Quality Assessment in Digital Libraries – Challenges and Chances

Sascha Tönnies, Wolf-Tilo Balke
L3S Research Center
Appelstraße 9a
30167 Hannover

{toennies, balke}@L3S.de

ABSTRACT

Today, more and more information provider such as digital libraries offer corpora related to a specialized domain. Beyond simple keyword based searches the resulting information systems often rely on entity centered searches. For being able to offer this kind of search, a high quality document processing is essential. In addition, information systems more and more have to rely on semantic techniques during the workflows of metadata generation, search and navigational access. But, due to the statistical and/or collaborative nature of such techniques, the underlying quality of automatically generated metadata is questionable. Thus, the quality assessment of information system's metadata annotations used for subsequent querying of collections has to be guaranteed. In this paper we discuss the importance of metadata quality assessment for information systems and the chances gained out of controlled and guaranteed quality.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital Libraries – *Systems issues*

General Terms

Algorithms, Experimentation.

Keywords

Digital Libraries, Information Quality.

1. INTRODUCTION

Recently the exponential growth of available information can be observed not only in the Web but also in highly specialized domains. Thus, all information providers have to face the problem of information overload. Relying on the work done in the virtual library for chemistry project (*ViFaChem II* [1]) we selected chemistry as an application area. A good example for the information growth in this domain is shown in a press release of the Chemical Abstract Service (CAS). A total of 50 million chemical substances have been indexed on September 2009 in the curated CAS registry, the worldwide most comprehensive registry of chemical substances. Remarkable is that only 40 million substances have been indexed just nine month before. In contrast, the CAS registry contained 10 million entries in 1990 and around 22 million entries in 2000.

Typical Web search engines are already able to handle such an amount of data in a fully automated way by indexing Web content using text-retrieval methods and structural properties of the underlying collection like link analysis. In contrast, there are information providers, e.g. libraries, still relying on manually created indexes due to high quality requirements. These providers have to face two serious problems even for focused collections. First, it is increasingly costly and time consuming to build up a proper index; second, even given an ideal collection, the indexing has to foresee all possible uses for each specific item. Also the information overload for the individual customer and the increasing specialization of (research) interests require indexes to be more specific in the choice of appropriate indexing terms.

Chemical documents have to be extended by two types of metadata. The first type, the bibliographic metadata like authors, affiliation, publisher and year, is obviously readily available in a library environment. The second and for our purposes more important type is chemical metadata, specified by chemical entities, reactions, concepts and techniques, contained in the original document. This chemical metadata is not readily available and must be extracted, collected and structured. Therefore, the development and application of technologies for automated metadata generation gain in importance. The advantage of using such techniques is twofold: First, document processing becomes less expensive and second, a higher degree of personalization is possible. In particular, the usage of semantic techniques has been proposed to bring a higher rate of automation into the indexing process. Commonly used semantic techniques in the domain of digital libraries are the usage of (bibliographic) ontologies, tagging and classification systems. Summarized, semantic techniques rely on statistical and collaborative methods to assess textual documents. However, due to the nature of statistical and collaborative methods, using such techniques may result in a loss of retrieval quality in comparison to handcrafted indexes. For information providers this potential loss in quality is a serious problem; if users cannot trust in the results, the added value of curated information systems over simple Web searches becomes questionable. Hence, before a semantic technique can be used, information providers have to gauge the impact of the technology's use in the retrieval process.

In this paper, we will discuss the problem of quality assessment for semantic techniques in information systems with focus on digital libraries and show first results on quality assessment.

GvD Workshop'10, 25.-28.05.2010, Bad Helmstedt, Germany.

2. CHALLENGES IN QUALITY ASSESSMENT

The development of a digital library is a multi-stage process containing the following steps: Preprocessing of underlying documents, metadata enrichment, indexing of a collection and metadata, and personalized document retrieval. Given that each of these steps may have a loss in quality, the total quality of the overall information is questionable.

