Conception of a Workflow for the Semi-automatic Construction of a Thesaurus for the German Printing Industry

Anette Siebenkäs, Bernd Markscheffel

Technische Universität Ilmenau
Fachgebiet Informations- und Wissensmanagement
Helmholtzplatz 3 (Oeconomicum), 98693 Ilmenau, Germany
{anette.siebenkaes, bernd.markscheffel}@tu-ilmenau.de

Abstract

During the BMWI granted project "Print-IT", the need of a thesaurus based uniform and consistent language for the German printing industry became evident. In this paper we introduce a semi-automatic construction approach for such a thesaurus and present a workflow which supports users to generate thesaurus typical information structures from relevant digitalized resources with the help of common IT-tools.

Keywords: Information organization, Information retrieval, Thesauri

1 Introduction

In Germany, the printing industry is largely based on small and medium-sized enterprises (SME) (bvdm 2012) which can benefit from networks to use synergy effects including their target groups. Therefore, a research project named "Print-IT" granted by BMWI was established. The project partners are TU Ilmenau, HTWK Leipzig, SID Leipzig and several SME's of the printing industry. Main aim of this project was to build an integration plat-

In: F. Pehar/C. Schlögl/C. Wolff (Eds.). Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19th—21st May 2015. Glückstadt: Verlag Werner Hülsbusch, pp. 217—229.

form where SME's can provide order- or product-specific goods and services (Eine, Stelzer 2014). Studies describe problems in communication between printing organizations and with the costumers (Canon 2012). The complicated technical terminology used is a result of the complexity and heterogeneity of the printing products and methods. It is necessary to standardize the vocabulary for an effective service bundling. For this purpose, we want to propose the building of a thesaurus for the printing industry. The application of information technologies could bring time efficiency and quality enhancement for the thesaurus construction and maintenance. In particular, we address the following research questions:

- Which information technological tools can be applied?
- How far is the automatic or semi-automatic construction possible?
- Which impact does the kind of source literature have on the workflow?

The development of the workflow is based on literature research (Webster & Watson 2002) and a deductive analysis of literature samples. Basics for thesaurus construction are drawn from the ISO 25964-1 (2011), Broughton (2010), Burkart (2004) and Wersig (1978). Tools for supporting the semi-automatic construction found in research projects in literature were tested as examples (Nohr 2001, Lepsky 2006, Gödert 2012, Gödert et al. 2012).

2 Thesaurus characteristics and system conception

A thesaurus is a "controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms. The purpose of a thesaurus is to guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given subject" (ISO 25964-1 2011: 12).

We use the thesaurus system conception of Wersig (1978: 324 ff.) as an orientation to show what is to be decided before starting the thesaurus construction.

Development approach: The thesaurus for the printing industry has to be a completely new development.

Specification: The subject of the thesaurus is the terminology of the printing industry, which is structured as shown as in fig. 1.

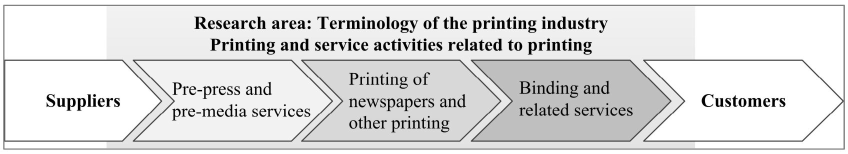

Research area: Terminology of the printing industry
Printing and service activities related to printing
Suppliers
Pre-press and pre-media services
Printing of newspapers and other printing
Binding and related services
Customers

Figure 1. Research area and the printing industry (EU 2008)

Conceptual overview: The extent of the thesaurus depends on the number of descriptors and texts of scope notes. An aspect for planning the information system holding the thesaurus is the required storage space. However, nowadays, a textual database does not represent a storage problem.

Problem analysis for structure: It is common to adapt the structure of the thesaurus to the one of the subject domain. This includes such decisions like the handling of compound and multi-word terms, deciding about pre- or post-coordination (ISO 25964-1 2011: 9 and Broughton 2006: 91), or the handling of synonyms and quasi-synonyms (Broughton 2006: 71 f.).

