Tagging and Automation – Challenges and Opportunities for Academic Libraries

Kai Eckert
Computer Science Institute, University of Mannheim, Mannheim, Germany

Christian Hänger and Christof Niemann
Department of Digital Services, University of Mannheim Library, Mannheim, Germany

Abstract

Purpose - comparing and examining the quality of the results of tagging, intellectual and automated indexing processes.

Design/methodology/approach - analysis and graphical representation of annotation sets using the software "Semtinel".

Findings – a combination of tagging, intellectual and automatic indexing is probably best suited to shape the annotation of literature more efficiently without compromising quality.

Research limitations/implications – exploratory study on the base of three journals.

Originality/value – the paper presents the open source software Semtinel offering a highly optimized toolbox for analysing thesauri and classifications.

Keywords - indexing, classification, tagging, information retrieval, visualization technique, comparative study

Paper type – research paper

1. 1. Introduction

While the amount of scholarly information is growing rapidly, academic libraries have to face the fact that researchers have ongoing problems in finding the relevant information they are searching for. Using traditional OPACs, they often do not find electronic information such as eBooks or articles in eJournals. Conventional integrated library systems do not have the necessary categories, making the storage and presentation of non-book materials difficult. Google Scholar supplies too many hits without any relevance to the researcher's field of interest, because the metadata of electronic texts such as eBooks or articles in eJournals have not been annotated by information specialists. A significant example is the collection of a total of 250,000 digitized books available in German academic libraries through the National Licenses Programme (Nationallizenzen) funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). The National Licenses Programme has, for example, licensed the collections "Early English Books Online" and "Eighteenth Century Collections Online". Compared to the holdings of Mannheim University Library, which offers 2,200,000 books to its clients, the nationally licensed collections add another 12 percent to the on-campus holdings. The usability of such sizeable additional content depends heavily on the implementation of integrative search engines as well as on the efficient exploitation of the collection's contents.

What can we do? Mannheim University Library is considering the introduction of a comprehensive search solution containing all available electronic and printed information (similar to an extended "Google Scholar" for the students and researchers of the university). The Ex Libris Group offers the research system "Primo" based on the search engine "Lucene", which enables the presentation of heterogeneous metadata. Primo facilitates the integration of eBooks, the nationally licensed collections as well as the annotating and tagging of individual records.

But simply offering such a comprehensive search solution to students and researchers does not solve the problem of the lack of indexing of most metadata. It is therefore necessary to search for alternative methods of indexing. One possibility is "collaborative tagging" (also known as

folksonomy, social classification and social indexing among other terms), which indicates the practice and method of collaboratively creating and managing tags to annotate and categorize content. In contrast to traditional subject indexing, metadata is not only generated by experts but also by creators and consumers of the content itself. Usually, freely chosen keywords are applied instead of a controlled vocabulary. The most popular applications based on collaborative tagging are Flickr [i] for storing photos or Del.icio.us [ii] for storing websites. CiteULike [iii], Connotea [iv] and BibSonomy [v] are bookmarking services for academic purposes, organizing individual and common access to scientific information. The idea is that customers tag items which are part of digital collections such as the electronic texts of the above-mentioned "Early English Books Online".

An alternative is automated document indexing. Currently, Mannheim University Library and the department of Computer Sciences of the University of Mannheim are researching methods of automated indexing using the Collexis search engine provided by the German company SyynX. The search phrase is indexed using one or more thesauri as well as free text. In the next step the result is matched with available document sources. All results are shown in order of relevance and accompanied by additional metadata. The documents are taken from journals provided by Elsevier in the fields of Economics and the Social Sciences. For our experiments we used the German Thesaurus for Economics ("Standard Thesaurus Wirtschaft", STW), which is commonly used in Germany for indexing economic literature. This project has been financed by the DFG and is called „Automated indexing and semantic search applications for economic journal articles".

2. 2. State of the art

We would like to give a review of the research dealing with the recent phenomenon of "tagging". Most scholars consider the advantages and disadvantages of this kind of indexing. All discussions start with the conventional way of annotating documents using controlled vocabularies from thesauri or classification systems. When a new subject heading like "Web 2.0" evolves, librarians have to integrate this new term in the existing thesauri or classification systems. This process is often handled in

a very conservative manner, as the indexers wait to see whether a new term will gain more importance or not. Their aim is to keep all parts of the system in balance regarding their size and relevance. Consequently, our example "Web 2.0" has not yet been included in the Regensburg Union Classification scheme (Regensburger Verbundklassifikation, RVK). The RVK has been developed by librarians from the University of Regensburg, Germany, and is utilized by about 20 other German university libraries.

In contrast to that, one main advantage of collaborative tagging is the absence of delay between the publishing of a document and its annotation, because a controlled vocabulary is neither necessary nor used (Mai, 2006, p. 17).

In addition, thesauri and classifications often represent the scholarly paradigms of their date of origin. For example, the classification used by Bielefeld University Library was created in the late 1960s. Its main feature is a strong focus on economic and social aspects within the historical classes – an approach typical for the research interests of historians at that time. User generated annotations do not have this problem, because they represent current perspectives as well as the thematic landscape of publications at a given moment. They can follow changes of interest within subject areas dynamically (Quintarelli, 2005).

On the other hand, the lack of controlled vocabularies is also the biggest disadvantage of tagging. Indexing with free vocabulary will result in ambiguous terms using synonyms or homonyms in different contexts. Take for example a search for the computer language Python, which will also yield hits including the snake or the ancient potter. Abandoning indexing by librarians will have negative consequences for the quality of information retrieval using library search tools (Guy and Tonkin, 2006; Gordon-Murnane, 2006).

3. 3. Related Work

In the following section, we give a short overview of recent efforts concerning the handling, analysis and integration of tagging results:

Heymann and Garcia-Molina discovered a simple but remarkably effective algorithm for converting a large corpus of tags (annotating objects in a tagging system) into a navigable hierarchical taxonomy of tags (Heymann

and Garcia-Molina, 2006). The algorithm leverages notions of similarity and generality that are present in the user generated content. Based on the similarity to certain nodes, the tags are placed within the hierarchical system.