Preprocessing. Chemists often search for documents containing particular chemical entities or reactions. Therefore, an important part of document preprocessing is entity recognition, which is a difficult task when working on proprietary document formats, e.g. PDF documents. Due to the unstructured representation of PDF documents it is very difficult to gain high quality during entity recognition. PDF documents store all characters using the absolute position within the document and thus all paragraphs are split into single line paragraphs. Since entity names usually are quite long, the probability that names are split into several parts is rather high. For example, entity extractors have a hard time figuring out whether different parts of a word belong to the same entity or are entities in their own right (for example the chemical name *4-(aminomethyl)cyclohexamine* separated into *4-aminomethyl* and *cyclohexamine*). Another difficulty is the processing of chemical formulas containing superscript and subscript letters which become lost during text conversion using any available software on the market. Even simple document elements like tables and figure captions cannot be processed in high quality. The resulting quality is even worse if the document is still not digitized and OCR software has to be used, because the digitized result will always contain additional OCR errors. These errors are insignificant when building up a full-text index since standard IR techniques are not really affected by unsystematic (OCR) errors. But considering entity centered domains it might already be an interesting factor in the process of tokenization and entity recognition, also affecting the overall retrieval quality [2].

Metadata enrichment. The important part of introducing a significant quality control for information systems lies in the metadata part used for querying the system. The quality of (handcrafted) metadata for traditional libraries may be measured in terms of completeness, correctness and relevance [3], [4]. For instance, the classification of a resource according to the Library of Congress subject headings (LCSH) or the Dewey Decimal System (DCC) can be measured using these criteria. In contrast, considering a semantic technique such as collaborative tagging where users categorize resources with a free vocabulary, such measures are difficult to apply. Thus, semantically enriched metadata has to be evaluated regarding the quality of metadata itself. In the domain of collaborative tagging systems, some work investigating the quality of tags has already been done, see e.g. [5]. But in general, research in the field of quality assessment for semantic techniques is still rare. A good example is [6] where measures for the quality of automatically generated taxonomies for resource classification are investigated. Further approaches like the comparison with other knowledge spaces, e.g., DBpedia and currated databases are possible and have to be evaluated. But a major problem in the quality assessment of metadata generation is the absence of 'gold standards' for benchmarking.

Indexing. Besides simple text indexes for metadata and full texts, chemical information service providers offer specialized indexes built up by identifying and indexing all chemical structures from a document collection in structure databases. The resulting databases can be accessed through graphical interfaces. By drawing a chemical structure a domain expert can thus formulate a query, which in turn will be parsed by the chemical query parser and matched against entities' fingerprints stored inside the structure database. The amount of manual work required for building up and maintaining such indexes results in high costs. Today, CAS offers high quality data at a price of about 30,000 USD/year for a single user subscription. Obviously for the growing open access movement this type of indexing documents is not a viable option. Also the widely used Web search interfaces, e.g. Google or Yahoo! cannot retrieve and search for structural data stored in a database.

Document retrieval. Besides the topical focus, the major success factor for effective information access is the respective user interface and the (generally metadata based) query facility. In terms of suitable interfaces information visualization is becoming increasingly prevalent for understanding and explaining information. Currently, faceted navigation is a popular technique for supporting exploration and discovery of digital libraries and document collections. Facets refer to different kinds of categories used to characterize information items in each corpus. However, the large-subject-space problem is still unresolved and makes innovative, yet understandable extensions of the faceted model essential [7].

Quality assessment. As today's digital libraries more and more rely on automatic enriched (semantic) metadata it seems obvious that the quality assessment of a digital library is not trivial. The traditional quality measures of digital libraries, i.e., the attractiveness of the collections, the technology's ease of use and the user satisfaction [8] may not be sufficient anymore, as all measure the user's experiences and imply high quality (manually generated) metadata. Even though the user satisfaction should be the overall intention, users may look favorably upon the novelty of the interface rather than assess the retrieval effectiveness. Furthermore, during the assessment, the user may not know which information he misses, because of low data quality.

