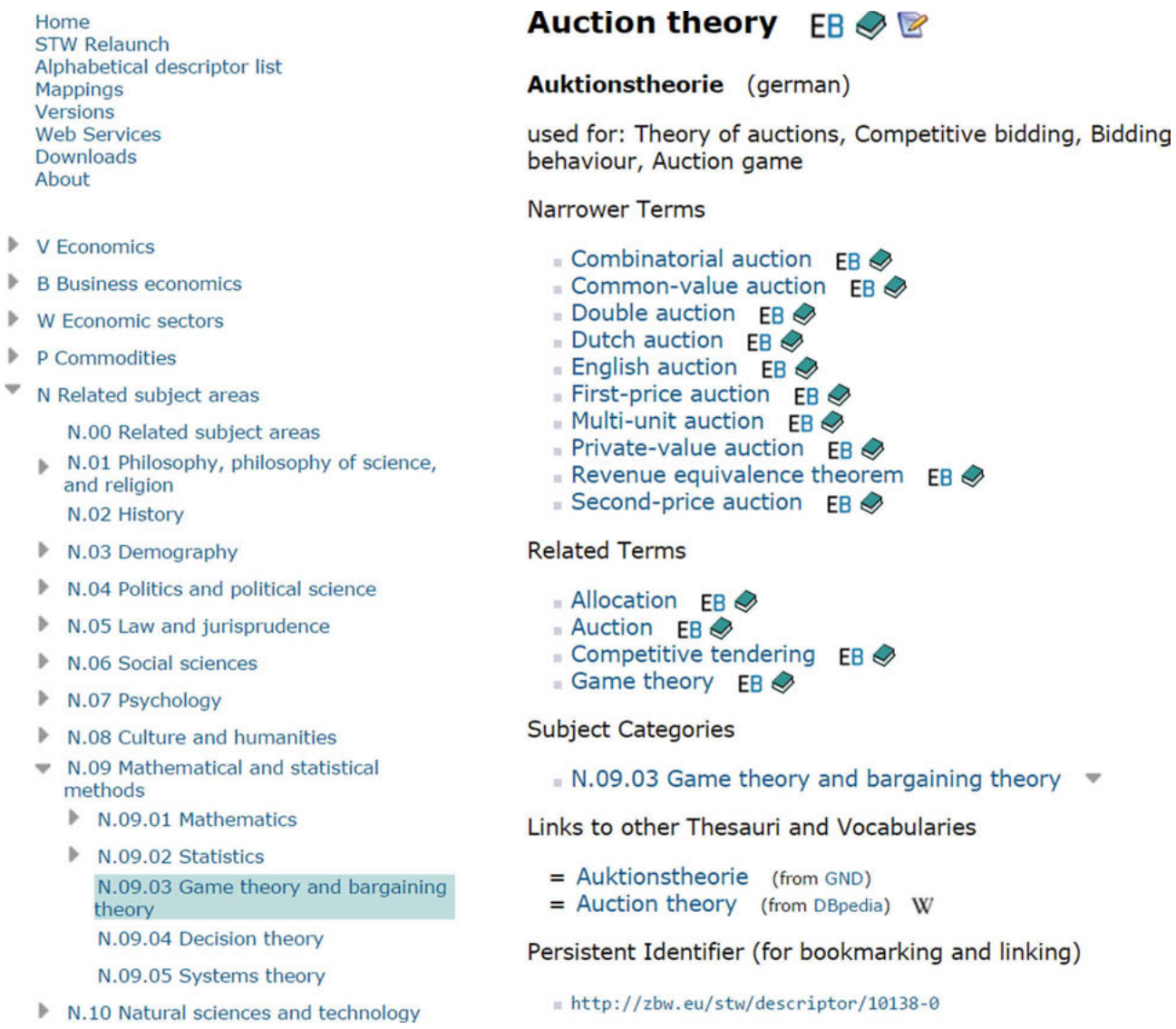


Figure 2. The STW descriptor page is linked extensively to support associative browsing. The subject categories navigation tree on the left allows for systematic access. The "EB" icon triggers a search for publications in EconBiz with the corresponding descriptor. The "W" icon close to the DBpedia concept links to the relevant Wikipedia page.

other hand to mean a type of organization. In the STW all terms are selected based on domain-specific and cross-institutional criteria as well as the expected level of occurrence in search queries (Deutsches Institut für Normung 1463, part 1). Thus, particular emphasis is put on the user-friendliness of the selected descriptors.

Apart from that, the STW 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 (see Figure 2), it allows STW users to browse the descriptors of a certain subject field thematically. While terminological control and semantic relations of a concept indicate the narrower content-specific relationships of a concept, the subject categories point to the larger domain-specific context of it. The first level consists of seven main subject groups or sub-thesauri. They are divided according to the sub-disciplines and sub-areas in the economic sciences. In addition to the usual continental European subdivision of the economic sciences between economics and business economics, the STW contains a sub-thesaurus of economic sectors and one of commodities. With regard to their subdivisions, the latter ones follow current classifications of products and economic sectors used in official statistics. Subject categories within universal authority files usually do not meet these highly domain-specific requirements of specialized information demands. As well as having the sub-thesauri "General descriptors and Geographic names," the STW is supplemented by a sub-thesaurus for "Related subject areas." The selection of descriptors from these subject areas

reflects the economic sciences' perspective, with, e.g., a particular focus on statistical and mathematical methods used across its sub-fields.