Other authors have investigated the frequency scale of tags: usually only few tags are chosen by many users to describe a given article (Vander Wal, 2005; Shirky, 2005). A graph containing the number of the tags annotated to a resource on the x-axis and the rank of a tag on the y-axis performs a so called long tail.

Peters and Stock want to solve some of the problems of tags (e.g. lack of precision) by introducing methods of Natural Language Processing (NLP) (Peters and Stock, 2008, p. 84). In their opinion tags should be normalized/standardized by using thesauri or lexica and after this process the user will choose the term he wants to tag. Additionally, they present criteria for tagged documents to create a relevance ranking from tag distribution, for example.

Heckner, Mühlbacher and Wolff carried out an empirical study of tagging behaviour in the scholarly annotation system Connotea and selected 500 tagged articles covering information and computer technology (Heckner *et al.,* 2008, p. 15). They set up a model for linguistic and functional aspects of tag usage and the relationship between tags and a document's full text. Their results describe the typical tag as a single-order noun, taken from the title of the article and directly related to the subject.

Finally, Razikin et al. investigated the effectiveness of tags as resource descriptors, determined through the use of text categorisation using Support Vector Machines (Razikin *et al.,* 2008, p. 59). For this, they randomly collected 100 tags and 20,210 documents. Their results were ambivalent: some tags were found to be good descriptors while others were not. "Given that tags are created for a variety of purposes, the use of tags to search for relevant documents must therefore be treated with care".

4. 4. Analysis of the structure and the quality of tags using the "Semtinel"-software

As a result of the considerations discussed above, we can conclude that we have to control the quality of tags if we want to use them for the appropriate exploitation of resources. In the following sections we will describe a method for investigating the structure and the coherence of tags to facilitate this control. We will compare the quality of automatic and user-based annotation to that of indexing done by librarians. This procedure is part of the tagging project mentioned earlier, which is financed by the DFG. The intention of the project is to provide a reference for document annotation – whether automated or created by user and/or librarian tagging.

4.1 Dataset

The data for this exploratory investigation consists of 372 articles included in three economic journals published by Elsevier:

- Journal of Financial Economics (ISSN: 0304-405X),
- Journal of Accounting and Economics (ISSN: 0165-4101) and
- Journal of Health Economics (ISSN: 0167-6296).

Every instance (article) in the dataset contains the name(s) of the author(s), the title of the article as well as a short abstract. Every article was annotated by librarians (1547 tags), users (591 tags) and through automatic exploitation (4135 tags). All of the annotated tags exclusively derive from the German Standard Thesaurus for Economics (STW).

4.2 Method

Our method is based on the free and open-source software Semtinel [vi], which is currently being developed as part of the same DFG-project concerning automated indexing (Eckert, 2007; Eckert *et al*., 2007, 2008a, 2008b, 2008c). Semtinel provides a highly optimised toolbox with various statistical analysis methods, as well as the possibility to get an in-depth view on concrete annotation results. It offers IC Difference Analysis, a

thesaurus evaluation based on a combination of statistical measures and appropriate visualization techniques. The basic idea is the comparison of the information content (Resnik, 1995) of a given concept calculated for two different sources. The results of this analysis for the whole thesaurus hierarchy are visualized using the treemap algorithm, as presented by Shneiderman (Shneiderman, 1992). This visualization makes it easy for the expert to browse the whole thesaurus and examine interesting parts in detail. Each concept of the thesaurus is represented by a rectangle whose dimension indicates the number of its subnodes. The more sub-nodes, the less specific is a concept while including other concepts.

