Using Semantic Technologies in Digital Libraries – A Roadmap to Quality Evaluation

Sascha Tönnies[1] and Wolf-Tilo Balke[1,2]

[1] L3S Research Center, Appelstraße 9a, 30167 Hannover, Germany
[2] IFIS TU Braunschweig, Mühlenpfordstraße 23, 38106 Braunschweig, Germany
toennies@L3S.de, balke@ifis.cs.tu-bs.de

Abstract. In digital libraries semantic techniques are often deployed to reduce the expensive manual overhead for indexing documents, maintaining metadata, or caching for future search. However, using such techniques may cause a decrease in a collection's quality due to their statistical nature. Since data quality is a major concern in digital libraries, it is important to be able to measure the (loss of) quality of metadata automatically generated by semantic techniques. In this paper we present a user study based on a typical semantic technique used for automatic metadata creation, namely taxonomies of author keywords and tag clouds. We observed experts assessing typical relations between keywords and documents over a small corpus in the field of chemistry. Based on the evaluation of this experiment, we focused on communalities between the experts' perception and thus draw a first roadmap on how to evaluate semantic techniques by proposing some preliminary metrics.

Keywords: Digital Libraries, Information Quality, Semantic Technologies.

1 Introduction

Digital Libraries provide a vast amount of digitized information ranging from collections of cultural heritage to specialized topic centered portals. One of the essential differences between digital libraries and unstructured collections such as the Web, is the focus on information quality. In contrast typical Web search engines base their indexing on text-based measures from information retrieval and structural properties of the collection, e.g. link analysis, whereas digital libraries usually use indexes (manually) crafted from document metadata. Since metadata can express concepts not explicitly occurring in the document, (or leave out concepts explicitly mentioned, but not relevant for the document) the use of a metadata index generally leads to better precision and recall in information services. In addition, library indexes usually rely on controlled vocabularies providing improved retrieval features such as word sense disambiguation or cross language retrieval.

Hence, digital libraries provide an added value over unstructured document collections by offering meaningful access paths. However, given the exponential increase in newly published items even for focused collections, librarians face two serious problems. First it is increasingly costly and time consuming to properly index new items

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 168–179, 2009.

(leading to a delay in actually offering the item to customers); second in an ideal collection, the indexing has to foresee all possible (future) uses for a specific item. Moreover, the information overload for the individual customer and the increasing specialization of (research) interests force indexes to be more and more specific in the choice of appropriate indexing terms. In fact, the vision of today's digital libraries is to provide *personalized information spaces* for each individual customer.

To this end, semantic technologies have been recently proposed to bring a higher rate of automation into the indexing process. In essence semantic technologies rely on statistical methods to assess textual documents and to some degree are therefore capable of mining 'hidden' information from collections. The advantage is twofold, first document processing becomes less expensive and a higher degree of personalization is possible. Though, due to the nature of statistical methods, using these semantic techniques may not result in the same retrieval quality as manual crafted metadata. Second, for libraries, this potential decrease in quality is a serious concern; if users cannot trust in the results, the added value over simple Web searches becomes questionable. Hence, before a specific semantic technique can be adopted for use, libraries need a way to gauge the impact of the technology's use in the retrieval process.

In this paper we discuss the open problem of quality assessment for semantic techniques in digital libraries and provide a roadmap for developing quality assessment measures. We will illustrate the use of our measures specifically in the field of chemistry. The selection of chemistry is driven by the current development of the virtual topical digital library for chemistry within the ViFaChem 2 project[1]. The ViFaChem 2 project is a tight cooperation between the L3S Research Center of the University of Hannover and the German National Library of Science and Technology Hannover (TIB). The project investigates and deploys innovative value-adding services for information provisioning in the area of chemistry. To this aim chemical document corpora are annotated by bibliographic and entity-based metadata using semantic technologies. The project's vision is the creation of personal information spaces that offer a variety of relevant resources tailored to the individual user's understanding of the topic.

This paper is organized as follows: the following section will discuss related work in the field of quality assessment for (semantic) digital libraries. In Section 3 we conduct a user study in a chemical digital library and evaluate communalities in experts' interactions with automatically generated metadata in the form of related keywords. Preliminary metrics for measuring the quality of semantic technologies are then derived and discussed in section 4. We close with a short summary and outlook.

2 Related Work

In this section we will first discuss the current state of the art in assessing the quality of classical (mostly manually maintained) digital libraries and then turn to the extension to evaluating semantic technologies. A short case study shows how evaluations of such technologies are actually carried out today.