A possible approach would be a combination of the users' satisfaction, and each individual quality aspect during the acquisition of documents, the metadata generation and the retrieval process. Adding up each single quality value may not be sufficient for an overall quality assessment. It seems obvious, that some errors affect each other in such a way, that the overall loss of quality will be higher than the addition of the single values. Though, the interaction of the individual aspects has to be investigated.

3. USE CASE – BUILDING A CHEMICAL DIGITAL LIBRARY

The *ViFaChem II* project focuses on using knowledge about chemical workflows as a basis for creating a digital library portal. The overall vision is a personalized knowledge space for the individual practitioner in the field of chemistry. Building on (automatically derived) ontologies structuring the domain, openly accessible topical databases, and specialized indexes of substances derived from a set of user-selected documents, a personalized knowledge space can be created that promises to help users combating the information flood. The characteristic of a chemical document is that the relevant chemical structure information is not just encoded in text but also in images. Within this project a digital library for chemical documents was build up and integrated into in the Chem.de portal. In this context we already investigated some quality challenges.

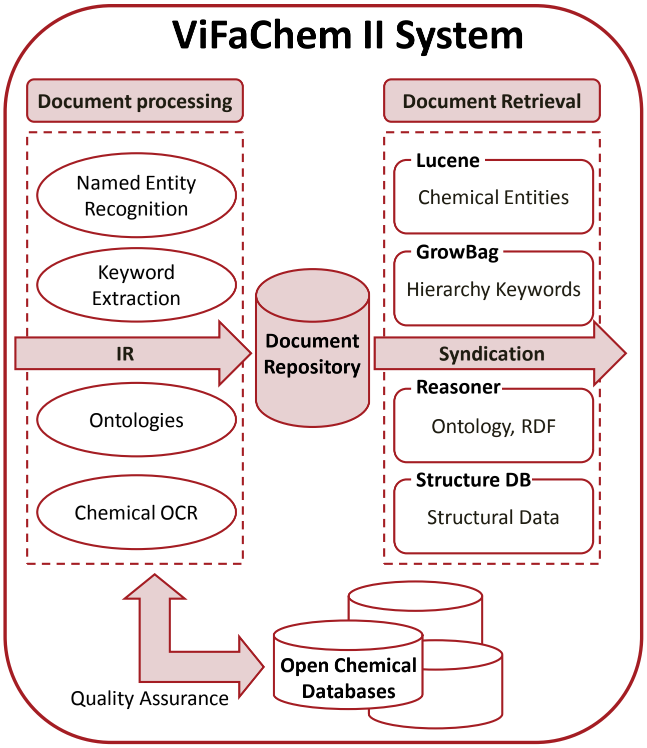


Figure 1. ViFaChem II architecture

Preprocessing. The document collection at the German National Library of Science and Technology (TIB Hannover) contains several hundreds of thousands journal articles, conference proceedings, research reports and online resources**.** For our chemical digital library we indexed, among others, the collection of chemical documents from the journal *Archive for Organic Chemistry* (ARKIVOC)[1] which is one of the most renowned open access sources for organic chemistry. This journal publishes all papers as PDF documents. Since we have to use many tools for deriving metadata for the use in our system, first all the different document types have to be converted into one general interface format. We rely on SciXML (a XML derivate) because the named entity recognition framework (Oscar3) expects that format as input.

One of the first challenges we had to target was the quality loss during the conversion from the PDF format to the SciXML format. We observed that problem during our first experiments; we had a very low entity recognition rate in comparison to our manual annotated corpora, caused by a very poor conversion quality of the PDF documents. Further investigations showed that this was a structural problem of the PDF document format, and that all available conversion tools ran into the same problems during the conversion step. Thus, we were not able to use any of the available tools out of the box and developed a Java-based framework enabling the *ViFaChem II* document processor to convert a document into an object model, verify the model and serialize it as a SciXML file resulting in considerable quality improvement also for keyword extraction.