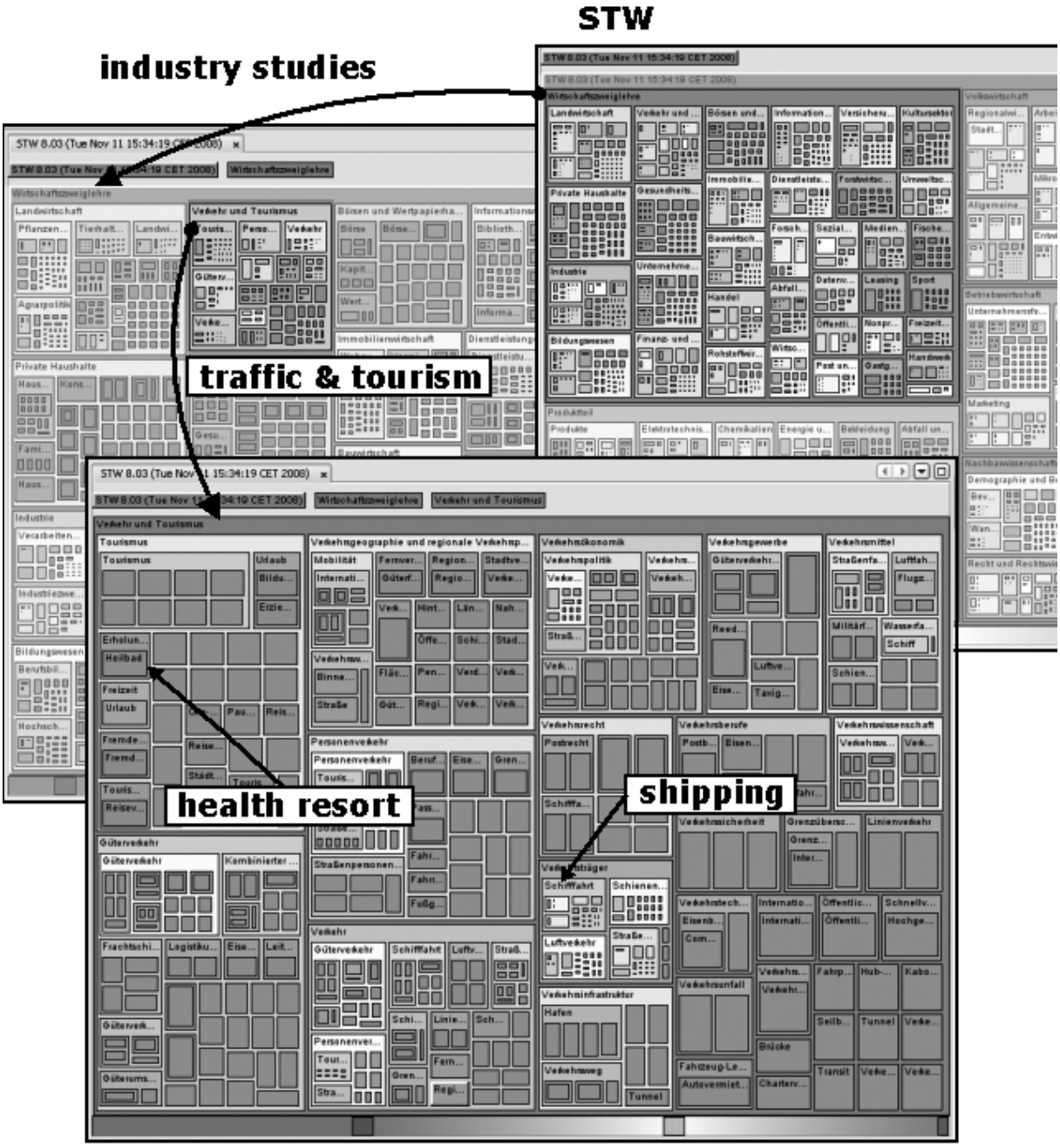

Figure 1. Every rectangle represents a concept. A zoom into its sub-concepts is provided by double clicking, while a simple click gives further information and defines a selection.

Figure 1 gives you an idea of the treemap visualization technique: the first level of the STW contains the concept „Wirtschaftszweiglehre" („economy branches"); the second level within the "economy branches" contains the concept "traffic & tourism". The third level already contains more concrete concepts like "health resort" or "shipping". Double-clicking on a concept allows an exploration of the classification without losing the overview of the relationships between the individual concepts. The deeper we browse, the more specific the concepts.

We can also differentiate between concepts annotated too often or too infrequently respectively through the red and blue coloured gradation of the rectangles: blue indicates too little, red too much usage of a concept, usually compared to another annotation source that serves as a reference. It is also possible to analyse only one annotation source by means of a heuristic approach, which calculates an expected value for the given concept based on the notion of the intrinsic information content, as presented by Seco (Seco *et al.*, 2004). The intrinsic information content depends on the position of a concept in the thesaurus hierarchy. As a rough guide, the deeper a concept resides in the hierarchy, the more specialized it should be and the higher is its expected information content.

4.3 Experiments

In our experiments, we explored our datasets in two steps: first, we had a closer look at the librarians' annotations alone, using the above mentioned heuristic prediction. Peculiarities of the annotations as well as deficiencies in the concept scheme could be found in this step.

Second, we identified the differences between user tagged annotations and automatically assigned annotations by comparing them directly.

Librarians' Annotations

This dataset contains 1547 valid keywords according to the STW concept scheme. If we compare their frequencies with the expected values, we get the treemap shown in Figure 2.

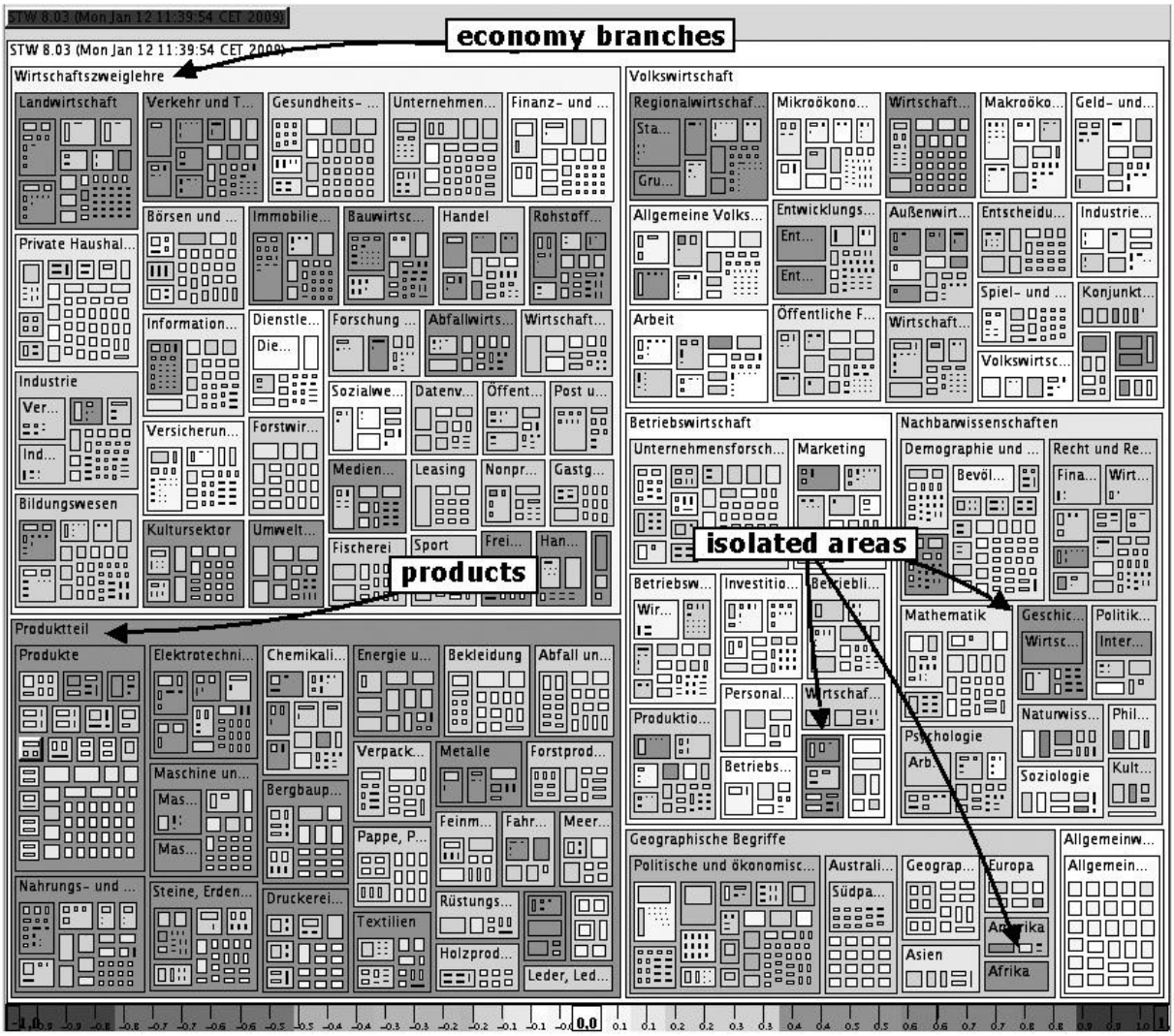

Figure 2. IC Difference Analysis, Librarians' Annotations compared to expectation.

