Kempf, Andreas Oskar; Neubert, Joachim

Book Chapter — **Published Version**

STW Thesaurus for Economics

Mitglied der

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Leibniz-Gemeinschaft

STW Thesaurus for Economics

by <u>Andreas Oskar Kempf</u> and <u>Joachim Neubert</u>

(This article is an updated version of an article from 2016, see colophon).

Table of contents:

Abstract:
The STW Thesaurus for Economics is an ISO-25964-compliant thesaurus in German and English, applicable and freely reusable in the broader field of economics. This paper relates the history of the STW, describes its structure and usage, and illustrates the possibilities of its reuse. The article covers the entire development of the STW from the early stages, when it was primarily a subject-indexing tool, to its use in newer applications, such as web services, and its integration into research-based approaches for the automation of subject indexing.

1. Introduction

STW is an acronym for "Standard-Thesaurus Wirtschaft". The STW Thesaurus for Economics [1] is the world's most comprehensive bilingual → thesaurus for economics, developed as a general → indexing and retrieval tool for subject-specific information in economics. With almost 6,000 subject headings in English and German and more than 20,000 synonyms, it covers all economics-related subject areas and, on a broader level, the most important related subject fields. The ZBW - Leibniz Information Centre for Economics, Germany's National Library of Economics, publishes and continuously updates the STW to keep current with the

latest changes in economics terminology.

Section 2 provides a brief history of the STW, while section 3 presents the STW in its current form, referring to its publication on the web, and describes its maintenance and development. Section 4 presents an example to illustrate the STW's structure and its specific construction principles. Section 5 discusses how the STW interlinks with other vocabularies via mappings since interoperability and integration into linked data applications are becoming increasingly important. Section 6 describes how the STW is used with regard to subject indexing and information retrieval. Section 7 shows how third parties utilize the STW. Finally, section 8 concludes with considerations for the future. This article leans on a more detailed *KO* article (Kempf and Neubert 2016).

2. A brief history of the STW

The STW resulted from the need for a common standardized indexing and search language for business information and economic literature. During the 1990s, leading German providers of business and economics information services maintained their own documentation languages. When information technology made it possible to bring together the content of different library catalogs and databases and to publish it on CD-ROM, users were disappointed to find that integrated content search was unsupported.

Four leading public and private German providers of business and economics information services jointly developed the STW in a project funded in part by the German Federal Ministry of Economics, during the years 1995-1997. The Library of the Kiel Institute of the World Economy (IfW Kiel), founded in 1919, and the Information Centre of the Hamburg Institute for the International Economy (HWWA), founded in 1908 as information center of the Hamburg Colonial Institute, were the two predecessors of today's ZBW - Leibniz Information Centre for Economics [2]. The different indexing languages ranged from a simple → keyword list to two fully-fledged thesauri, as in the case of the two predecessor institutions of today's ZBW. Each indexing language reflected the specific collection focus of the participating institutions, whether based on empirical or theoretical literature, on economics or on business economics and practice. Depending on the history of the individual institutions involved, the collection focus could also change significantly.

2.1 Precursors of the STW

The history of the STW, in fact, reaches back much further. In the 1970s, the situation in other countries convinced West German authorities at the federal level of the importance of information and documentation. This triggered the launch of several large-scale funding programs to establish specialized information centers all over the country. In addition, it became obvious that subject catalogs and classifications, which also existed, no longer were satisfactory solutions for subject indexing and access, due to increasing specialization and changes in the content structure of publications.

The IfW Kiel had begun developing its subject catalogs in the 1950s, and the HWWA as far back as the 1920s. In order to combine the ordering principles of

subject catalogs and classifications, both predecessors of today's ZBW undertook extensive preparatory work to convert their subject catalogs into thesauri.

The vocabularies of the two precursor thesauri, Sachkatalog (Bibliothek des Instituts für Weltwirtschaft an der Universität Kiel 1980) and Thesaurus Wirtschaft (HWWA-Institut für Wirtschaftsforschung 1987), formed the basis of the STW. The thesaurus published by the library of IfW Kiel in 1980 consisted of more than 6,000 concepts, both in German and English. The conversion of the subject catalog into a thesaurus set the stage for a future standardization project. Even at that time, there was a strong national and international demand for a common standardized documentation language in economics.