[1] http://www.L3S.de/vifachem

2.1 Evaluating Quality in Digital Libraries

What defines a high quality digital library? In 2000, Saracevic was one of the first authors to consider this problem [27]. He argues that any evaluation basically raises issues such as the criteria, the measures, the context and the methodology. However, his analysis shows that there is no agreement regarding the exact elements of these issues for digital library evaluation. Trying to fill some gaps in this area, Fuhr et al. developed a new description scheme using four major dimensions: collection, technology, users and uses [7]. Based on this dimensions, a questionnaire was developed and the need for an appropriate test collection was stated, similar to the TREC and CLEF initiatives. Extending this work, Gonçalves et al. [11] proposed an actual quality model for digital libraries which is deeply grounded in the formal 5S framework [12]. Exposing several digital library key concepts, several dimensions of quality were added to each concept. For each of these dimensions, the variables to measure, together with the respective S were identified.

The first comprehensive study on digital libraries evaluation frameworks is presented in [8]. The attractiveness of the collections, and the technology's ease of use are identified as key factors in assessing the quality of a digital library. Moreover, the importance of the user satisfaction is emphasized. The model presented is the interaction triptych model which defines three components of the digital library: the system, the content, and the user. In addition three axes of evaluation were provided: usability of user interaction with the system, usefulness of the content for the user, performance of managing the content by the system. Recent research is trying to adopt Web metrics, originally developed for evaluating e-commerce applications, for evaluating digital libraries [18]: preliminary results discuss, e.g., the usage of session length for evaluating the customers' satisfactions with the portal.

2.2 Extending Measures to Semantic Digital Libraries

With upcoming semantic digital libraries like JeromeDL [20] the question of quality has to be extended: what defines a high quality *semantic* digital library? Kruk et al. do not really answer this question when evaluating JeromeDL against a standard digital library measuring several traditional aspects like precision / recall and the user satisfaction [21]. The conducted user studies imply that the individual user's satisfaction seems to be higher when using semantic technologies. However, it has to be pointed out that the results shown in [21] cannot be generalized, since semantic techniques are just as good as the underlying metadata.

Particularly in the domain of collaborative tagging systems, some work investigating tag quality has been performed. According to [10] the distributions of different tags for each individual document tend to stabilize over time, i.e. more and more users add meaningful tags whereas irrelevant tags are not amplified. This result is confirmed in [13] and the authors show in addition, that tags follow a power law distribution. Considering these properties of collaborative tagging systems, it seems likely that tag data can, indeed, be a reliable source of information.

For searching and metadata creation within tagging systems, [15] proposes the exploitation of co-occurrence of users, resources, and tags. This is done using a graph model to represent the folksonomy. In [1] tag data is explored for the purpose of Web search through the use of two tag based algorithms: one exploiting similarity between tag data and search queries, and the other one utilizing tagging frequencies to determine the quality of Web pages. Chan examined a huge number of query terms posed to Powerhouse and concludes that the combined usage of folksonomies with taxonomies increases the recall of the information seeking process [3]. In contrast [25] found out that the use of only document terms yielded slightly better F-measure than using terms and tags together. The authors' results suggest that not all tags are useful descriptors for resource sharing. This leads to the question which kind of tags have a high quality: Bischoff et al. [2] showed that it is worthwhile having a common tag classification scheme for different collections – allowing tags to be compared tags used in different tagging environments. The experiments show that more than 50% of all existing tags bring new information to the resources they annotate and that a large amount of tags are accurate and reliable. A general algorithm for measuring the quality of tags is proposed in [19]. The authors decoupled the relationship between users and tag-resource pairs modeling the tag-resource pairs as nodes and co-user relationship as edges of a graph. This structure allows every two tag-resource pairs used by the same user to have different quality. The algorithm then propagates quality scores iteratively through the graph after being initialized with a set of seed nodes.

In categorization systems, especially in the ontologism field, much work has been done, and several metrics for assessing the quality of an ontology have been proposed, e.g. QOOD [9], OntoMetric [23], and OntoQA [28]. However, all these metrics remain purely on the structural level of the ontology, which is according to [29], not sufficient. In particular, the semantic quality, in terms of correctness, has to be addressed and the authors propose the development of semantically aware ontology metrics. As a first step the authors define the normalization of ontologies and introduce the term of stable metrics. The measurement of the semantic of on ontology becomes vital considering automatically generated ontologies.