Selection of a structure-adequately processing: We preferred SKOS (Simple Knowledge Organization System) as output format to provide interoperability. We expected app. 2,800 descriptors, derived from the count of keywords in the "Handbuch der Printmedien" (Kipphan 2000: 1214 ff.).

Development strategies: The construction of a thesaurus can follow different methods. Wersig (1978: 236–239) describes a pragmatic, a systematic and an individual method as a border case of the systematic method. The systematic method starts with establishing the terminological structure followed by filling it with the corresponding terms. The pragmatic method uses strategies and tools of indexing to gain descriptors from current literature sources. It is recommended to combine these methods (Wersig 1978: 236–239).

Potential division of labour and cooperation: The thesaurus management software should support the maintenance of the thesaurus in the future (see ISO 25964-1 2011: 98 ff. and table 1).

Planning manpower, resources and scheduling: The time for building a thesaurus depend on the amount of literature, the quality of literature sources, the used tools and other features.

3 Tools for thesaurus construction and maintenance

Thesaurus management software: For the construction and maintenance, a thesaurus management software (TMS) is needed. Derived from the electronic support described by Burkart (2004: 153 ff.), we present the following requirements for a TMS relating to Wersig (1978: 235 f.) in table 1.

Table 1. TMS advantages and requirements

Advantages	Requirements
Import of vocabulary	Import interfaces
Check for plausibility at input, check for errors at import	Thesaurus specific check and control procedures
Establish back links automatically, automatic correction in all records of one descriptor, simple-to-use features to change structure by splitting or merging	Program functionalities to support and check back (cross references) links, import/export and merge functionalities
Presentation of the thesaurus alphabetically, as a list or a graphic	Variable output configuration, select/sort/print properties Additional: Output as SKOS file
Hypertext structure for easy navigating, integration in web application	Web application module or export format as Hypertext or XML

We used MIDOS*Thesaurus* from PROGRIS for this task. This TMS fulfils the requirements, supports versatile input- and output features, and is easy to use. For further information we refer to Gödert et al. (2012: 382 ff.).

General Tools: For computer-aided processing, the sources have to be in digital form. With the help of scanner and optical character recognition software (OCR), an editable document can be created. Misinterpretation of the OCR can be corrected manually or with the aid of an indexing system (Lepsky 2006: 173). For further editing of the digital sources to extract the information for the thesaurus, text editors or word processors like Notepad®, Microsoft Word®, and so on, are useful. The requested import format for an indexing system or the TMS can be realized with an editor, too.

Indexing tools: For the pragmatic method of the thesaurus construction, an indexing software tool (IDXS) is needed. The indexing process is taken in two steps (Nohr 2001: 15 and ISO 25964-1 2011: 5):

1. content analysis: Conceptual determination of the content of a document (identify the concepts),
2. content representation: Attachment of terms of a controlled indexing language to summarize the content.

In this work, we used *content analysis* for generating the vocabulary. We focused on indexing by extraction methods. As an example, we chose the open source software LINGO, which supports statistical and linguistic approaches. The indexing process can be controlled by configuration script files. The dictionaries can easily be edited and updated (Lepsky, Vorhauer 2006). For applications and configuration samples of the LINGO IDXS see: Gödert et al. (2012: 269 ff.), Gödert (2012: 5 ff.), Lepsky (2006: 174 ff.).

4 Steps of thesaurus construction in the workflow

Step 1: Selection and analysis of sources

We selected an example literature source which provides a good and correct overview on the subject domain: the “Handbuch der Printmedien”. The authors of the book did extensive research and interviews with experts of the printing industry and their customers. The book is build very systematically. A digitized edition in pdf-files is also available (Kipphan 2000: VI & VII).

Apart from the content, a closer look shows differences in information density within *one* source (Lohmann 2000: 17). Parts of high information density in books are table of contents, abstracts or indexes. Glossaries and dictionaries provide another information density. Considering further computer-aided processing, it is useful to divide sources into their components for bundling. These now called “*source components*” will be classified as followed (Tab. 2).