In 1987, the information center of the HWWA published its thesaurus, which was based on its former subject catalog of about 13,000 subject headings. Taking into consideration the core standards and guidelines of that time, the structure and essential principles of the later STW were already recognizable in this predecessor thesaurus.

The conversion of the subject indexing tools was also accompanied by far-reaching institutional transformations. Both institutions initially developed independently from institute libraries to supra-regional subject libraries. This meant not only an increase in the number of staff, but also the establishment of separate scientific services with specialized scholarly staff, partly with a doctorate in economics who developed and maintained the thesauri.

3. The STW in its current form

At first, the standardization of content description in business and economics information primarily addressed a German-speaking scientific community. As a result, the first version of the STW, published in 1998, was available only in German. With this essentially German-speaking user group in mind, the STW claimed to serve as a standard, as reflected in the German name of the thesaurus. As with other disciplines, however, the scientific discourse within sub-disciplines of the economic sciences increasingly became more international. In 2007, the ZBW exclusively assumed responsibility for the maintenance and development of the STW and consequently directed its information services towards an international, English-speaking, scientific community and a greater level of interoperability in the modeling of concepts. As of version 8.02, released in 2007, all descriptors in the STW are bilingual, in German and English.

In addition, the publication of the STW on the Web and its first complete overhaul decisively shaped the current appearance of the thesaurus.

3.1 Web publication

The goal of the web-based version of the STW in 2009 was its reusability by people as well as by machines (Neubert 2009). The web version applied two main technologies: SKOS, the Simple Knowledge Organization System for defining concepts and their interrelations, and RDFa, a technique for embedding machine-readable data into human-readable web pages, as described in section 4. At the time that both technologies were in their final states of standardization by the World

Wide Web Consortium, the Linked Open Data on the web movement was just emerging. A few extensions (e.g., for the category system of the STW) were added to cover all semantic structures of the STW. The web pages and all related services derive from one set of publicly available SKOS files obtained and converted from the maintenance system. This ensures consistency across diverse methods of content publication.

The highly interlinked and completely bilingual web site provides persistent identifiers (URI) for each concept, independent of language and version. When new versions of the STW are published, all older versions remain accessible [3]. The STW maintains obsolete concepts as "deprecated", and frequently points to a replacement, so institutional users can trust that once-published concept identifiers remain resolvable. The STW web version includes a change tracking functionality, which provides users with news and updates. The underlying framework, while developed for the STW, can be transferred to any other SKOS vocabulary (Neubert 2015).

3.2 Maintenance

The editorial board of the STW consists of four subject librarians with qualified expertise in economics and a knowledge organization specialist serving as editor-in-chief. The editorial board regularly verifies and decides on changes and updates to the vocabulary and the structure of the thesaurus. Suggestions for new descriptors come from ZBW's subject librarians as well as from colleagues from other institutions that use the STW for subject indexing. A new version is released annually.

Editorial procedures for updating correspond to those depicted in ISO 25964-1 (see clauses 13.6.2 and 13.6.3). An easy mechanism for suggesting changes is provided for all subject indexers and for external users. The procedures cover all types of changes as listed in ISO 25964-1 (see clause 13.6.4). The decision to include a concept in the STW is primarily based on the subject-specific relevance and the frequency with which this concept is made a topic in the literature. Other decision criteria include currency, the need for a concept, provenance, → interoperability, unambiguousness, and the level of specificity. The ability to form a concept by post-coordination is also a decision criterion, in order not to let the thesaurus become too extensive. In particular, in the sub-thesauri for "Economic Sectors" (W) and for "Commodities" (P) preferred terms often need to be combined with other preferred terms. However, to keep the number of results for some concepts manageable and to avoid false associations, sometimes a pre-combined form is used, e.g. `Waste trade`.

Over the last more than 20 years, the STW has undergone constant revision due to profound changes in the field of economics, as for example developments in the financial sector, as well as new methods applied in economics. After more than 15 years of permanent updating, the STW had been completely revised over a period of several years. These changes resulted in the addition of over 750 concepts and the elimination of nearly 1,100 concepts. Other changes consisted of new and updated entry terms and enhancements to the category system.