2.3 Use Case Study: Evaluating the Semantic GrowBag

Let us consider a typical way of accessing digital collections. Metadata in the form of descriptive terms is often used to describe and summarize documents, and navigational access. Such terms can either be provided by the documents' authors, or be derived from controlled vocabularies, e.g. by the publisher. The collections then allow users to browse documents based on the keywords organized by some categorization system or thesaurus, i.e. searches can be broadened by choosing more general terms or focused by using more specific terms. However, creating and maintaining the underlying categorization systems is primarily done manually with very high efforts and they are often only available for specific domains.

To limit these efforts recently semantic techniques to automatically created categorization systems in the form of taxonomies have been proposed. Examples are statistical evaluation of term co-occurrences [26], language models [4], or syntactical contexts [14]. Although such techniques allow the automatic creation of taxonomies, the suitability of the resulting classification system for actually searching documents is problematic. How can the quality of such generated taxonomies be assessed? For Web search rephrasing queries in different terms is acceptable, however users of digital libraries expect clear and efficient navigation paths. Hence, the measuring of classification systems' quality becomes a vital part in the adoption of semantic technologies.

The actual measurement widely varies in semantic technology research ranging from manual inspection (of random partitions) of the taxonomy to comparison of the entire taxonomy with some kind of 'gold standard'. For instance, in the area of (bio-) medical collections the MeSH taxonomy [16] provides an often used benchmark: when putting an implementation to the test it is run over a focused collection e.g. the Medline corpus [17] and the resulting taxonomy is compared to the corresponding MeSH entries and their respective relationships. For example, in [6] a technique called Semantic GrowBag (based on term co-occurrences, for details see [5]) is used to compute more than 2000 individual taxonomies over Medline documents. It is interesting to notice that for deriving sensible topical taxonomies a minimum of about 100,000 documents was necessary, since statistical methods only provide meaningful results using a sufficiently large sample. For evaluation, the average percentage of accordance or discrepancy with respect to MeSH is presented. Still, it is not clear what these percentages mean in terms of the libraries usability when the respective taxonomies are used as classification system for navigational access.

3 Experiments over a Digital Collection of Chemical Documents

We conducted a user study by observing experts, in our case practitioners in the field of chemistry, when working over a topic restricted document collection with metadata automatically created by semantic technologies. The aim of the study was first to get a deeper understanding of the process of evaluating metadata and assessing the individual expectations second the actual helpfulness of the metadata provided.

For the experiments we used a corpus of 1000 documents randomly extracted from the Journal of Synthetic Organic Chemistry published by Thieme Publishers, Stuttgart Germany. For the metadata extraction, we focused on the author keywords which were subsequently used for automatically creating folksonomies. The actual graphs were calculated by the Semantic GrowBag technique [6] investigating higher order co-occurrences of the keywords in relation to the respective documents. A term A is considered to be 'more general' than some term B, if B usually occurs together with A, whereas A also occurs in other contexts. In that case a directed edge is added from A to B. Together with the graph structure the Semantic GrowBag technique also allows a confidence assessment for each relationship visualized by bold (strong) or dashed (weak) arrows. The Semantic GrowBag uses a biased page rank algorithm to determine this confidence. In Fig. 1 '*amino acids*' is considered more general than '*amino alcohols*' which is indeed justified by amino alcohols being a subclass of amino acids. Note, however, that a relationship as given by the GrowBag graphs does not always express a subclass (or 'is-a') relationship, but just points out that in terms of usage as reflected by the document collection the parent term is more general than the child term.

We extracted a total of 680 graphs (e.g. Fig. 1), each representing the semantic environment for all sufficiently discriminative keywords. The page rank of each term (the number in brackets) in the graphs was also used to create the related tag clouds for the keywords (e.g. Fig. 2). The respective size of each term in the tag cloud is proportional to the page rank value of the term in the GrowBag graph. Please note that in principle the tag cloud contains all information which is available in the graph (terms and their respective page rank) just the hierarchical structure (edges) is missing.

For the actual experiments we randomly chose three query terms for each expert to evaluate the quality of the given graphs and the respective tag clouds. All experts were asked to think aloud after being exposed to the individual graph or tag cloud and provide feedback on how they assessed the quality and which metadata items were considered to be sensible for the average user of the respective collection. Moreover, after reviewing the metadata for each query term, the experts were asked about their expectations in terms of organization of the metadata and the respective correctness and completeness of the automatically created metadata vocabulary.