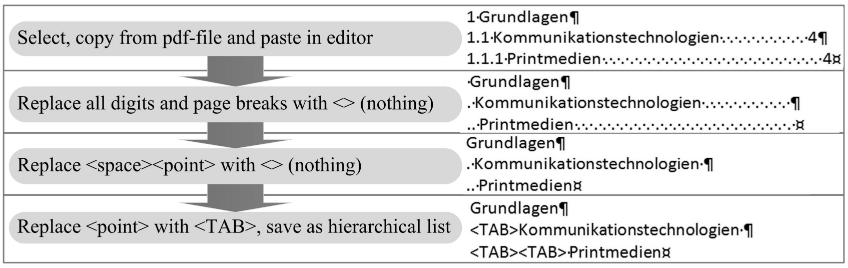
Table 2. Definition of source components and their use for the thesaurus

Source component	Definition	Use for thesaurus and special aspects
Contents (table of contents)	"a list of the chapters or sections given at the front of the book or periodical" (Oxford 2010: Table of contents)	compound terms, hierarchical overview of the terminology, nouns in singular or plural, numbers and other junk
Index (back-of-the-book index, index list)	"an alphabetical list of names, subjects, etc. with reference to the pages, on which they are mentioned" (Oxford 2010: Index)	Descriptors, compounds, hyphen plus adverb for narrower term, references to synonyms, base forms, but complex structure,
Glossary (dictionary entries)	"an alphabetical list of words relating to a specific subject, text, or dialect, with explanations; a brief dictionary" (Oxford 2010: Glossary)	Descriptors, scope notes, synonyms, hierarchical and semantic relations, keyword in base form, formatted in Italic
Abbreviations (list of abbreviations)	"an alphabetical list of shortened forms and their corresponding long forms of words or phrases" (Oxford 2010: Abbreviation).	Descriptors and its synonyms, multi-word phrases, compounds, short form/long form in base form, no hierarchies
List	"a number of connected items or names written or printed consecutively, typically one below the other" (Oxford 2010: List).	List of descriptors for a special topic like "products", terms in base form or plural, but no relations
Text (body text, continuous text)	all text that is not in lists, indexes, figures or tables	Compounds, descriptors, filling and stop words, flectional terms
Graphic (figure, picture, diagram)	Here, a figure in a book is labelled. We use the term "graphic" for its digital data characteristic.	Descriptors, hierarchies and other relations can clearly be shown in a diagram

Step 2: Transformation

The vocabulary is collected by extracting the contents of the source components to a file, making changes until it fits into the TMS input filter with preservation of the relation information. Then, the edited components are imported into the TMS and saved as part thesauri. So, the vocabulary collection persists of several part thesauri. The top terms can be defined at all times, for the TMS provides the structural transfer and inheritance.

We developed editing methods for each component. These are represented in the workflow as "method X" (see example in fig. 2). In our research, we tested several samples from Kipphan (2000) and other sources like web shops as components and recorded the steps with screen shots and verbal descriptions. For TMS input, we selected two formats, the “hierarchical list” and the “thesaurus list” (Tab. 3).



Select, copy from pdf-file and paste in editor

```
1·Grundlagen¶
1.1·Kommunikationstechnologien·.·.·.·.·.·.·.·.·.·4¶
1.1.1·Printmedien·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·4¤
```

Replace all digits and page breaks with <> (nothing)

```
·Grundlagen¶
.·Kommunikationstechnologien·.·.·.·.·.·.·.·.·.·¶
..·Printmedien·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·.·¤
```

Replace <space><point> with <> (nothing)

```
Grundlagen¶
.·Kommunikationstechnologien·¶
..·Printmedien¤
```

Replace <point> with <TAB>, save as hierarchical list

```
Grundlagen¶
<TAB>Kommunikationstechnologien·¶
<TAB><TAB>·Printmedien¤
```

Figure 2. Sample method for editing component “Contents” (right) with MS Word®.