More recently, the ZBW undertook a pilot study to identify which additional sources are best suited for the further development of the thesaurus. The study considered the following sources: publication titles, author keywords, abstracts, and

log files of search queries in ZBW's search portal *EconBiz* [4]. The text material underwent various filtering processes before a subset was recommended to the editorial board as candidates for new descriptors. The pilot study showed that author keywords were best suited as an additional source for descriptor candidates, followed by search queries and publication titles, both almost equal in ranking with each other. The results also showed that abstracts were the least suitable (Prange 2016). In the near future, the next steps will be to apply the results of this study to include other sources in addition to the existing ones for new descriptor candidates.

4. Design and structure

The STW is compliant with the latest ISO-thesaurus standard, i.e. ISO 25964 (2011). Some core construction principles of this standard characterize the STW in particular.

The STW is very user-friendly, which is evident by the considerable number of entry terms. Synonyms and quasi-synonyms are selected for their efficiency in guiding the choice of terms for indexing and searching and by anticipating the degree of discrimination required at the time of searching (see ISO 25964-1 2011, Clause 8.3). Specific terms might be subsumed by a broader concept, especially in the Related subject areas (see ISO 25964-1 2011, Clause 8.4).

Furthermore, the STW consists of an elaborate systematic structure in the form of domain-specific subject categories on as many as four different hierarchical levels. These categories, as well as the descriptors, are concepts according to the definitions of ISO and SKOS. (Different from ISO terminology, → "descriptor", here, is not used as a synonym for a preferred term.) A navigation tree on the web page (see Figure 1) allows STW users to browse the descriptors of a certain subject field thematically. While terminological control and semantic relations of a descriptor indicate the narrower content-specific relationships of that descriptor, the connected subject categories point to the larger domain-specific context.

The first level of the subject categories 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 for "Economic sectors" and one for "Commodities". The latter two rely on 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 a sub-thesaurus "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 perspective of the economic sciences, with, for example, a particular focus on statistical and mathematical methods used across their subfields.

The → notation code for subject categories consists of letters and numbers. The capital letter denotes the sub-thesaurus of the STW; the number refers to the partial subject field to which the descriptor belongs. A descriptor can be assigned to more than one subject category. Each subject category also has its own page. Numerous subject category pages contain links to the notation of other classification schemes, in particular to the Journal of Economic Literature

Classification Scheme (JEL).

Figure 1: On the left, the subject categories navigation tree allows for systematic browsing. On the right, the subject category page lists all the descriptors assigned to the corresponding subject category, and it contains all relevant information about it.

All concepts in the STW, represented by terms, are subject to terminological control. All of the information required for this is summarized on the corresponding descriptor page, as shown in Figure 2. The preferred term appears at the top of each descriptor page in English and German. Usually, the most common term is the preferred term. "Used for" lists the synonyms and quasi-synonyms that serve as lead-in entries. In some cases, a scope note in italics follows to define or clarify the semantic boundaries of a concept. Where necessary, an additional note may serve to clarify further by showing what is excluded from the concept.

Further down, the page lists descriptors in a hierarchical or associative relationship to the corresponding descriptor. The indication of broader and narrower terms allows users to navigate semantically to more general or to more specific terms and vice versa and to trigger a search with the corresponding terms in *EconBiz*.

The category Related Terms contains terms with a content-related relationship to the respective descriptor but which do not have a hierarchical relationship to it.

The descriptor is then assigned to one or more subject categories, which are represented by the corresponding notation codes. Below are links to descriptors of other thesauri or entries to other vocabularies.

Home
STW Relaunch
Alphabetical descriptor list
Mappings
Versions
Web Services
Downloads
About

▸ V Economics
▾ B Business economics
  B.00 Business economics
▸ B.01 Management and business organization
▸ B.02 Corporate finance and capital budgeting
▸ B.03 Accounting
▸ B.04 Human resource management
▸ B.05 Materials management and logistics
▸ B.06 Production management
▸ B.07 Marketing
▸ B.08 Corporate tax management
▾ B.09 Business information systems
  B.09.01 Information system components
  B.09.02 IS development and management
  B.09.03 Corporate information systems
▸ B.10 Operations research
▸ W Economic sectors
▸ P Commodities
▸ N Related subject areas
▸ G Geographic names
▸ A General descriptors