3.1 A Case Study

In this case study, we describe a typical expert's interaction with a generated graph (Fig. 1) / cloud (Fig. 2) for the query term '*amino alcohols*' to illustrate the conduction of our user study. A first expert was asked about the graph representation and a second about the cloud representation. The graph and cloud contain the same terms and just differ in the visualization and connections between terms.

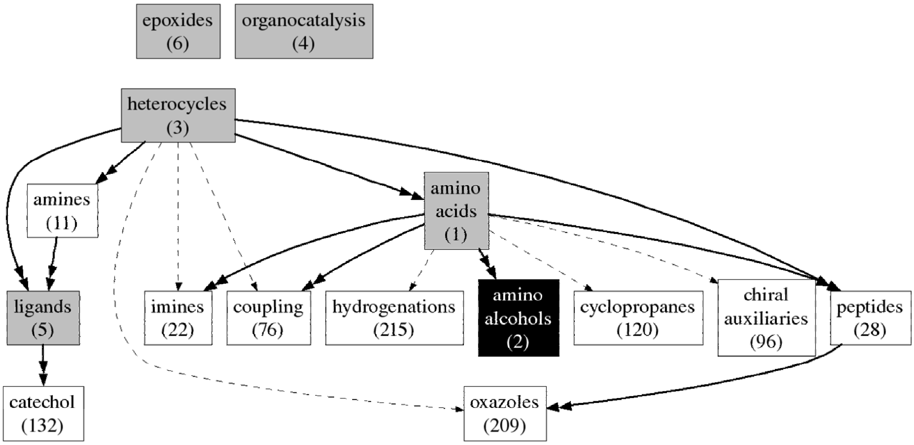

Fig. 1. The generated GrowBag graph for the keyword '*amino alcohols'*

Given the graph as shown in Fig. 1, the expert immediately pointed out that the query term represents a class of *chemical entities*; therefore, he expected to see several *attributes of this class*, typical *reaction names* where amino alcohols are used, *technical uses* and some specific terms from an *analytic* point of view. Following these expectations he clustered the elements into the following groups:

- reactions: 'coupling' and 'hydrogenations'
- classes: 'cyclopropanes', 'oxazoles', 'heterocycles', 'peptides', 'imines', 'amines', 'amino alcohols', 'amino acids', and 'epoxides'
- general concepts: 'chiral auxiliaries', 'organocatalysis', and 'ligands'
- instances: 'catechol'

In a next step, the expert noticed that there are significant differences in the generality of the terms, e.g. ‘*heterocycles*’ has been seen as a very general term whereas ‘*cyclopropanes*’ is a more specific term. For the last step of interaction, the relationships were analyzed: the expert considers some useful, e.g. ‘*peptides*’ are connected via their building blocks ‘*amino acids*’ with ‘*amino alcohols*’ which fits better than a direct connection to 'amino alcohols' and others not useful, e.g. ‘*catechol’* which represents a ‘*hydroxyl benzene*’ with no obvious connection to ‘*amino alcohols*’.

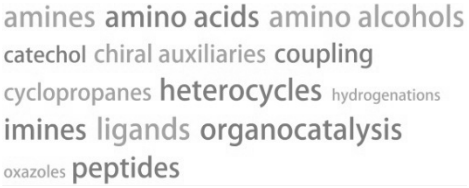

Fig. 2. The generated Tag Cloud for the keyword *amino alcohols*

After giving the equivalent tag cloud (Fig. 2) to the expert, it was interesting to note that the interaction was to a large degree identical with the graph-based representation. The expert started with predefined categories and tried to assign the terms, second the generality of the terms was judged and third the terms were linked to the query term. It has to be pointed out that the expert working on the tag cloud had much more problems during the last step, due to the way of visualization. For instance, he was surprised about the font size of ‘*cyclopropanes*’ and ‘*oxazoles*’. Due to the fact that ‘*cyclopropanes’* is not related to the query term, he expected the font size to be much smaller than, e.g., the size of the heavily related term ‘*oxazoles’*.

3.2 Experimental Results

The evaluation of our observations showed that all practitioners made three major steps during the interaction with the offered metadata.