Table 3. Components and TMS input format as the aim of editing

Component	Input format	Example
Contents	hierarchical list	Descriptor (= keyword) <CRLF> <tab> subordinate term (level 1) <CRLF> <tab><tab> subordinate term (level 2) <CRLF><CRLF>
List		Descriptor (= upper term)<CRLF> <tab> subordinate term 1 <CRLF> <tab> subordinate term 2 <CRLF><CRLF>
Index	thesaurus list	Descriptor (= keyword) <CRLF> <tab>NT<tab> subordinate term <CRLF> <tab>RT<tab> related term <CRLF> <tab>UT<tab>generic term <CRLF><CRLF>
Glossary		Descriptor (= keyword) <CRLF> <tab>SN<tab> scope note <CRLF> <tab>UF<tab>synonym <CRLF><CRLF>
Abbreviations		Descriptor (= long form) <CRLF> <tab>UF<tab> short form of term <CRLF> <tab>UF<tab>synonym <CRLF><CRLF>
Text	input file for IDXS, index file	Descriptor 1 <CRLF> Descriptor 2 <CRLF>…
Graphic	Manually input data in TMS	Descriptor 1 <CRLF> Descriptor 2 <CRLF>…

Step 3: Control

We performed the vocabulary control mostly by *keeping* the already existing information about relations. After importing and joining the part thesauri in the TMS, an intellectual control has to follow. The IDXS needs updated dictionaries, so these have to be updated with every new descriptor list.

Step 4–6: Presentation, test and maintenance

With MIDOS*Thesaurus*, several formats can be selected, configured, printed, and exported as XML, HTML, txt, rdf with header, footer, and column formatting. A web application can be built and configured with style sheets. The "completeness" of the thesaurus can be tested with indexing an amount of continuing text from related sources. If the dictionaries are all updated, no unknown word should be found by processing the IDXS. The TMS supports the maintenance of the thesaurus. Files can be merged or parted. The TMS prevents duplicating descriptors at import. Properties can be passed to narrower descriptors (Gödert et al. 2012: 382ff.).

In table 4, we give a short overview over the main workflow steps.

Table 4. Steps of the thesaurus construction

Step	Approach / Method	Realization / Tool
Selection and analysis of sources	Select and analyse sources, prepare source components, digitize non-digital sources by scanning and OCR, correct errors manually or with IDXS	Semi-automatic with scanner, OCR, IDXS, editor
Trans-formation	Define top terms, collect vocabulary with preservation of the information about relations, extract and select potential descriptors by automatic indexing, create/update dictionaries	Intellectual, semi-automatic with editor, IDXS, IDXS dictionaries
Control	Import into TMS, control and error corrections	Semi-automatic with TMS
Presentation	Output in XML, SKOS, Web application	Automatic with TMS
Test	Test of "completeness" by indexing new text	Automatic with IDXS
Maintenance	Further maintenance and enhancement	Semi-automatic (TMS)

In this paper, we can only give a condensed presentation. Fig. 3 to 4 show the main part of the workflow as eEPC (extended event-driven process chain).