The last step in the ViFaChem document processing is the chemical OCR. Chemical OCR analyzes images contained within the document and tries to convert these images back to structural information. This information can be stored in a structural database and be used for, e.g. substructure searches. The problem of such semantic technique is the very poor quality as measured in [9]. Thus, this technique should definitely not be used in any automatic information retrieval process today.

Metadata enrichment. The personalized retrieval process of the Chem.de portal relies on classical bibliographic metadata and semantic enriched metadata, i.e. chemical metadata. Whereas the classical bibliographic metadata is derived from the catalog system of the TIB Hannover und thus has high quality due to the review process in the library, the quality of the automatically generated semantic metadata has to be assessed.

First experiments rely on the author keywords which were subsequently used for automatically creating folksonomies. The resulting tag clouds were calculated by the Semantic GrowBag technique [10] investigating higher order co-occurrences. To assess the quality of these graphs, we conducted a user study with domain experts. All experts were asked to think aloud after being exposed to the individual graph and provide feedback on how they assessed the quality and which metadata items were considered to be useful for the average user of the respective collection. Moreover, after reviewing the metadata, the experts were asked about their expectations in terms of correctness and completeness of the automatically generated metadata.

The study resulted in three major observations:

1. Domain experts always started from a (reasonably similar) cognitive classification of possible entities. They expected to find relevant terms with respect to all expected classes.
2. Considering the given metadata all experts expected to find a similar degree of generality / specificity of the keywords. The respective degree was derived relative to the general understanding of the respective domain.
3. Assessing the type of relationship between each keyword and the query term all experts tried to embed the terms in a common context. With increasing broadness of the context, the satisfaction with the keywords decreased.

Based on these observations, we proposed three measures namely degree of category coverage, semantic word bandwidth and relevance of covered terms. Although our preliminary results address the sensibility of our measures, a detailed investigation using several document corpora is still needed to reflect different topics and sizes. In addition, automatically building folksonomies is just one possible semantic technique for an assisted information retrieval and many more are possible, e.g., author networks, personalized document ranking and automated classification of documents. For all of these semantic techniques information providers should try to find possibilities for quality assessment, to fulfill their mission of high quality standards.

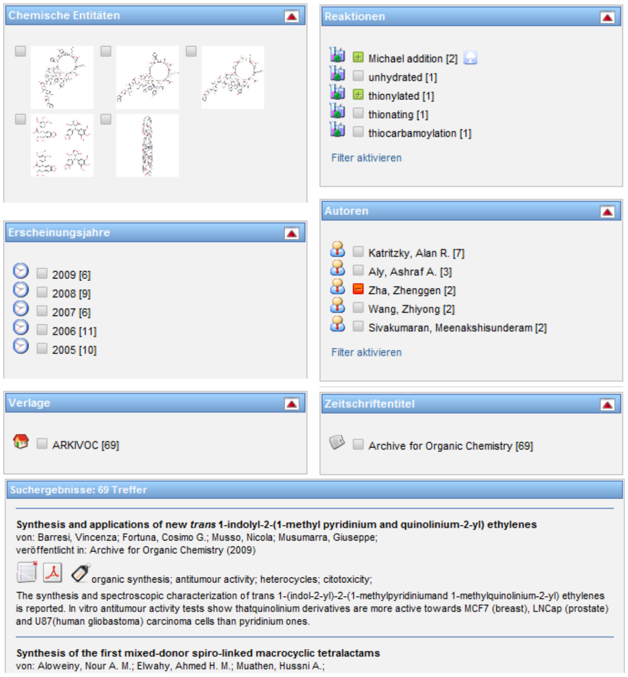
Besides the chemical entities, also reaction names are extracted. These names are linked to the Chemical Entities of Biological

[1] www.arkat-usa.org

Interest (ChEBI)[2] and the Name Reaction Ontology (RXNO)[3] by simple string matching algorithms.