The features that are immediately striking are the blue area in the lower left part ("products"), the heterogeneous impression of the area top left ("economy branches") and the blue isolated areas in the right half of the screen ("Africa", "history", etc.).

As mentioned above, the concept "products" appears to be underrepresented. This is hardly surprising given the nature of the

journals comprising our dataset, where no products in the sense of the thesaurus (such as textiles, chemicals etc.) are mentioned. The keyword "business administration" is an example of a relatively homogeneous concept. A closer look provides the screenshot shown in Figure 3.

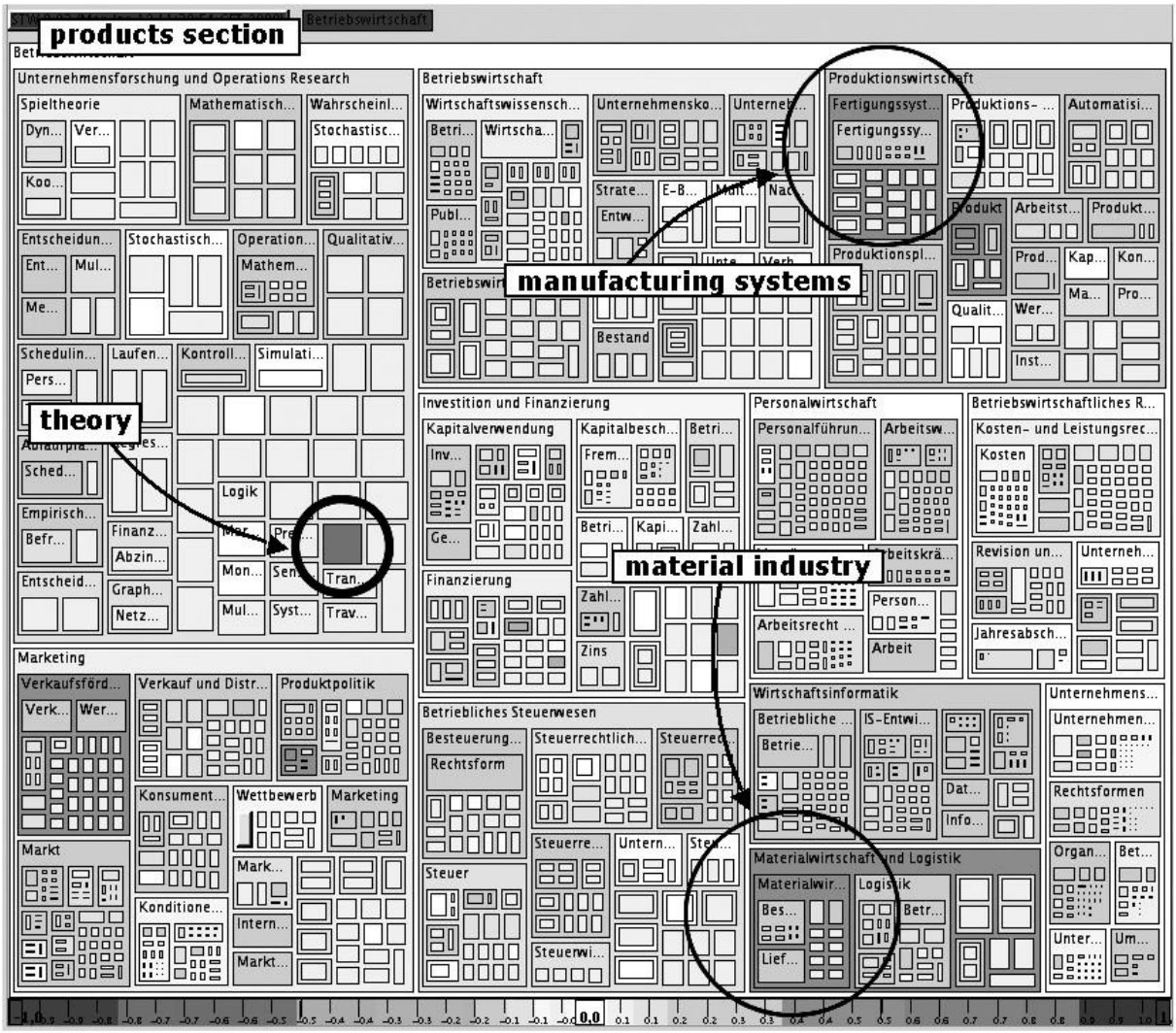

Figure 3. A closer view of the "business" concept.

As expected and corresponding to the "products" section, the concepts of „manufacturing systems", „product" or „material industry" are underrepresented, while one specific rectangle on the left is coloured in deep red. It is the general keyword "theory", which is used by the librarians to annotate theoretical approaches in the given articles and which in our dataset adds up to 171 articles (about 46%). The keyword "equity offering" in the finance section (the small green rectangle selected

in Figure 4) is also worthy of note. Its disproportionate frequency can again be explained by the thematic orientation of the annotated journals.