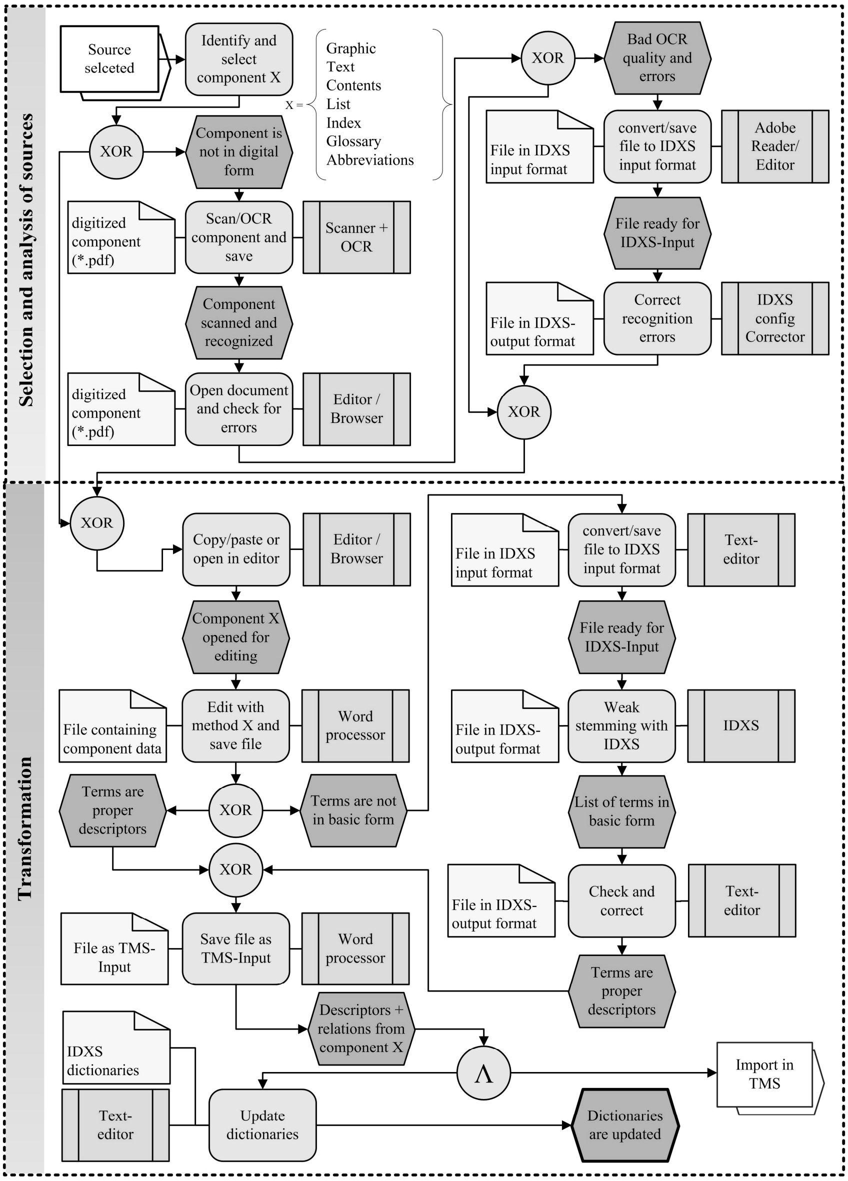

Selection and analysis of sources
Source selceted
Identify and select component X
X =
Graphic
Text
Contents
List
Index
Glossary
Abbreviations
XOR
Component is not in digital form
digitized component (*.pdf)
Scan/OCR component and save
Scanner + OCR
Component scanned and recognized
digitized component (*.pdf)
Open document and check for errors
Editor / Browser
XOR
Bad OCR quality and errors
File in IDXS input format
convert/save file to IDXS input format
Adobe Reader/ Editor
File ready for IDXS-Input
File in IDXS-output format
Correct recognition errors
IDXS config Corrector
XOR
Transformation
XOR
Copy/paste or open in editor
Editor / Browser
Component X opened for editing
File containing component data
Edit with method X and save file
Word processor
Terms are proper descriptors
XOR
Terms are not in basic form
XOR
File as TMS-Input
Save file as TMS-Input
Word processor
Descriptors + relations from component X
IDXS dictionaries
Text-editor
Update dictionaries
Λ
Import in TMS
Dictionaries are updated
File in IDXS input format
convert/save file to IDXS input format
Text-editor
File ready for IDXS-Input
File in IDXS-output format
Weak stemming with IDXS
IDXS
List of terms in basic form
File in IDXS-output format
Check and correct
Text-editor
Terms are proper descriptors