Information technology EB

Informationstechnik (german)

used for: Information and communication technologies, ICT (Information and communication technologies)

For IT equipment, software, networks, etc. in a company or organization USE IT infrastructure

Broader Terms

▪ Electrical technology EB

Related Terms

▪ Computer EB
▪ Computer science EB
▪ Data processing EB
▪ ICT industry EB
▪ Office technology EB
▪ Telecommunications engineering EB

Subject Categories

▪ B.09 Business information systems ▾
▪ N.10.04.01 Key technologies ▾
▪ N.10.05.01 Information technology and systems architecture ▾
▪ P.10.04 ICT equipment ▾
▪ W.19 ICT industry ▾

Links to other Thesauri and Vocabularies

= Informationstechnik (from GND)
= information technology (from Wikidata) W
= Information technology (from DBpedia)
= information technology (from EuroVoc)
~ information technology applications (from EuroVoc)
= information technology (from TheSoz)
= information technology (from TheSoz)
~ communication technology (from TheSoz)
= information technology (from AGROVOC)

Persistent Identifier (for bookmarking and linking)

▪ http://zbw.eu/stw/descriptor/16757-5

Figure 2: On the left, the subject categories navigation tree allows for systematic browsing. On the right, the descriptor page contains all relevant information about a descriptor. The "EB" icon triggers a search for publications in *EconBiz* with the corresponding descriptor. Links to mapped thesauri (e.g., to Eurovoc or the Thesaurus for the Social Sciences) are indicated in the lower right corner.

5. Interoperability

The STW interlinks with various other vocabularies via mappings to STW descriptors as well as STW subject categories. The mappings can be differentiated according to the different mapping approaches that were used, e.g. intellectual, semi-automatic, automatic (Kempf and Neubert 2016). More importantly, different parties created the mappings either as a one-sided and one-time-approach, in the case of Agrovoc, or in the form of continuous updates by a joint editorial team, in the case of the mapping to the German Integrated Authority File (GND) [5]. This results in different levels of maintenance. In addition, institutions involved in the mapping process can choose to only consider and display the mappings to a certain extent, depending on the particular subject focus of the institution. In order to facilitate interoperability, the STW development team created a method to overwrite individual relationships in a third-party mapping and apply it to externally created mappings to the STW.

More recently, descriptors of individual sub-thesauri have been mapped to the community-driven knowledge base *Wikidata*, which not only supports many language-specific editions of *Wikipedia*, but has also evolved to be a global data hub, linking to more than 5,000 external databases by identifier. This includes thesauri

and classifications, for which a specific mapping to the STW would not be feasible. A derived mapping, via bilaterally linked *Wikidata* items, could prove useful.

6. Usage

The STW facilitates → knowledge organization and supports information retrieval in a variety of ways.

6.1 Subject indexing

Besides the institutions that jointly developed the STW, various other institutions use the STW for subject indexing (see section 7). In the academic search portal *EconBiz*, provided by the ZBW, a significant portion of its nearly eleven million textual resources is annotated intellectually with descriptors from the STW, according to in-house rules. A large percentage of the full text within the ZBW's digital repository *EconStor* (http://econstor.eu) is also annotated by descriptors taken from the STW.

In recent years, new indexing approaches using the STW have emerged due to digitization, as the number of publications has increased and personnel resources have decreased. Apart from intellectual subject indexing, the STW is also used for → automatic subject indexing at the ZBW.

Efforts to automate the subject indexing process began around the year 2000. After initial experience with commercial systems, the ZBW decided to initiate an individual non-commercial in-house project to develop machine learning solutions for subject indexing. In a research-based project (AutoIndex) undertaken between 2014 and 2018, a ZBW-specific solution for subject indexing was developed based on open source machine learning solutions. This combines several associative and lexical methods in a fusion approach, which leads to improved performance. The solution inputs resource titles and author keywords and generates suggestions for descriptors from the STW that adequately summarize the resource in question. Various rule-based, post-processing routines ensure the quality of the metadata. So far, several data releases have been issued. The ZBW developed a web-based reviewing tool to evaluate the data intellectually (Toepfer and Seifert 2017). Currently under investigation is to what extent neural networks are suitable for optimizing the coordination between the individual methods in the fusion approach.