Indexing. Our chemical digital library has different indexes used for different kinds of document retrieval. All extracted chemical entities are converted into chemical structures and are stored within a structure database. This enables the user, to search for documents by drawing a chemical entity as query. In addition, we build up a Lucene based text index containing trivial names of the identified entities and the full-texts. To provide a high recall we had to solve several problems. Chemical substances can have many different and often ambiguous textual representations, like several trivial names, InChI codes or SMILES. In chemical documents besides structure images usually only trivial names are used for brevity and improved readability. We developed a workflow allowing the automatic enrichment of chemical metadata from publicly accessible databases for each occurring chemical entity. In this way, it is possible to provide a simple keyword based search interface with in Chem.de. Our experiments show that the resulting retrieval quality of our enriched index is almost as good as chemical exact structure searches and significantly better compared to a full text search [11].

Document retrieval. A user can retrieve documents by either doing a keyword based search over the text index or a (drawn) structure search over the structure database. A chemical entity search will result in a hitlist of chemical entities. The user can than select the chemical entities of interest to retrieve the documents containing the selected entities. The resulting document hitlist can be further filtered by chemical and bibliographic facets as shown in Figure 2.



Chemische Entitäten

□ □ □ □ □

Reaktionen

⊞ Michael addition [2]
□ unhydrated [1]
⊞ thionylated [1]
□ thionating [1]
□ thiocarbamoylation [1]

Filter aktivieren

Erscheinungsjahre

□ 2009 [6]
□ 2008 [9]
□ 2007 [6]
□ 2006 [11]
□ 2005 [10]

Autoren

□ Katritzky, Alan R. [7]
□ Aly, Ashraf A. [3]
⊟ Zha, Zhenggen [2]
□ Wang, Zhiyong [2]
□ Sivakumaran, Meenakshisunderam [2]

Filter aktivieren

Verlage

□ ARKIVOC [69]

Zeitschriftentitel

□ Archive for Organic Chemistry [69]

Suchergebnisse: 69 Treffer

Synthesis and applications of new *trans* 1-indolyl-2-(1-methyl pyridinium and quinolinium-2-yl) ethylenes
von: Barresi, Vincenza; Fortuna, Cosimo G.; Musso, Nicola; Musumarra, Giuseppe;
veröffentlicht in: Archive for Organic Chemistry (2009)

organic synthesis; antitumour activity; heterocycles; citotoxicity;

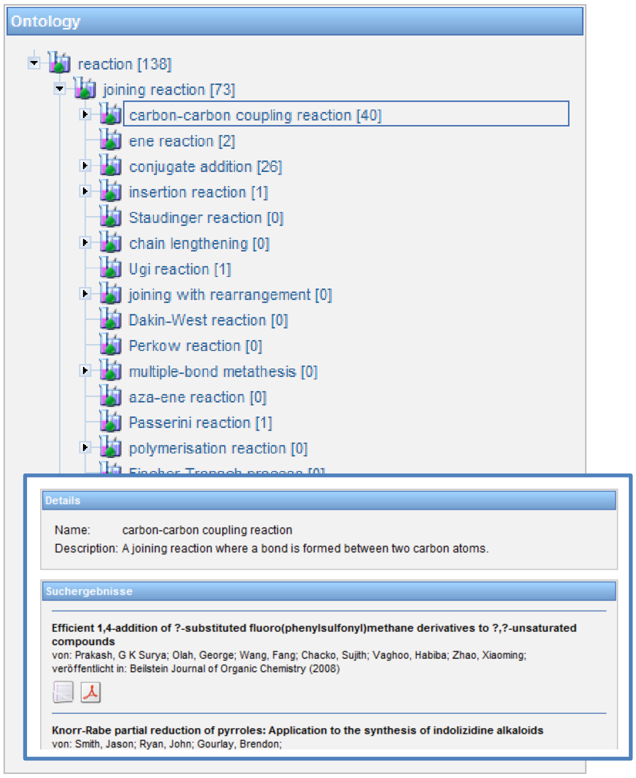
The synthesis and spectroscopic characterization of trans 1-(indol-2-yl)-2-(1-methylpyridiniumand 1-methylquinolinium-2-yl) ethylenes is reported. In vitro antitumour activity tests show thatquinolinium derivatives are more active towards MCF7 (breast), LNCap (prostate) and U87(human gliobastoma) carcinoma cells than pyridinium ones.