All experts started with some initial expectation for the categorization of metadata terms. First, they categorized the query term, e.g. as a substance class and then settled on semantically related subcategories based on the main category. It was interesting to see, that these subcategories varied slightly based on the background of the expert. For instance, an expert in the domain of medical chemistry also mentioned the pharmacological impact, whereas a process engineer mentioned environmental perils and toxicity. This observation leads to the conclusion that a categorization of the terms, as it is done, e.g. for the faceted browsing, is indeed useful for the customers and that the structure of a tag cloud may not always be sufficient for visualizing this kind of semantic metadata. It seems that the distribution of terms over relevant categories is one useful metric for measuring the quality of the generated metadata. In our experiments over 90% of the expected categories were indeed filled by matching keywords.

In the second step, the experts tried to understand the content of the graph / cloud. For this purpose they evaluated the terms regarding their respective generality / specifity. This was done without considering the query term. This step has been used by the experts to eliminate outliers in terms of very general or very specific keywords. In particular during our experiments the experts considered 32% of the provided keywords as being too general / specific for the respective graph / cloud.

The last step was the evaluation of the semantic closeness regarding the query term. During this evaluation step the visualization of the metadata affected the experts. Working on the graph, every term was judged individually and depicted relationships were readily taken as explanations. The experts which worked on the cloud did not have these relationships and, therefore, were confused about some terms. Even worse, the font size of the term influenced the experts far more than the confidence in the GrowBag graph. These observations imply the usage of different visualizations: using a cloud for well connected terms and using a graph for the others. In summary, the experts used their individual knowledge to understand the occurrence of the terms and if they could not make a direct connection between a keyword and the query term, they tried to connect the term via some other occurring terms in the graph. If this also failed, they considered the term as wrong or irrelevant for the query. In our experiments this happened with 12% of the occurring terms: this means that 88% have been classified correctly.

4 Towards Measuring Semantic Information Quality

The experiments in the previous section provide some ideas regarding the quality measurements for a semantic technology. Generally speaking, quality can be defined as *correctness* of information. For the field of chemistry this is especially true for data maintained in typical databases like molecular weights or boiling point of substances. However, with respective to semantic, e.g. given by author keywords, the actual correctness is somehow difficult to assess. Observing the expert we found that experts gorge the correctness rather in terms of helpfulness of a keyword and the understandability of the keywords' relationships to a query term. According to the three steps observed during the experiment, we found some communalities between experts. Based on this we will now discuss three preliminary quality metrics that of course have to be further evaluated in future work.

4.1 Degree of Category Coverage (DCC)

The evaluation of the experiments showed that all experts from the start have an implicit course topic map, together with possible classifications for entities in mind. Although the topic map differed with the individual interests of the expert, it is interesting to note that the basic entity classification was very similar (in a way, reflecting the typical cognitive instruments of a chemist). According to this implicit classification, each expert tried to categorize the metadata terms automatically created by the semantic technology. The choice of categories under consideration slightly differed according to the query term and the experts expected at least the closest categories to be filled with keywords found in the graphs, respectively clouds.

This leads to the *degree of category coverage* metric which has to measure how many of the expected categories are actually filled with terms. The more categories are filled the better the result quality is.

With $C := \{c \mid c$ relevant category in the topical classification$\}$ we define:

$$f(c) = \begin{cases} 1 & \text{if there is at least a single term } t \text{ in category } c \\ 0 & \text{else} \end{cases}$$

In addition, the metric also has to measure how many of the given terms do not fit to at least one of the expected categories. The more terms can be allocated, the better the result quality is.

With $T := \{t \mid t$ term from a given metadata subset$\}$ we define:

$$g(t) = \begin{cases} 1 & \text{if term } t \text{ belongs to some category from } C \\ 0 & \text{else} \end{cases}$$

This results in:

$$DCC = \frac{\sum_{i=1}^{|C|} f(c_i)}{|C|} + \frac{\sum_{j=1}^{|T|} g(t_j)}{|T|}$$

4.2 Semantic Word Bandwidth (SWD)

The Semantic Word Bandwidth (SWD) should reflect the results of the second interaction step: the experts estimated the overall generality / specificity of the given terms. Of course this bandwidth can only be evaluated with respect to the highest possible bandwidth. The smaller the bandwidth, the more focused is the set of related keywords.