In addition, the STW is currently integrated within a commercial web-based tool for machine-assisted support of intellectual verbal and classificatory subject indexing for testing purposes. In testing situations, the tool facilitates the process of subject indexing by presenting suggestions from other data sources that already contain subject-specific information for the resource in question. These suggestions are based on the previously mentioned vocabulary mappings to the STW (see section 5). In production scenarios, suggestions will also come from those in-house research-based automated subject indexing solutions as outlined above (Kasprzik 2020).

6.2 Retrieval

The STW supports a variety of classical retrieval functions. The STW web pages can trigger a direct search via the "EB" icons on the page for the appropriate concept; this enables the semantic network of descriptors and the subject category system to create a high-level entry point into the publications of the *EconBiz* portal to explore and identify relevant subjects.

Enhancements to the search engine index improves search functionality within the EB portal. For the STW controlled fields, index entries are produced with the English and German preferred terms, all synonyms of the descriptors, all preferred terms, and synonyms of exact matches from the various mappings to STW descriptors. The data and code for generating these enhancement lists are freely available [6].

This index enhancement results in quick retrieval and works well for large document collections. (EB comprises nearly 11 million documents.) It is also inflexible, so users have very little influence on the outcome. Another search interface was implemented within the smaller collection of scientific papers in the ZBW repository *EconStor* [7]. This search interface uses the "econ-ws" web service to match the query against the full text index of the STW and identifies thesaurus descriptors within the query. This enhances the identified descriptors dynamically with their direct and indirect synonyms upon user request. The search interface also displays narrower and related descriptors for (parts of) the query, for the selection of thematically related search terms, again including synonyms (see Figure 3).
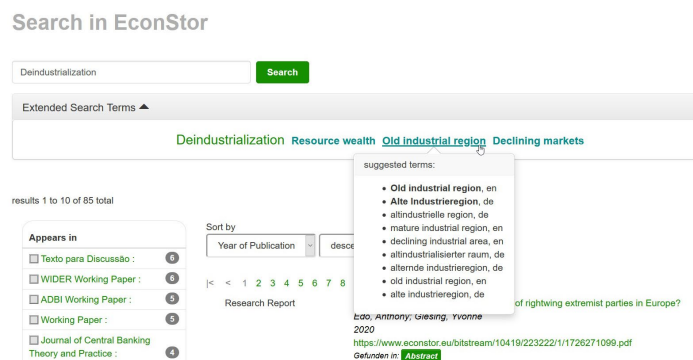

Figure 3: An "Extended Search Terms" section in the *EconStor* search results page offers users a thesaurus-enhanced search for matching query terms. By clicking on the `Deindustrialization` link, the user may enhance the search by all available synonyms, increasing the number of results from 85 to 130. `Old industrial region` is another related concept suggested by the STW, which can be searched with multiple synonyms.

7. Reuse

The number of institutions using the STW is steadily growing. The library of the of the German Institute for Economic Research and the Swiss Economic Archive are two such institutions [8]. The STW's free and easily accessible web version has led to increased usage. The STW's liberal license (since 2014 ODbL) authorizes free use for any purpose, provided the ZBW is attributed and that the resulting publication is openly available under the same conditions. All SKOS files of the thesaurus and its mappings can be downloaded from the web site. The free web service [9] provides an API for use in programs, and a SPARQL endpoint [10] allows unrestricted

experimental queries. Additionally, a MARC-XML dump of the STW from the catalog system can be provided on request; the downloaded data may differ from the web publication and will not include the category system or the English scope notes.

8. Considerations for the future

For many years, the STW was used exclusively for subject indexing. Over time, it was integrated into retrieval processes. Currently, the STW is getting more and more integrated into processes of automated subject indexing. This will result in new requirements for maintenance, further development, and data management of the vocabulary.