In recent years digitization, increasing numbers of publications, and decreasing personnel resources have led to a new situation in subject indexing. Complementary to the traditional two-step indexing process which includes the intellectual understanding of a document and the translation of its content into a documentation language, various new indexing approaches have emerged which all involve thesauri in one way or another. With regard to the STW, the following alternative, partly machine-processed approaches are already applied or under active development.

4.1 Mappings from and to other vocabularies

Nowadays, inter-vocabulary mapping enables the exploitation of subject indexing information to a much greater extent than ever before, as reflected in the latest international standard on thesauri and interoperability, ISO 25964-2 (International Organization for Standardization 2013), including verbal as well as classificatory subject indexing.

For STW, enhanced interoperability has been achieved thanks to vocabulary mappings established in the past, allowing STW content descriptions to be translated into subject information using other vocabularies (and vice versa). Here we can distinguish between different mapping types including mappings on the level of STW descriptors and on the level of the STW subject categories on the one hand, and on the other hand between different mapping procedures including intellectual as well as automatic and semi-automatic mapping approaches. So far, the STW has intellectually been mapped to the subject headings' part of the German Integrated Authority File (GND) and the Thesaurus for the Social Sciences (TSS) of GESIS-Leibniz Institute for the Social Sciences. The original mappings had been built as part of the publicly funded terminology mapping initiative named KoMoHe between 2004 and 2007 (Mayr and Petras 2008). In the meantime, the ZBW and the German National Library (Deutsche National Bibliothek) entered into an agreement to maintain and further develop the cross-concordances cooperatively. Every modification made in one of the two vocabularies is traced back in the cross-concordances. Thanks to the mapping of the STW to GND, descriptors taken from the STW and assigned to bibliographic records in the cataloging system are translated automatically into GND subject headings. It is for this practical reuse scenario that inter-vocabulary mapping marks an important step towards truly collaboratively organized subject indexing beyond library boundaries between central specialized libraries on the one hand and the national library on the other. In addition, other libraries which use the GND can reuse directly what has originally been indexed by using the STW (Dolud and Kreis 2012). Furthermore, mapping candidates and synonyms from other vocabularies are available for use as candidates for further development of the thesaurus. They can be used either as future descriptors or as additional synonyms to serve as entry terms for search enhancement as described later in sections 5.1 and 5.2.

In contrast to the concordance between STW and GND, the concordance between STW and TSS has regularly been updated semi-automatically as part of the library track of the Ontology Alignment Evaluation Initiative (OAEI) in the years 2012, 2013 and 2014. After evaluating the performance of multiple matching tools against a reference set of correct mappings intellectually built up in the past, the tools were used to generate new mapping candidates to speed up the intellectual work of domain experts (Kempf et al. 2014) and to facilitate sustained updating of high-quality vocabulary crosswalks. Further mappings on the level of STW descriptors, which were built up automatically, are mappings to the AGROVOC thesaurus of the Food and Agriculture Organization of the United Nations (verified afterwards by domain experts) and to the DBpedia, as already mentioned in section 3. The latter allows us to provide links from the descriptor pages to the corresponding Wikipedia pages (Figure 2), supporting the users with additional and often exhaustive information about the concepts. Unlike the intellectual and semi-automatically established concordances, which include exact, broader/narrower and related mapping relations, the automatically built concordances include only exact and close matches.

Another mapping procedure under active development is based on the STW subject categories. Unlike the mappings on the level of STW descriptors, such a mapping allows for broader domain-specific classification of an information resource as will be explained in section 4.2. Further reuse scenarios of mappings will be described in the sections 4.4 and 5.1.

4.2 User-generated indexing

The second alternative subject indexing approach focuses on a new user group of thesauri for indexing. Whereas in the past, thesauri were designed for information professionals trained in indexing and searching, today there is a demand for vocabularies that even untrained users should find intuitive. The intention is to encourage economists themselves to use concepts of the STW when indexing their own research papers, this way enabling a participatory culture in scientific publishing, which is usually referred to as Open Science or Science 2.0. This is currently attempted in two ways.

The first one is the free publication of academic literature in economics on the ZBW's open access repository

EconStor (https://www.econstor.eu/). It provides access to and free download of more than 100,000 working papers in the economic sciences. Documents are ingested by automatic harvesting processes or, to a smaller extent, by self-upload. During the self-uploading process on EconStor, scholars are requested to assign descriptors from the STW, which is incorporated as an autosuggest service in the online form for step-by-step bibliographic and content description of the publication. In addition, scholars who want to upload their publications are requested to assign classes taken from the Journal of Economic Literature classification system (JEL), which is the internationally established standard method of classifying scholarly literature in the economic sciences, and which is also offered to scholars as an autosuggest service in the online form. In the same way the STW is indicated in the metadata schema of the "da|ra Registration Agency for Social and Economic Data" as a controlled vocabulary to describe economic research data (Helbig et al. 2014).