Synthesis of the first mixed-donor spiro-linked macrocyclic tetralactams
von: Aloweiny, Nour A. M.; Elwahy, Ahmed H. M.; Muathen, Hussni A.;

Figure 2. Parts of the advanced search interface of the Chem.de portal

Each facet's entry can be selected to be included or excluded. If one entry is included, the document hitlist will only contain documents linked to the facet entry. If one entry is excluded the documents contained in the hitlist are explicitly not containing the facet entry.

The document retrieval process also includes ontology based document retrieval (see Figure 3): All documents containing named reactions are linked to the respective ontology term of the RXNO ontology. Thus, a user can retrieve documents by browsing the RXNO ontology.



Ontology

reaction [138]
joining reaction [73]
carbon-carbon coupling reaction [40]
ene reaction [2]
conjugate addition [26]
insertion reaction [1]
Staudinger reaction [0]
chain lengthening [0]
Ugi reaction [1]
joining with rearrangement [0]
Dakin-West reaction [0]
Perkow reaction [0]
multiple-bond metathesis [0]
aza-ene reaction [0]
Passerini reaction [1]
polymerisation reaction [0]
Fischer-Tropsch process [0]

Details

Name: carbon-carbon coupling reaction
Description: A joining reaction where a bond is formed between two carbon atoms.

Suchergebnisse

Efficient 1,4-addition of ?-substituted fluoro(phenylsulfonyl)methane derivatives to ?,?-unsaturated compounds
von: Prakash, G K Surya; Olah, George; Wang, Fang; Chacko, Sujith; Vaghoo, Habiba; Zhao, Xiaoming;
veröffentlicht in: Beilstein Journal of Organic Chemistry (2008)

Knorr-Rabe partial reduction of pyrroles: Application to the synthesis of indolizidine alkaloids
von: Smith, Jason; Ryan, John; Gourlay, Brendon;

Figure 3. Ontology based document browsing

4. CHANCES OF QUALITY ASSESMENT

Besides the enormous complexity of examine the overall quality of information systems, the quality assessment will also result in chances for information providers such as digital libraries. Today, the major difference between digital libraries and a simple Web search engine as information provider is the given quality. Web search engines do currently not focus on quality but on fast and effective information retrieval. For instance, Google is indexing millions of books for the book search project[4] without considering the requirements discussed in [12]. Digital libraries instead do still rely on information quality. This competitive advantage can only be retained if the quality standard can be guaranteed in the future.

Of course, even if semantic techniques may help in the future to gain automatically generated semantic metadata digital libraries have to spend a lot of money for the quality assessment. But through the assessment of the quality of service it may be also possible to establish new business models. For instance, digital

[2] http://www.ebi.ac.uk/chebi/

[3] http://www.rsc.org/ontologies/RXNO/index.asp

[4] http://books.google.com

libraries can provide higher quality of service to premium customers who will pay money for the services in contrast to the standard customer. That way, the semantic techniques used for the retrieval process could be adapted to the customer's need. For instance, the digital library can adopt a more general ontology used within the retrieval process to the specific domain of the customer and thus gain higher quality.