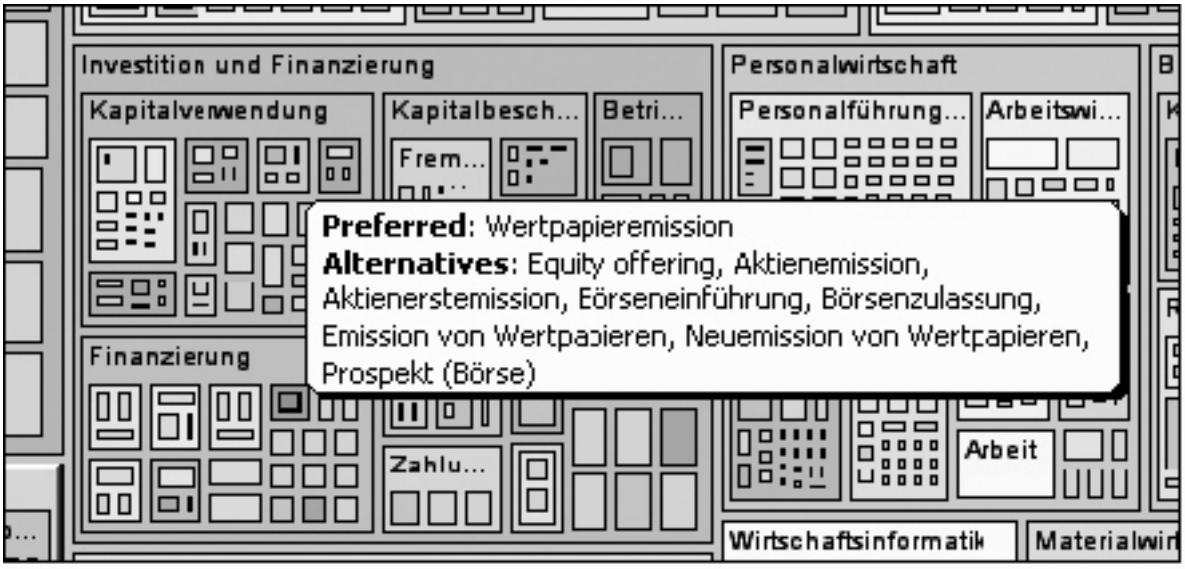

Figure 4. The selection of a rectangle provides further information about a concept.

The area in the thesaurus showing this thematic bias best is the above-mentioned concept "economy branches". The heterogeneous picture in the overview shows that the sub-concepts perfectly match the prediction in their sum of annotations. Nevertheless, a closer look reveals that the distribution within the economy branches is not well balanced and reflects exactly the thematic foci of the journals comprising our dataset: "health care", "finance and banking" "insurance" and "stock exchange" are dominant, while concepts such as "agriculture", "transport" or "feedstock industry" are practically non-existent.

Automatic annotations

In a previous publication (Eckert *et al.*, 2007), we examined the quality of automatic indexing by comparing the annotations with the heuristic prediction. As we worked on the same dataset, we retraced some of the findings. However, in this section, we will directly compare the results of the automatic indexing system with the annotations made by the librarians. We used the thesaurus-based Collexis search engine [vii] as the

indexing system, which led to 4135 automatically assigned keywords. As input we used the abstracts attached to every article in the dataset.

The sheer number of keywords alone lets us suspect that this form of annotation covers more areas than the ones previously presented. This peculiarity is also reflected in the IC-Difference analysis. The overall view reveals that the “products” section is used now for annotations. Looking at this concept in detail, it quickly becomes apparent where this supposedly "new perspective" on the database of financial journals originates from (see Figure 5).

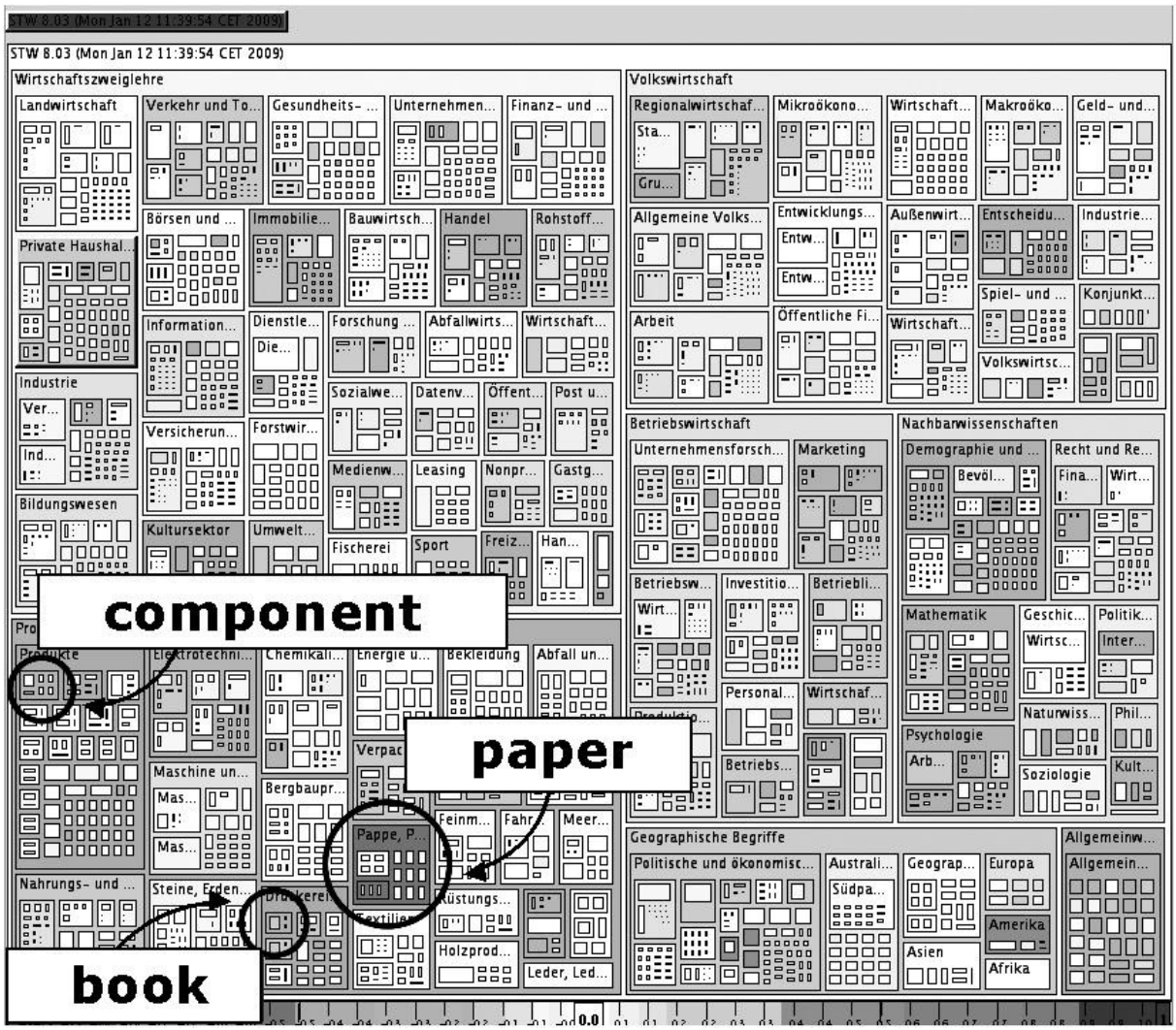

Figure 5. The concepts “component”, “paper” and “book” were annotated misleadingly by automatic indexing. The method is susceptible for failures ascribed to the misunderstanding of synonyms. On the other side, geographic locations are visibly underrepresented in the automatic annotations.