The second way to encourage researchers themselves to use concepts of the STW relies on building up mappings between vocabularies researchers are already quite familiar with. Complementary to keywords that are usually assigned freely, economists themselves widely use the JEL when classifying their own publications. When offered a mapping between JEL and STW on the level of its subject category system, economists may be encouraged to provide a more fine-grained content description of their publications by using the STW as a standardized controlled vocabulary. At the moment, a semi-automatic mapping between both vocabularies is being built up, so far using the web-based interactive mapping platform AMALGAME, which has been developed in the context of the Europeana connect project. In the SKOS-based mapping process, both vocabularies are getting enriched. On the part of the STW, the subject categories are enriched by the descriptors assigned to a subject category with synonyms of these STW descriptors in German and English and with all the exact-matching descriptors and their synonyms from those vocabularies which have already been mapped to the STW in the past. On the part of the JEL, classes are enriched by keywords attached to them in the JEL classification codes guide. Preliminary results (see Kempf et al. 2015) confirm that the use of equivalence relations from other mappings leads to a promising selection of mapping candidates, which are evaluated by a domain expert before acceptance.

4.3 Author keywords

A third alternative indexing approach on the basis of vocabulary mappings is the reuse of uncontrolled keywords assigned by the authors themselves. This can be done in two ways. On the one hand, a good number of author keywords can be converted into STW descriptors through the comparison of keywords with descriptors from the STW based on stemming. On the other hand, author keywords may be reused for thesaurus enrichment.

Content-descriptive metadata extracted automatically from bibliographic records are a potential source of suggestions for new candidate STW descriptors or synonyms (Bahls and Rebholz 2015). In the first place, this should support further development of the STW. Secondly, it might lead to a mapping between the STW and author keywords. That approach however may be limited by the ambiguity of natural language in the keywords assigned by authors, as illustrated above with the example of "investment theory." If author keywords do not explicitly find their way into the thesaurus, they may instead be stored in the background.

4.4 Automatic indexing

Furthermore, the STW is used for research and development of a semi-automatic indexing approach. In the future, electronically available full texts will be analyzed automatically by using text and data mining approaches and will be annotated by the use of STW descriptors. Additionally, texts which are not available in electronic form, but which are already provided with content-descriptive metadata, will get annotated automatically by using the STW. This can be achieved by the help of mappings either to other controlled vocabularies or to author keywords as described in section 4.3. So far, experiments on a dataset of open access publications indexed by using the STW are done on the basis of a novel combination of graph-based concept activation methods and kNN (K-nearest neighbors) as a concept selection method (Große-Bölting et al. 2015).

5.0 Thesaurus-enhanced retrieval

Thesauri, in general, aim to support a variety of classical retrieval functions, in the first place, through the use of the core structuring elements inherent to the thesaurus concept: the semantic network of descriptors and secondly, if present, the thesaurus subject category system.

For retrieval purposes, the semantic network structure of a thesaurus can be utilized in many different ways. Synonyms, broader and associated descriptors can be suggested as search terms to the users in cases where no or only few results have been achieved. Narrower concepts can be offered to specify the search query and to narrow down search results. In this way, a thesaurus meets the requirements for a connection between content description and user guidance.

Mere search by title is rather likely to lead to suboptimal search results, for document titles often fulfill marketing purposes to give a reading incentive. Title search is often misleading and produces a considerable amount of ballast. Free text searches, in turn, are often of rather limited use because of semantic variations in the way concepts are expressed and of the multilingualism of the data pool.

The ZBW provides advanced retrieval facilities for economics literature for its users. The EconBiz portal http://econbiz.eu, with more than nine million items from the ZBW itself and various other data providers, is the main entry point for research. The portal is complemented by the digital repository EconStor (as described above). The STW provides search support in both services, but in different ways with different user interfaces. Their respective advantages will now be discussed.

5.1 Index enhancement by search engine

Due to their diverse provenance, the title records in the EconBiz portal have multiple fields with controlled and free keywords. The main sources for controlled terms are the STW and the subject headings' part of the GND. Fast access to the title records is provided by a customized open-source search engine (Solr). The creation and update of the indexes for this search engine can be (and indeed is) tuned in several ways. For the field with STW descriptors, index entries are produced with not only the English and German preferred terms but also all synonyms of the descriptors. Thus, the record is found even if it does not contain the searched term, as long as this term occurs as a synonym in the thesaurus. Since the STW index is merged into the main search index which is addressed by EconBiz' search box, the synonym enhancement is active for all "simple" searches—without any special user interaction.