High quality metadata also result in good options in terms of suitable interfaces of information. A promising example of a beneficial usage of high quality semantic metadata is the *GoPubMed* portal[5] providing ontology-based literature search over around 19 million biomedical research journals in the Medline collection. This portal relies on the manually curated MeSH[6] and gene ontology[7] and thus can offer enormous capabilities in semantic document retrieval.

Generally speaking, the biggest change for digital libraries will be the transparency of the whole information retrieval process. Thereby, the user can understand, how the search result is generated and to what extent the underlying data quality affects the retrieval process. For metaphor purpose a color-coded visualization based on a traffic light may express information quality. In this way, the user knows which quality he can expect from the information and can decide if the given quality is acceptable for his task at hand.

5. FUTURE WORK

Currently, quality is only examined for a few semantic techniques. Therefore, we will investigate different semantic techniques using manual inspection together with appropriate quality measures. Applying these quality measures in a real digital library will be the next step. This will result in an investigation on how the individual quality measures will affect the outcome of other semantic techniques and whether it is possible to tweak them with the quality input.

For the retrieval part of the digital library, the influence of the quality assessment on the user has to be investigated. This implies the personalized creation of retrieval workflows based on the users' quality requirements and the visualization of different quality aspects in the digital library interface.

6. ACKNOWLEDGMENTS

This work was partially supported by the German Research Foundation (DFG) within the *ViFaChem II* project.

7. REFERENCES

[1] S. Tönnies, B. Köhncke, O. Koepler, and W. Balke, "Building Chemical Information Systems - the ViFaChem II Project," Datenbanksysteme in Business, Technologie und Web (BTW 2009), 13. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), GI, 2009.

[2] A. Abdulkader and M.R. Casey, "Low Cost Correction of OCR Errors Using Learning in a Multi-Engine Environment," 10th International Conference *on Document Analysis and Recognition*, IEEE, 2009, pp. 576-580.

[3] D.M. Nichols, C. Chan, D. Bainbridge, D. McKay, and M.B. Twidale, "A lightweight metadata quality tool," *International Conference on Digital Libraries*, 2008.

[4] T. Margaritopoulos, M. Margaritopoulos, I. Mavridis, and A. Manitsaris, "A conceptual framework for metadata quality assessment," *International Conference on Dublin Core and Metadata Applications*, 2008.

[5] K. Bischoff, C.S. Firan, W. Nejdl, and R. Paiu, "Can all tags be used for search?," Conference on Information and Knowledge Management, 2008.

[6] S. Tönnies and W. Balke, "Using Semantic Technologies in Digital Libraries – A Roadmap to Quality Evaluation," 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009, Berlin, Heidelberg: Springer Berlin / Heidelberg, 2009, pp. 168-179.

[7] M. Hearst, "UIs for Faceted Navigation: Recent Advances and Remaining Open Problems," 2008.

[8] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C. Klas, L. Kovács, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Sølvberg, "Evaluation of digital libraries," International Journal on Digital Libraries, vol. 8, 2007, pp. 21-38.

[9] A. Valko and P. Johnson, "CLiDE Pro: A chemical OCR tool," Proceedings of the 8th International Conference on Chemical Structures (ICCS), 2008.

[10] J. Diederich and W. Balke, "The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems," 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), 2007.

[11] S. Tönnies, B. Köhncke, O. Koepler, and W. Balke, "Exposing the Hidden Web for Chemical Digital Libraries," 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL), Surfers Paradise, Gold Coast, Australia: 2010.

[12] S. Tönnies and W. Balke, "User-centered Content Provisioning over Large Collections of eBooks," Proceedings of the 2009 2nd ACM Workshop on Research Advances in Large Digital Book Repositories, BooksOnline 2009, Corfu, Greece, October 2, 2009, 2009.

[5] http://www.gopubmed.org/web/gopubmed/

[6] http://www.nlm.nih.gov/mesh/

[7] http://www.geneontology.org/