Figure 3. Step 1 Selection and analysis of sources and step 2 transformation

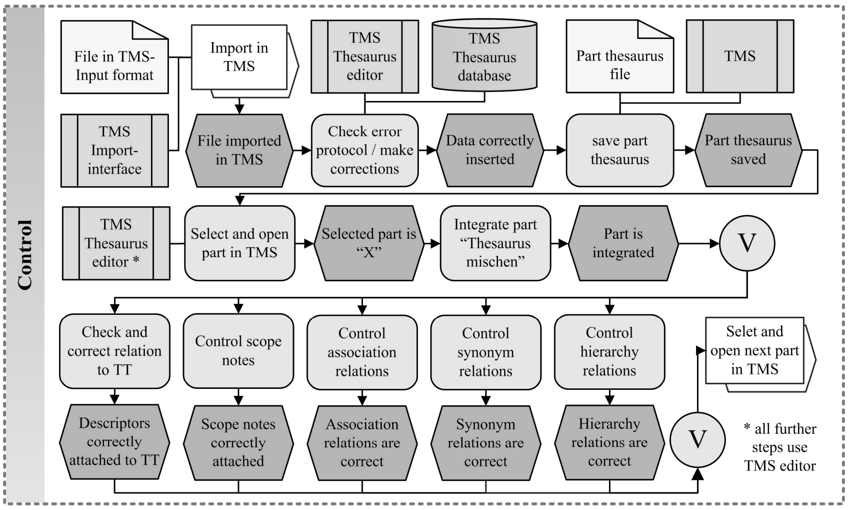

Control
File in TMS-Input format
Import in TMS
TMS Thesaurus editor
TMS Thesaurus database
Part thesaurus file
TMS
TMS Import-interface
File imported in TMS
Check error protocol / make corrections
Data correctly inserted
save part thesaurus
Part thesaurus saved
TMS Thesaurus editor *
Select and open part in TMS
Selected part is "X"
Integrate part "Thesaurus mischen"
Part is integrated
V
Check and correct relation to TT
Control scope notes
Control association relations
Control synonym relations
Control hierarchy relations
Selet and open next part in TMS
Descriptors correctly attached to TT
Scope notes correctly attached
Association relations are correct
Synonym relations are correct
Hierarchy relations are correct
V
* all further steps use TMS editor

Figure 4. Step 3 control

6 Discussion and conclusions

In this work, a workflow for the semi-computer aided construction of a thesaurus is presented as an overview. The developed workflows are tested with the use of common tools for text processing and editing, with the LINGO indexing, and the MIDOS*Thesaurus* management software systems. We generated descriptors by indexing, editing and importing parts of the reference book "Handbuch der Printmedien". The components were imported in the TMS one by one and saved as separate data files. At the end, we joined the part thesauri by merging as a test and tested several output formats, too.

- *Development of the workflow:* We did not find similar workflow in literature. The goal of most of the research projects found was the creation of metadata by automatic indexing approaches, development of appropriate tools, improvement of library software and retrieval tests.
- *Influence of the kind of source:* We found, not only the source but the component of the source is important for the workflow. We characterize the source components by content and composition. The components hold information in differing structure and concentration about the concepts and their relations. For use in the thesaurus, the sources have to be digi-

tized, with correct technical content, and show the most of the structure of the vocabulary. The transforming process should save the relations.

- *Possibility of automatic or semi-automatic thesaurus construction:* Many steps of the workflow are only intellectually to handle. A good support and time sparing can be achieved by implementing little macros for iterative processes. The use of an indexing system to evaluate the thesaurus relations was shown. We found the assistance of a modern word processor or editor very useful. With well-founded knowledge of the capacities of the office tools, editing the source components is not very time consuming. We found a disadvantage in the dependence on the quality and correctness of the sources. The computer can support the editing but cannot replace the "necessary special knowledge at the thesaurus work" (Wersig 1978: 239 ff.).

  For automatic thesaurus construction with methods described in Salton, McGill (1987: 88 ff.), there must be a sufficient and representative document collection to prove and relations found with these approaches are only approximate and therefore fuzzy, too. We can confirm that the thesaurus construction is an iterative process: Terms and relations from various source components can be inserted continually, depending on the TMS functionalities.
- *Use of IT-tools:* We took care to use standard software. The indexing system has many configuration features. Therefore, it was used in many research projects, like the commercial software MIDOS*Thesaurus,* too. The use of the LINGO indexing system was not successful for all our purposes. If the dictionaries are updated continuously, new terms can be found by indexing new text corpora. But inserting these words in the thesaurus structure is difficult without information about the hierarchical relations. The detection of multiword phrases should be an approach for finding relations with approximation methods (so-called automatic generated concept associations). We did not have enough vocabulary to check these approaches.
- *Further development of the thesaurus for the printing industry:* A basic structure of the thesaurus is established with approximately 500 descriptors and several relations between them. In extracting, editing and importing further sources the thesaurus can be filled and checked as described below. As a general approach, the workflow is useful for building other thesauri, too. The method of editing depends on the kind of component and its *form* of content, not on the meaning of the content.