The main advantage of this solution to thesaurus-based index enhancement is speed. It delivers search results very quickly, since special processing is done during index creation and update. Furthermore, the index, with its various entries originating from free text and controlled keyword fields, can be tuned in a rather uniform way. Boosting rules can be applied, which for example rank results higher when a search term appears in the title field of the record or in a thesaurus-enhanced keyword field. A disadvantage of the index enhancement approach is the fact that new versions of the thesaurus require a time-consuming rebuild of the index. This effort is accepted, since an index rebuild can be run in the background without affecting the user.

Currently, the STW subject field is enhanced with STW synonyms, and the GND subject field is enhanced with GND synonyms. However, preparations are ongoing to take advantage of the vocabulary mappings described above. For the GND field, this will be a two-step approach where in the first step the GND/STW mapping and, in a second step, all the other STW mappings mentioned above will be used. For each GND descriptor, matching exactly with a STW descriptor, all preferred and non-preferred terms from all exact matching entries in other vocabularies will be used to enhance the index of the GND field. The same inclusion of terms obtained from matching descriptors in other vocabularies will be applied to the STW field, thus providing additional entries for search. This will support users who are only familiar with one particular vocabulary to bridge the gap to the other vocabularies also used in parts of the EconBiz data, as well as users who select their search terms without any knowledge of the existence of controlled vocabularies at all.

5.2 Search enhancement and alternative concept suggestion

A different approach has been taken at the EconStor digital repository where about half of its 100,000 papers are indexed with STW descriptors by researchers using the self-upload functionality (see above) or by ZBW indexers. In EconStor, the use of the STW for search is not woven into the index of the underlying DSpace system. Instead, visible options are presented to users. Each search phrase a user enters is sent not only to DSpace but also to a STW-based web service. The service returns matching descriptors, by executing a full-text search on the thesaurus, and for each of these descriptors all related and narrower descriptors. In the same response, all synonyms for these descriptors (and, again, all synonyms of exactly matching descriptors in other vocabularies) are returned. Thus, a cloud of semantically related "tags," namely the STW descriptors, can be displayed together with the generic search results of the query. When the user picks such a tag, all of the attached synonyms are inserted into the search box, connected by "OR," and the query is executed again. In the example shown in Figure 3, "deindustrialization" was searched. This resulted in 381 hits. When the user clicks the "Deindustrialization" tag, synonyms are included, e.g. "de-industrialisation" or the interchangeably used German terms "deindustrialisierung" and "entindustrialisierung," and the new search earns 656 hits. Since the search is executed against an index which includes full text, it also covers documents which are not indexed by using STW descriptors. Additionally, besides the concept directly matched in the thesaurus, related and narrower concepts are suggested to the user. According to their particular research interests, a search for the descriptors "De-


EconStor >

Search Results

Search:

All of EconStor

for (("Deindustrialisierung") OR ("Deindustrialization") OR ("deindustri: GO

Deindustrialization Resource wealth
Old industrial region Declining markets

Results 1-10 of 656.

Item hits:

Date	Title	Authors
2006	Deindustrialisierung: Eine neue 'britische Krankheit'?	Scheuer, Manfred / Zimmermann, Guido
2011	Manufacturing productivity, deindustrialization, and reindustrialization	Tregenna, Fiona
1994	Deindustrialisierung als Wachstumsbremse?	Klodt, Henning
1998	The causes of welfare state expansion: deindustrialization or globalization?	Iversen, Torben / Cusack, Thomas R.

Figure 3. STW-expanded EconStor search, after searching for "deindustrialization" and clicking the link labeled "Deindustrialization"

clining markets" or "Old industrial region" may be chosen (Figure 3). When clicked, these descriptors in turn are searched with all their synonyms, and the upcoming result page of the new search reveals more possibly relevant concepts, like "Industrial geography" or "Regional revitalization."

The discovery of related concepts, which normally have no syntactical similarity to the original search terms, is a big advantage of this approach. Whether these concepts are relevant for the current research focus can only be decided by the user herself, not by clever algorithms. As they could lead to surprising results, they are not included automatically. Therefore, user interaction is still needed.

While the EconBiz search follows a Google model (simple search box, fast and internally optimized results), the EconStor search supports a more explorative approach. In both scenarios, the STW thesaurus has proven to be well suited.

5.3 Search results faceted by upper level subject categories

STW descriptors are grouped into about 500 subject categories[3], which form the classification hierarchy in the seven sub-thesauri described above. Descriptors often belong to more than one subject category—the concept "productivity" for example exists in economics as well as in business economics. Historically, this hierarchy was used only to organize the thesaurus itself. With the extensive changes during the STW overhaul, simple alphabetical listing of new or deprecated descriptors became more and more impractical, and subject categories assumed a new function. The change reports referenced in section 3.0 and the graphical overviews allowing exploration of the vast amount of changes by area of interest are intersected by the about 100 second-level subject categories to which the descriptors are attached.