In the near future, maintenance and the further development of the vocabulary will be much more → data-driven. Data from additional sources will need to be considered systematically in order to enrich the existing vocabulary with additional synonyms and to get extension proposals for terms that represent new concepts. This includes the filtering, clustering, and systematic evaluation of uncontrolled keywords, which subject indexers always already assign to a document in case they cannot find a suitable descriptor in the STW. It also includes the matching of author keywords with the STW and their technically supported analysis and evaluation. The pilot study and the results of the latest review of the different automatic subject indexing approaches used show the potential of this text type, on the one hand to generate new descriptors from it, on the other hand to achieve good automated subject indexing when using the STW as background knowledge. Data that is already available in machine-readable form needs to be systematically incorporated into the further development of the vocabulary.

Another important task will be to develop an effective indexing suggestion system that takes into account different sources of already existing third-party subject-specific data.

As the STW becomes more integrated into automated subject indexing solutions, enhanced interoperability will be an important goal for the future. As a result of the mapping to *Wikidata*, additional semantics and information about a concept become available, and indirect mapping relationships to concepts in third-party vocabularies can be derived. This leads to methods for pattern recognition in texts that, combined with technologies for evaluating explicit semantics, will improve the results of automated subject indexing solutions.

Endnotes

1. http://zbw.eu/stw.

2. The other two institutions that participated in the cooperation project were GBI Gesellschaft für Betriebswirtschaftliche Information (now: GBI-GENIOS, http://genios.de) and the ifo Institute (https://www.ifo.de/en).

3. http://zbw.eu/version.

4. http://econbiz.eu.

5. https://zbw.eu/stw/mapping/.

6. https://github.com/zbw/sparql-queries.

7. https://www.econstor.eu/.

8. For a selection of institutions using the STW see http://zbw.eu/en/stw-info/uses.

9. http://zbw.eu/beta/econ-ws.

10. https://zbw.eu/beta/sparql-lab/about/#stw.

References

Bibliothek des Instituts für Weltwirtschaft an der Universität Kiel, Zentralbibliothek der Wirtschaftswissenschaften in der Bundesrepublik Deutschland. 1980. *Sachkatalog*. Kiel: Bibliothek des Instituts für Weltwirtschaft an der Universität Kiel.

HWWA-Institut für Wirtschaftsforschung. 1980. *Thesaurus Wirtschaft*. Hamburg: Verlag Weltarchiv GmbH.

ISO (International Organization for Standardization). 2011. *Thesauri for Information Retrieval: Information and Documentation*. Part 1 of *Thesauri and Interoperability with Other Vocabularies: ISO 25964-1*. Geneva: International Organization for Standardization.

Kasprzik, Anna. 2020. "Putting research-based machine learning solutions for subject indexing into practice". In *Proceedings of the Conference on Digital Curation Technologies (Qurator 2020), Berlin, Germany. January 20th to 21st 2020*, eds. Adrian Paschke et. al. http://nbn-resolving.de/urn:nbn:de:0074-2535-7.

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. https://doi.org/10.5771/0943-7444-2016-3-160.

Neubert, Joachim. 2009. "Bringing the 'Thesaurus for Economics' on to the Web of linked data". In *Proceedings of the WWW2009 Workshop on Linked Data on the Web (LDOW 2009), Madrid, Spain, April 20, 2009*. Eds. Christian Bizer, Tom Heath, Tim Berners-Lee and Kingsley Idehen. Aachen, Germany: CEUR Workshop Proceedings, paper 7. http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf.

Neubert, Joachim. 2015. "Leveraging SKOS to trace the overhaul of the STW Thesaurus for Economics". In *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2015), São Paulo, Brazil, 1-4 September 2015*. Published by the Dublin Core Metadata Initiative. http://dcevents.dublincore.org/IntConf/dc-2015/paper/view/332.

Prange, Alexander. 2016. *Evaluierung verschiedener Quellen zur Anreicherung kontrollierter Vokabulare durch Generierung von Vorschlägen*. Master thesis. Kiel: Christian-Albrechts-Universität zu Kiel, Leibniz-Informationszentrum Wirtschaft.

Toepfer, Martin and Christin Seifert. 2017. "Descriptor-invariant fusion fusion

architecture for automatic subject indexing". In *Proceedings of Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society. Washington, D.C.: 31-40.

Version 1.0 published 2021-01-07

Article category: `KOS, specific (domain specific)`

The present article is an updated version of Kempf and Neubert (2016). https://doi.org/10.5771/0943-7444-2016-3-160