The erroneous annotations result from the missing word sense disambiguation for certain ambiguous terms - a problem that is partly enforced by the fact that the STW is mainly a German thesaurus and only one English term is available for each concept. These terms are in many cases more ambiguous than their German counterparts. For example, the concept “Baufertigteil”, which is a pre-fabricated section of a building, has the ambiguous term “component” assigned as its English equivalent.

Consequently, the “Baufertigteil” is assigned wrongly in every single case where different kinds of abstract components are mentioned in the articles.

Similarly, the term "paper" can mean a treatise of lesser extent as well as the product gained from lumber, which in this case was incorrectly annotated. Such ambivalent concepts that cause problems in an automatic indexing process can easily be identified using Semtinel. A similar picture emerges for instance around the concept of "Analysis", which belongs to the sections "neighboring sciences" and "mathematics". Evidently articles annotated with “analysis” mainly cover the field of *scientific analyses* and only rarely *mathematical analysis*. For the automatic keyword detection this distinction does not exist.

Another important notion is the lack of “geographic locations” assigned by the automatic indexing system. The reason for this is that geographical terms virtually do not appear verbally in the abstracts used for the automatic indexing process. In the related articles, which are mainly written for the domestic market, there seems to be no necessity to mention the name of the country explicitly. However, with annotations assigned for foreign users as well as a conscientious librarian, geographical information will surely be part of the keyword chain. Hence, the terms “USA”/“America” dominate our treemap in the sub-categories “NATO countries”, “industrialized countries”, etc. The concept “USA” generates the deep blue appearance of the whole area almost entirely on its own. In comparison, “Europe”, “Asia” or “Africa” hardly occurs.

Therefore, we can conclude that associative knowledge such as similarities of a concept to certain theoretical edifices or to more general concepts can hardly be found by automatic indexing. It becomes clear at this point that the counting of words and/or the comparison of strings cannot produce any additional knowledge beyond the identification of similarities.

User-contributed annotations

Another additional source of annotations beside the automatic ones examined above is user-contributed annotations, commonly referred to as "tagging". For our experimental setup, we used a more restricted form of tagging in order to allow a comparison with other sources. Therefore concepts of the STW were made available for tagging.

For this first exploratory study, we asked an undergraduate of Library and Information Sciences to assign adequate STW concepts to our documents without preparatory training. As a result, we got 579 annotations, roughly a third of the annotations made by the librarians. Thus, we expected that some details would be missing. Figure 6 shows the overall view of the IC Difference Analysis, as anticipated mostly colored in blue, showing that all areas aside from "general terms" are used less frequently.