While this still only affects the thesaurus and its development, experiments are under way to project the category system on to the indexed content. This would overcome the shortcomings of the currently implemented faceting by keyword. For users, some keywords (like "Germany" or "theory") are not really helpful. Others are very fine-grained and not suitable for narrowing possibly large result sets to a field of subjects relevant to the query. The faceting by upper level subject categories, derived from the descriptors actually used, and the exclusion of particularly the "General descriptors" and the "Geographic names" sub-thesauri will hopefully result in a meaningful intersection of result sets, according to the structures of the knowledge domain as modeled in the thesaurus.

```
SELECT DISTINCT ?term ?concept ?prefLabel
WHERE  {
  ?concept text:query ('labour') .
  {
    ?concept skos:prefLabel ?term
  } UNION {
    ?concept skos:altLabel ?term
  }
  FILTER regex(str(?term), "^labour", "i") .
  ?concept a zbwext:Descriptor .
  ?concept skos:prefLabel ?prefLabel .
  FILTER (lang(?prefLabel) = "en") .
}
ORDER BY lcase(?term)
LIMIT 10
```

Figure 4. Example of an optimized SPARQL query, as used for the "suggest" service. It exploits an installation-specific text index to get all occurrences of the the searched keyword "labour" very quickly, and later on restricts the results in a more expensive operation to the strings starting with this keyword.

One minor and one major problem occur. The minor one results from the fact that many descriptors belong to multiple subject categories, and therefore will be included in multiple facets. This is not really a problem—sometimes it is perfectly legitimate to have a title in different facets, and sometimes the result is noisy, but still better than an un-faceted result set. The major problem is that not all content is indexed with STW descriptors. As with any faceting system, this may exclude relevant results in a way not transparent to the user. One approach to mitigate this effect will be the exploitation of mappings, whereby, for example, GND subject headings may be mapped via their matching STW descriptors to STW subject categories. When available, the mapping of JEL classes to STW subject categories as discussed above can be used for the same purpose in an even more direct way. Technically, it should even be possible to assign STW subject categories from author keywords or from automatic keywords extraction. The practical usefulness of this approach, however, will have to be evaluated carefully.

6.0 New uses for Linked Open Data

Many of the new forms of use of the STW described above are inspired or entirely enabled by the rise of linked open data on the web. Datasets like the GND and its predecessors or mappings between thesauri have existed long before. Yet they were available only by special agreements between institutions, and in a format which made every dataset and every mapping a special case for integration. Typically this required substantial efforts in handling and custom programming. Thus, integration of such resources was done rarely; often it lasted only for the lifespan of a particular project. The LOD wave very much sped up developments, in that it empowered single institutions or even individuals, such as Ed Summers (2008), who published the Library of Congress' subject headings as a prototype on his personal website in 2008, to publish their own data and make use of data published by others. This allowed network effects to come into play. Today, it is almost common that thesauri are published in SKOS format, and very often can be downloaded without further requirements. Building on this, a mapping from the labor law thesaurus by Wolters-Kluwer Germany to the STW has been created without any effort from the side of ZBW.

However, SKOS download files impose considerable costs on the data consumer, particularly for less tech-savvy users lacking knowledge in specific Semantic Web technologies. That is also true for public SPARQL endpoints, which can be queried in a very flexible, SQL-like way for particular data but require a deep knowledge of the data structures and the SPARQL syntax.

6.1 econ-ws web services

To overcome these hurdles for adoption, the ZBW has early on published the STW data additionally in the form of web services services (Neubert 2012) at http://zbw.eu/beta/econ-ws. These web services execute predefined and pre-optimized SPARQL queries (Figure 4), each service for a particular use case. The results are delivered in XML or JSON format, which are familiar to web developers. The JSON-LD format, since 2014 a W3C recommendation, even goes a step further by combining the richness of the full RDF data with the simplicity of JSON. While many application programming interfaces require a sequence of fine-grained function calls, the design philosophy in econ-ws is that the whole, possibly complex information re-