Acknowledgments

We like to thank PROGRIS for the MIDOS*Thesaurus* licence use. All translations of citations from German books were done by the authors.

References

Broughton, V. (2006). *Essential thesaurus construction*. London: Facet Publishing.

Burkart, M. (2004). Thesaurus. In: Rainer Kuhlen, Thomas, Dietmar Strauch (Eds.). *Grundlagen der praktischen Information und Dokumentation.* München: Saur, pp. 141 ff.

bvdm (2012). Bundesverband Druck und Medien, Taschenstatistik 2012. http://www.bvdm-online.de/aktuelles/downloads.php?Action=Download&FileID=356 <2.5.2013>.

Canon (2012). Canon Deutschland: Studie zeigt Chancen für Druckereien zur Stärkung ihrer Kundenbindung. http://www.canon.de/About_Us/Press_Centre/Press_Releases/Business_Solutions_News/1H12/Canon_research_highlights_opportunities_printers.aspx <23.5.2013>.

Eine, S.; Stelzer, D. (2014). Erstellung eines Metamodells zur Entwicklung einer Kollaborationsplattform für die Druckbranche. In: *Ilmenauer Beiträge zur Wirtschaftsinformatik.* http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-29491/ilm1-2014200034.pdf <29.12.2014>.

EU (2008). *Statistical Classification of Economic Activities in the European Community*. http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC&IntCurrentPage=1 <29.12.2014>.

Gödert, W. (2012). Detecting multiword phrases in mathematical text corpora. http://arxiv.org/ftp/arxiv/papers/1210/1210.0852.pdf <18.1.2014>.

Gödert, W.; Lepsky, K.; Nagelschmidt, M. (2012). *Informationserschließung und Automatisches Indexieren – Ein Lehr- und Arbeitsbuch*. Heidelberg: Springer.

ISO 25964-1 (2013). Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval.

Kipphan, H. (2000). *Handbuch der Printmedien: Technologien und Produktionsverfahren.* CD-ROM. Berlin: Springer electronic media.

Lepsky, K. (2006). Automatisches Indexieren des Reallexikons zur Deutschen Kunstgeschichte. In: *Information und Sprache – Festschrift für Harald H. Zimmermann*, München: Sauer, pp. 169–178.

Lepsky, K.; Vorhauer, J. (2006). LINGO – ein open source System für die Automatische Indexierung des Deutschen. http://blackwinter.de/da/da/lit/lepsky_-_lingo_-_ein_open_source_system_f%C3%BCr_die_automatische_indexierung_des_deutschen.pdf <17.1.2014>.

Lohmann, H. (2000). KASCADE: Dokumentanreicherung und automatische Inhaltserschließung. Projektbericht und Ergebnisse des Retrievaltests. Düsseldorf: Universitäts- und Landesbibliothek.

Nohr, H. (2001). *Automatische Indexierung: Einführung in betriebliche Verfahren, Systeme und Anwendungen.* Potsdam: Verlag für Berlin-Brandenburg.

Oxford (2010). *The New Oxford Dictionary of English*, Kindle edition, 2nd edition. Oxford: University Press.

Salton, G.; McGill, M. (1987). *Information Retrieval – Grundlegendes für Informationswissenschaftler*. Hamburg, New York: McGraw-Hill.

Webster, J.; Watson, T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. In: *MIS Quarterly*, 2, xiii-xxiii.

Wersig, G. (1978). *Thesaurus-Leitfaden. Eine Einführung in das Thesaurus-Prinzip in Theorie und Praxis*. München: Saur.