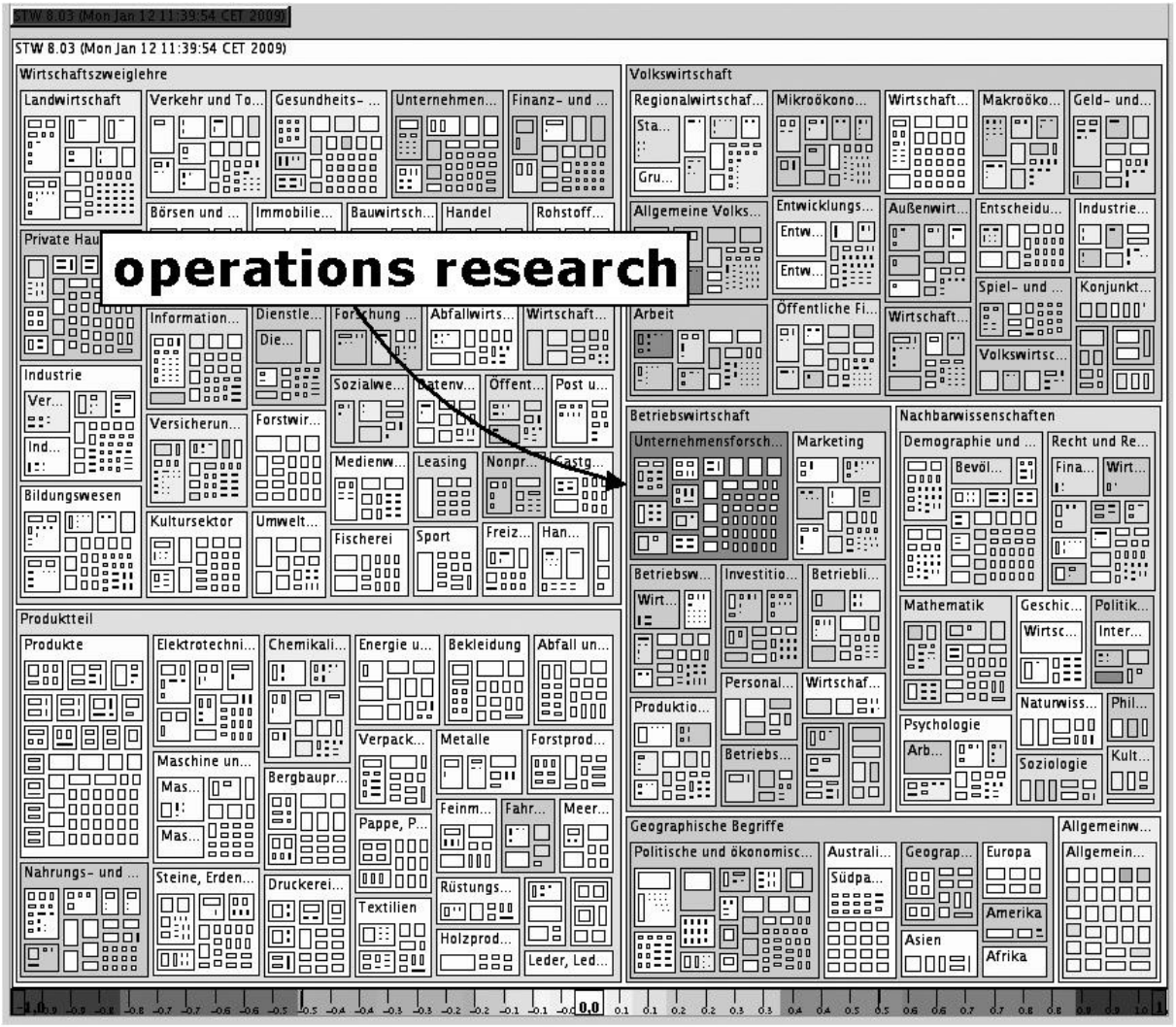

Figure 6. The comparison of user-contributed tags to professional annotations

The most striking area of the overview is the one including the concept "operations research". The concept primarily responsible for this finding is the category "theory". Although "theory" is a concept used very often by librarians, in our case it was never assigned by the undergraduate. The reason can be found by considering the training of the librarians, who usually evaluate a document according to its practical or theoretical focus. In the documents used by this study, this particular aspect is not often mentioned explicitly in the abstract and thus was completely ignored by the undergraduate.

The only area that is marked in red in the overview is the one concerning "general terms". Figure 7 shows these terms in detail. Whereas the "computer-aided methods" are underrepresented, just as with the automatic annotations, terms like "cooperation" or "evaluation" are used more often by the undergraduate. A closer look at the documents involved reveals two reasons: First, the librarians tend to use more specialized concepts in the thesaurus where available. For example, they assign "business cooperation" instead of "cooperation" and "corporate assessment" instead of "evaluation". Second, on several occasions the undergraduate used only one of the "general terms" to describe a concept. We can guess that he failed to find adequate terms in these cases and thereupon switched to a "general term" like "comparison".

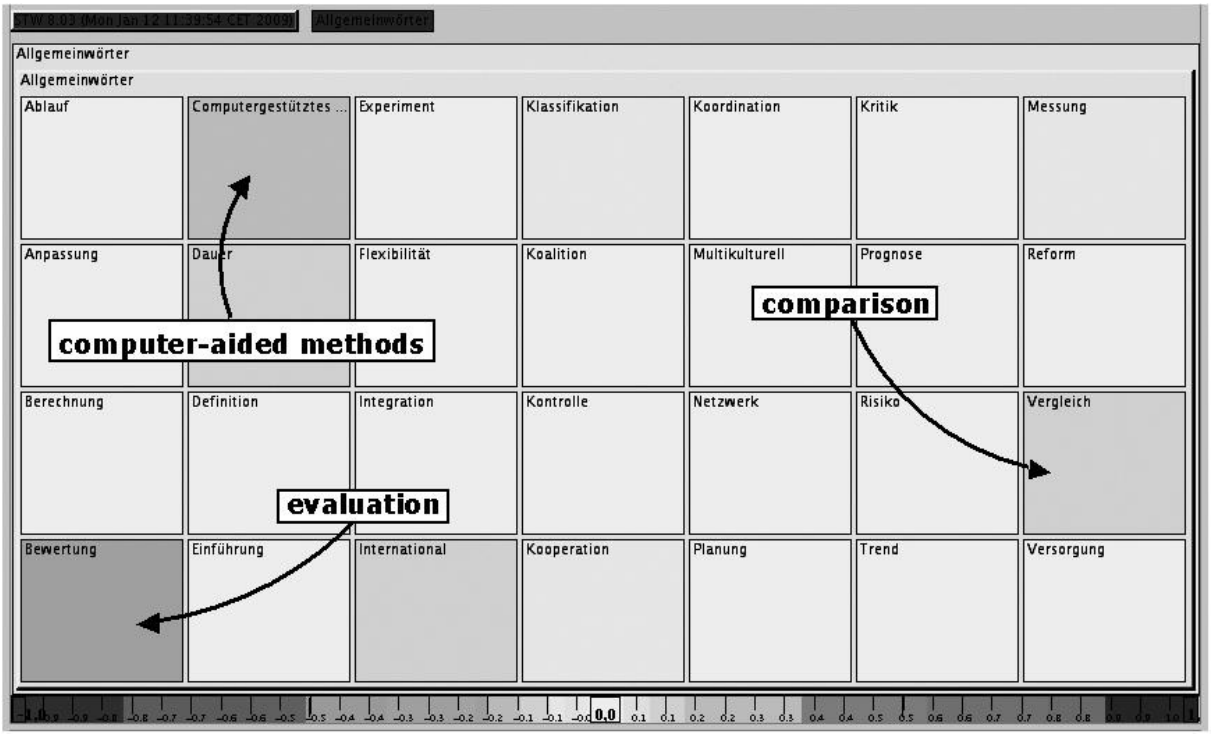

Figure 7. The "general terms" section

Generally, it can be stated that the undergraduate in our example did not make such obvious mistakes as the automatic indexing system (like assigning "paper" to every occurrence of the term). However, the results turned out to be similar in the sense that concepts were generally assigned when they occurred in the text explicitly. Due to the lack of the experience and specialized training of a librarian, the user did not have the same ability (or motivation) to read "between the lines". Nevertheless, the assignment of annotations by the user showed no severe mistakes despite the somewhat imprecise usage of terms. This problem may become less significant if access to all thesauri-concepts is facilitated by a more intuitive and easy-to-use method like the treemap visualization technique.

A last point has to be mentioned regarding the tagging approach. We compared the tagging results of only one user to the annotations of a professional librarian. The general success of tagging in the internet strongly depends on the "wisdom of crowds", the collective intelligence of lots of users. Whereas we have not yet enough data available to prove this effect, it can be expected that at least some of the weaknesses concerning the lack of appropriate annotations could be resolved simply by taking more users into account.