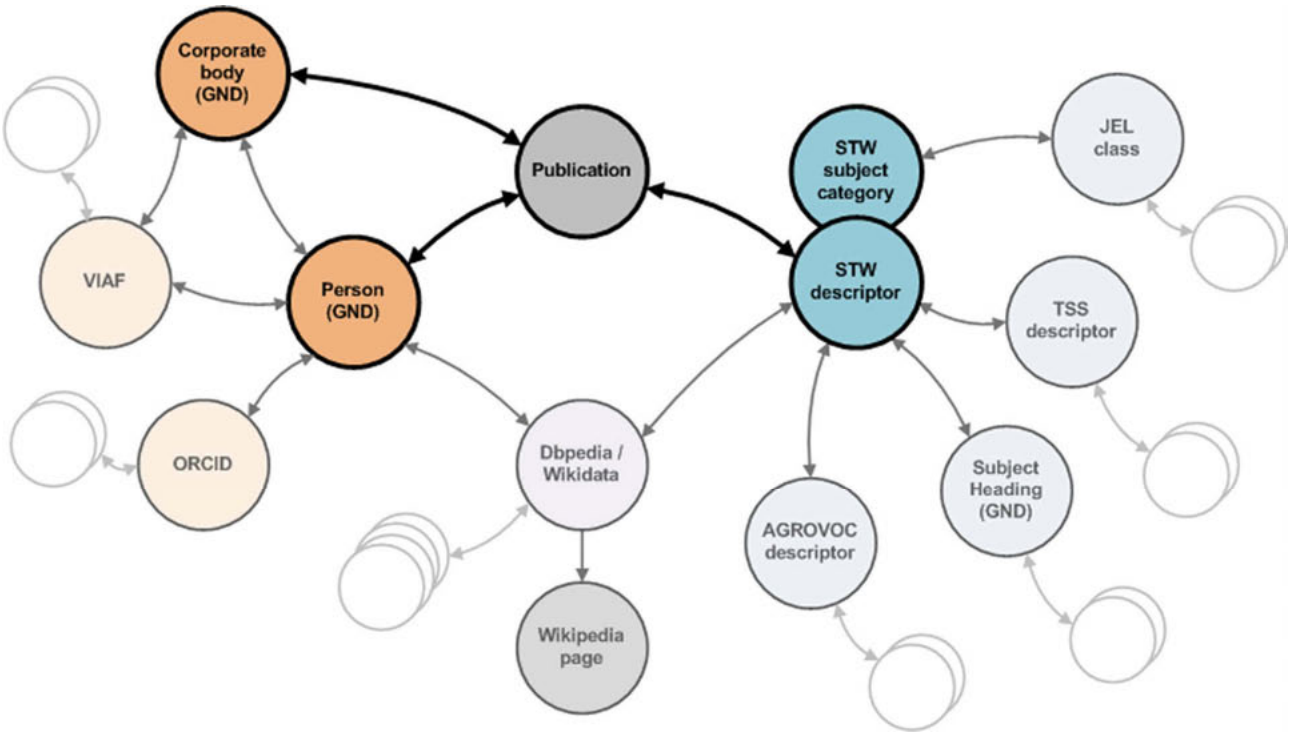


Figure 5. Publications as linking hubs between subjects and persons/corporate bodies.

source should be delivered in one request/response cycle. That is true even for the most complex of these services, which delivers all information which is used in EconStor for enriching search terms and suggesting alternative concepts as described above. The programmers of the EconStor web application were able to integrate the enhancement solely on the level of the result page, without touching the code of the underlying DSpace application. In the same way, integration into other retrieval applications working on economics data becomes possible. Since the executed query uses all synonyms, its use does not require that the searched content be indexed with the STW or any controlled vocabulary at all.

A similar service just delivers synonyms of a term, optionally including synonyms derived from the mappings to other vocabularies. It can be utilized in the same shallow form of integration into third-party web applications for improving their search results. The autosuggest service, on the other hand, supports data input. From a partial input of potential keywords, it looks up matching descriptors or synonyms, and allows selection of the correct concept, while in the background the identifier/URI is transferred and can be stored in the application. The suggest service supports the self-upload in EconStor described above and will assist users in several research data repositories currently under development, side by side with a suggest service for descriptors of the TSS operating on the same technical infrastructure.[4] We think that a simple way to select descriptors (and automatically transfer and save identifiers/URIs) is essential for the generation of high-quality links in the web of data, as well as for the proliferation of the controlled vocabularies.

To support the use case of user-generated indexing (section 4.2) in content management systems (CMS), as a proof of concept, a plugin for the "Web taxonomy" Drupal module (https://www.drupal.org/project/web_taxonomy) has been created. It accesses the econ-ws web service to auto-suggest STW descriptors and thus allows authors of blog entries or other CMS content to "tag" their pieces in a familiar way, yet with well defined and clearly linked concepts from a large controlled vocabulary. A similar plugin exists for the *Getty Art and Architecture Thesaurus*. These vocabularies are not imported into the CMS and, even more important, do not require local updates when new versions are published. This reduces the workload for the CMS site administrators considerably, and hopefully will encourage adoption.

6.2 Crossing the borders of books and vocabularies

When we look at bibliographic data not from a traditional librarian's point of view but from a linked data perspective, we can, for a moment, forget all the details of the books and articles it describes. We can see them as mere linking hubs in a network (Figure 5), connecting concepts to persons and corporate bodies.

Given a system where the concepts, persons and corporate bodies are represented not as strings but as entities, we can extend the coverage of our knowledge organization system from bibliographic metadata to authors and even (research) institutions, which, e.g., act as editors or as an author's affiliation.

We then can start asking questions like: What are the main subjects of an author? Who are the most productive authors for a given subject? Which institutions are relevant for a certain subject field?

And, indeed, we can build a system to answer such questions, currently as an early prototype of an EconBiz research dataset. When loaded into a triple store (in other words a database for linked data), we can translate the questions above into database queries, which group and aggregate data from different angles. Furthermore, since our own data are linked to multiple other datasets on the web, we can draw from these (e.g., for the nationality of an author or the seat of an institution), and even ask questions like: which institutions in the European Union do research on a certain subject?

To answer these kinds of questions, current research information systems (CRIS) have sometimes been built, with a lot of effort and possibly duplicate data ingestion. Libraries using knowledge organization systems that are well adapted to their domain are in a good position to contribute here. Without a knowledge organization system as a backbone, the desired aggregation for subjects or subject fields is simply unattainable.

7.0 Conclusion

As we have seen, thesauri nowadays are capable of forming a core component for a huge variety of new applications. The different ways in which they are integrated within these applications clearly reflect how the changing role of thesauri is interlinked with overall changes of modern information infrastructure services as a whole. First of all, due to resource constraints, it is becoming more and more impossible to index all the items of every major collection intellectually. Apart from that, 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 folksonomies. In the end, even though some of these portals and discovery systems are trying to create "closed gardens," they in turn are part of the much larger open web, which has already become the place to search for information.

Not only can an ever decreasing part of content relevant to information seekers be covered by subject indexing and classical terminological control, but subject indexing as a whole and its significance and use for browsing and retrieval is changing fundamentally. As we have seen by referring to the example of the STW, the process of subject indexing has changed into a complex interaction of different interdependent and interwoven indexing components, with user-generated and (semi-)automatic indexing alongside traditional intellectual indexing and inter-vocabulary mapping for third-party indexing transfer. At the same time, it is important to be aware that 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 applications increases. Since alternative process-based approaches rely on such a high-quality core dataset, decisive for the future is an effective interplay between intellectual and machine-based indexing (Stumpf 2015). Only a considerable amount of intellectual subject indexing guarantees staying in touch with the scientific field for further development and constant re-adjustment of a thesaurus according to the latest developments within a discipline. This normally requires a team of domain experts and information scientists working closely together, who can focus on the development of the vocabulary itself and have a clear idea of how the use of the terminology can change in the course of time. And it also requires subject indexers, who apply the vocabulary to new publications in the field. They are the ones who spot missing terms and concepts, they feel the pain when concepts lack sufficient discriminatory power, and over time, their indexing results build an indispensable feedback loop about the relevance of the concepts of the thesaurus in the field.

Application of thesauri for intellectual indexing as well as their development and routine maintenance are in the case of the STW and in many others subsidized by the taxpayer—and hardly would have been possible otherwise. We would argue that these investments have resulted and furthermore will result in improved services (well aware of the fact that, similar to other services academic libraries and scientific infrastructure bodies provide for their students, their research communities and the general public, the return on investment cannot be determined monetarily in any plausible way).

The various interacting subject indexing components described here lead to new browsing and advanced retrieval scenarios and applications. At the query stage, synonym enhancement and vocabulary mappings can lead to additional index entries and may be used to generate semantically related concept suggestions. When the results are presented, subject information taken from a controlled vocabulary can enable advanced faceted browsing. As we have seen with the STW, the power of descriptors and their semantic structure can be extended via subject categories. These effectively form an additional knowledge

organization system for indicating the overall domain-specific context of a publication, and they enable additional vocabulary mapping scenarios on a higher aggregation level.

Web services can effectively support the use of a thesaurus without requiring a full import and periodic update of large data files in a variety of applications, for retrieval as well as for data input. Thesaurus-driven search enhancement is not constrained by difficult-to-integrate specific thesaurus management software, but can be added as a loosely coupled component to retrieval systems. The simple selection of auto-suggested concepts during data input and the transfer of universal unique identifiers in the background are essential for the creation of high-quality links between information resources in the economic sciences.

These links form the basis of the open web of data, which far exceeds the bibliographic domain. Links connect researchers and institutions, publications and research datasets, with subjects and broader subject fields. The opportunities offered here on the web at large have been hardly touched. More than ever, it is true that only the most recent technological developments help thesauri to exploit their full potential.

Notes

1. In short STW—the acronym refers to the German name "Standard-Thesaurus Wirtschaft."
2. The institutions which helped to develop the STW and which among others still use the STW for subject indexing were the two predecessor institutions of today's ZBW-Leibniz Information Centre for Economics/ German National Library of Economics: Bibliothek des Instituts für Weltwirtschaft—Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Kiel, and Hamburgisches Archiv für Weltwirtschaft (HWWA), GBI Gesellschaft für Betriebswirtschaftliche Information (now: GBI-GENIOS), and the ifo Institute (now: CESifo Group Munich).
3. STW descriptors and subject categories both are defined as subclasses of skos:Concept.
4. E.g., the economics and social sciences data repository SowiDataNet (https://sowidatanet.de), or EDaWaX (http://www.edawax.de), which aims at making economics research reproducable.

References

Baca, Murtha and Melissa Gill. 2015. "Encoding Knowledge Systems in the Digital Age: The Getty Vocabularies." *Knowledge Organization* 42: 232-43.

Bahls, Daniel and Tobias Rebholz. 2015. "Evidenzbasierte Begriffs- und Synonymerweiterung des STW." In *Proceedings 104. Bibliothekartag, 26-29.05. 2015*. Nürnberg NCC.

Baker, Tom and Osma Suominen. 2015. "Global Agricultural Concept Scheme. The Collaborative Integration of Three Thesauri." Presentation at the DINI Annual Meeting, Bonn, Germany. http://de.slideshare.net/CIARD_AIMS/globgl-agricultural-concept-schemethe-collaborative-integration-of-three-thesauri

Berners-Lee, Tim. 2006. "Linked Data." http://www.w3.org/DesignIssues/LinkedData.html

Bertram, Jutta. 2005. *Einführung in die inhaltliche Erschließung. Grundlagen - Methoden - Instrumente*. Würzburg: Ergon Verlag.