5. Conclusion

In this paper, we presented an exploratory study concerning tagging and automated indexing as a possible source for subject annotations in addition to those traditionally provided by professional librarians. As a preliminary conclusion it can be stated that a combination of librarians' and automatic indexing, as well as tagging, is probably best suited to shape the annotation of literature more efficiently without compromising quality. Especially when adequate subject headings are missing – whether because of time (if the article in question was published very recently) or granularity (if the article will not be annotated in the usual process of a library) – the additional sources can usefully fill this gap. Despite their lower quality they can improve the search experience.

Nonetheless, according to this study the associative and abstract additional knowledge that a specialist contributes to the indexing process

cannot be generated either by automatic indexing or by user-contributed tagging.

After this initial exploration further studies will be needed to perform intensive reviews. With the ongoing growth of scholarly publications, it is indisputable that fast, informal and ad-hoc mechanisms like automation and tagging are needed to keep up with the increasing number of new publications. But we need greater in-depth knowledge about the weaknesses and strengths of both approaches to make the most of them and transform them into a valuable opportunity for academic libraries.

Notes

[i] http://www.flickr.com
[ii] http://delicious.com
[iii] http://www.citeulike.org
[iv] http://www.connotea.org
[v] http://www.bibsonomy.org
[vi] http://www.semtinel.org
[vii] http://www.collexis.com

References

Eckert, K. (2007), *Thesaurus Analysis and Visualization in Semantic Search Applications,* University of Mannheim, Mannheim, available at: http://ki.informatik.uni-mannheim.de/fileadmin/publication/Eckert07Thesis.pdf (accessed 9 May 2009).

Eckert, K., Stuckenschmidt, H. and Pfeffer, M. (2007), "Interactive Thesaurus Assessment for Automated Document Annotation", in *Proceedings of the 4th international conference on Knowledge capture (K-CAP '07)*, *Whistler, BC, Canada*, ACM Pr., New York, available at: http://ki.informatik.uni-mannheim.de/fileadmin/publication/Eckert07Thesaurus.pdf (accessed 9 May 2009).

Eckert, K., Pfeffer, M. and Stuckenschmidt, H. (2008a), "Assessing Thesaurus-Based Annotations for Semantic Search Applications", *International Journal on Metadata, Semantics and Ontologies*, Vol. 3, No. 1, pp. 53-67.

Eckert, K., Pfeffer, M. and Stuckenschmidt, H. (2008b), "Semtinel: Interactive Supervision of Automatic Indexing", in *JCDL '08: Proceedings of the 2008 conference on Digital libraries*, *16–20 June 2008, Pittsburgh, PA, USA,* ACM, New York, demo paper available at:

http://ki.informatik.uni-mannheim.de/fileadmin/publication/Eckert08Semtinel.pdf (accessed 24 February 2009).

Eckert, K., Pfeffer, M. and Stuckenschmidt, H. (2008c), *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), 14-19 September 2008, Aarhus, Denmark,* Springer, Heidelberg, to appear.

Gordon-Murnane, L. (2006), "Social bookmarking, folksonomies, and Web 2.0 tools", *Searcher, The Magazine for Database Professionals, Vol.* 14, No. 6, pp 26-38.

Guy, M. and Tonkin, E. (2006), "Folksonomies: Tidying up tags?", *D-Lib Magazine,* Vol. 12, No. 1, available at: http://www.dlib.org/dlib/january06/guy/01guy.html (accessed 9 May 2009).

Heckner, M., Mühlbacher, S. and Wolff, C. (2008), „Tagging tagging. Analysing user keywords in scientific bibliography management systems", *Journal of Digital Information,* Vol. 9, No. 2, pp. 1-19.

Heymann, P. and Garcia-Molina, H. (2006), "Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems", *InfoLab Technical Report* 10, pp. 1-5, available at: http://ilpubs.stanford.edu:8090/775/1/2006-10.pdf (accessed 9 May 2009).

Mai, J. M. (2006), "Contextual analysis for the design of controlled vocabularies", *Bulletin of the American Society for Information Science and Technology,* Vol. 33, No. 1, pp. 17-19.

Peters, I. and Stock, W. C. (2008), „Folksonomies in Wissensrepräsentation und Information Retrieval", *Information. Wissenschaft und Praxis,* Vol. 59, No. 2, pp. 77-90.

Quintarelli, E. (2005), "Folksonomies: Power to the people", paper presented at the ISKO Italy UniMIB meeting, 24 June 2005, Milan, Italy, available at: http://www.iskoi.org/doc/folksonomies.htm (accessed 9 May 2009).

Razikin, K., Goh, D.H.L., Chua, A.Y.K. and Lee, C.S. (2008), "Can Social Tags Help You Find What You Want?", available at: http://www.springerlink.com/content/5783577131036q60/fulltext.pdf (accessed 16 February 2009).

Resnik, P. (1995), "Using information content to evaluate semantic similarity in a taxonomy", in Mellish, C. (Ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Montréal, Québec, Canada, August 20-25, 1995,* Morgan Kaufman, San Mateo, Calif., Vol. 1, pp. 448-453.

Seco, N., Veale, T. and Hayes, J. (2004), "An Intrinsic Information Content Metric for Semantic Similarity in Wordnet", available at: http://eden.dei.uc.pt/~nseco/ecai2004b.pdf (accessed 16 February 2009).

Shirky, C. (2005), "Ontology is Overrated: Categories, Links, and Tags", available at: http://shirky.com/writings/ontology_overrated.html (accessed 1 February 2007).

Shneiderman, B. (1992), "Tree visualization with tree-maps: 2-d space-filling approach", *ACM Transactions on graphics*, Vol. 11, No. 1, pp. 92-99.

Vander Wal, T. (2005), "Explaining and Showing Broad and Narrow Folksonomies", available at: http://www.personalinfocloud.com/2005/02/explaining_and_.html (accessed 1 February 2007).

About the authors

Kai Eckert is based at the Computer Science Institute, University of Mannheim, Mannheim, Germany.

Christian Hänger is Head of IT-Group in the Department of Digital Services, University of Mannheim Library, Mannheim, Germany. He is the corresponding author and can be contacted at: christian.haenger@bib.uni-mannheim.de

Christof Niemann is a Researcher in the Department of Digital Services, University of Mannheim Library,