Deutsches Institut für Normung. 1987. *Erstellung und Weiterentwicklung von Thesauri*. Teil 1: Einsprachige Thesauri. DIN 1463. Berlin: Beuth Verlag.

Dolud, Lena and Constanze Kreis. 2012. "Die Crosskonkordanz Wirtschaft zwischen dem STW und der GND. Ein Instrument zur kooperativen Inhaltserschließung und zur Vernetzung im Semantic Web." *Dialog mit Bibliotheken* 24 no. 2:13-19. http://hdl.handle.net/11108/76

Gastmeyer, Manuela. 2000. "Der Einsatz des Standard-Thesaurus Wirtschaft im HWWA. Ein Instrument zur Qualitätssicherung von wirtschaftswissenschaftlicher Fachinformation." *Auskunft. Mitteilungsblatt Hamburger Bibliotheken* 20:108-30. http://zbw.eu/stw-info/pub/gastmeyer_stw_hwwa.htm

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

Helbig, Kerstin, Brigitte Hausstein, Ute Koch, Jana Meichsner and Andreas Oskar Kempf. 2014 *da|ra Metadata Schema: Version 3.1*. GESIS Technical Reports 2014/17. Köln: GESIS. http://www.da-ra.de/fileadmin/media/da-ra.de/PDFs/TechnicalReport_2014-17.pdf

Informationszentrum des HWWA. 1992. *Thesaurus Wirtschaft*, Band 1, 2. Hamburg: Verlag Weltarchiv.

International Organization for Standardization. 2013. ISO 25964-2: *Information and Documentation—Thesauri and Interoperability with other Vocabularies - Part 2: Interoperability with other Vocabularies*. Geneva: International Organization for Standardization.

Kempf, Andreas Oskar, Dominique Ritze, Kai Eckert and Benjamin Zapilko. 2014. "New Ways of Mapping Knowledge Organization Systems: Using a Semi-Automatic Matching Procedure for Building up Vo-

cabulary Vocabulary Crosswalks." *Knowledge Organization* 41: 66-75.

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

Mayr, Philipp and Vivien Petras. 2008. "Cross-Concordances: Terminology Mapping and its Effectiveness for Information Retrieval." In *Programme and Proceedings of the 74th IFLA World Library and Information Congress*. Québec, Canada. http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf

Neubert, Joachim. 2009. "Bringing the 'Thesaurus for Economics' on to the Web of Linked Data." In *Linked Data on the Web (LDOW 2009) Proceedings of the WWW Workshop on Linked Data on the Web, Madrid, Spain,April 2, 2009*, ed. Christian Bizer, Tom Heath, Tim Berners-Lee and Kingsley Idehen. CEURWorkshop Proceedings 538. http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf

Neubert, Joachim. 2012. "Linked Data Based Library Web Services For Economics." In *DC-2012—The Kuching Proceedings: Papers, Project Reports and Posters for DC-2012 in Kuching, Sarawak, Malaysia, 3-7 September 2012. DCMI* International Conference on Dublin Core and Metadata Applications. http://hdl.handle.net/11108/82

Neubert, Joachim. 2015. "Leveraging SKOS to trace the overhaul of the STW Thesaurus for Economics." In *DC-2015--The São Paulo Proceedings: Papers, Project Reports and Posters for DC-2015 in São Paulo, Brazil, 1-4 September 2015*. DCMI International Conference on Dublin Core and Metadata Applicatications. http://hdl.handle.net/11108/203

Stumpf, Gerhard. 2015. " 'Kerngeschäft' Sacherschließung in neuer Sicht. Was gezielte intellektuelle Arbeit und maschinelle Verfahren gemeinsam bewirken können." In *Fortbildungsveranstaltung für Fachreferentinnen und Fachreferenten der Politikwissenschaft und Soziologie*. Berlin: Hertie School of Governance. https://opus.bibliothek.uni-augsburg.de/opus4/files/3002/Stumpf_Sacherschliessung.pdf

Summers, Ed, Antoine Isaac, Clay Redding and Dan Krech. 2008. "LCSH, SKOS and Linked *Data." In DC-2008--Berlin Proceedings: Papers and Project Reports for DC-2008 in Berlin, 22-26 September 2008*.DCMI International Conference on Dublin Core and Metadata Applications. http://dcpapers.dublincore.org/index.php/pubs/article/viewFile/916/912

Suominen, Osma, Sini Pessala, Jouni Tuominen, Mikko Lappalainen, Susanna Nykyri, Henri Ylikotila, Matias Frosterus and Eero Hyvönen. 2014. "Deploying National Ontology Services: From ONKI to Finto." In *SEMWEB (2014)*. http://seco.cs.aalto.fi/publications/2014/suominen-et-al-deploying-onki-finto-2014.pdf