Considering categorizations where we can rely on some ISA hierarchy, e.g. taxonomies of chemical substances, it is quite simple to determine the bandwidth. In this case, we have to identify the depth within the hierarchy for each term. Using the maximum and the minimum depth of terms normalized by the total depth of the hierarchy (*maxdepth*) the semantic word bandwidth can be defined as follows:

$$SWB = \frac{\max_{t \in T}(depth(t)) - \min_{t \in T}(depth(t))}{maxdepth}$$

In cases where no ISA hierarchy is given, it is much more complex to estimate the semantic word bandwidth. For instance, considering substances (e.g. reactants or catalysts) involved in chemical reactions could be considered more specific but in any case this would need a complex ontology describing the relationships for reactions which can currently not be found in the market place.

4.3 Relevance of Covered Terms (RCT)

The last measure used by the experts tried to determine the usefulness of a term in relation to the query term. If we consider again some ISA hierarchy or an ontology, we may express the usefulness of a term in relation to a query term as the semantic similarity between those terms. The total relevance can then be established as the average similarity of keywords to the query term.

Practically, this can be done by analyzing the underlying ontology. All keywords are associated with concepts in the hierarchy. A direct method for measuring the respective similarity is then to find the minimum length of any path connecting the two concepts [24]. However, according to [22] this may not be sufficient for more general and larger ontologies, and thus, the similarity should be a function of the attributes path length, depth and local density.

Another possibility to measure the relevance of the covered terms may be reflected by using independent semantic techniques. In our example, the Semantic GrowBag uses statistical information to compute higher order co-occurrences of keywords. Thus, the relations shown in the graphs reflect some characteristics of the underlying document collection. The naïve way of interpreting the results is that all terms covered by one graph are somehow used together with the query term. If we assume that terms which are more related to the query term are also generally used more often in relation with some document, this should also be reflected by a simple Web search query. Thus, a two term query for a query term *qt* and a word w_1 which are closely related should result in more hits than a query for *qt* and some word w_2 that are not as closely related. Preliminary experiments based on our used graphs seem to support this assumption, e.g. a Google search for the query '*amino acids* AND *amino alcohols*' yields 39,800 hits and the query '*amino alcohols* AND *cyclopropanes*' only yields 2,540 hits.

5 Conclusions and Outlook

Semantic techniques are ubiquitous in modern information systems and digital collections. In this paper we dealt with the question whether the expected loss of quality due to the use of statistical techniques can be measured. We argue that the development of such measures is especially important for their safe and sustainable application in digital libraries which generally have higher quality constrains in comparison to, e.g. Web search engines. Putting the focus on automatic metadata creation as provided by related keywords, we conducted a user study in the field of chemistry observing some experts' interaction with the created 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 (DCC), semantic word bandwidth (SWB) and relevance of covered terms

(RCT). Although our preliminary results address the sensibility of the measures, a detailed investigation using several document corpora is still needed to reflect different topics and sizes. In addition, the quality of digital libraries does not only result in high precision but also in high recall. This is not faced in our metrics yet, but will be investigated in the future. Therefore, our future work will focus on the creation of suitable test corpora and will measure different semantic techniques using manual inspection together with appropriate quality measures.

References

1. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: WWW 2007: Proceedings of the 16th international conference on World Wide Web. ACM Press, New York (2007)
2. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Can all tags be used for search? In: CIKM 2008: Proceeding of the 17th ACM conference on Information and knowledge management. ACM Press, New York (2008)
3. Chan, S.: Tagging and Searching – Serendipity and museum collection databases. In: Proceedings of Museums and the Web 2007. Archive & Museum Informatics 2007, Toronto (2007)
4. Cimiano, P., Handschuh, S., Staab, S.: Towards the self-annotating web. In: Int. Conf. on the World Wide Web (WWW). ACM, New York (2004)
5. Diederich, J., Balke, W.-T.: The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 1–13. Springer, Heidelberg (2007)
6. Diederich, J., Balke, W.: Automatically Created Concept Graphs using Descriptive Keywords in the Medical Domain. In: Methods of Information in Medicine (METHODS), Schattauer, vol. 47(3) (2008)
7. Fuhr, N., Hansen, P., Mabe, M., Micsik, A., Sølvberg, I.T.: Digital Libraries: A Generic Classification and Evaluation Scheme. In: Constantopoulos, P., Sølvberg, I.T. (eds.) ECDL 2001. LNCS, vol. 2163, p. 187. Springer, Heidelberg (2001)
8. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., et al.: Evaluation of digital libraries. In: Int. J. on Digital Libraries, vol. 8(1) (2007)
9. Gangemi, A., Catenaccia, C., Ciaramita, M., Lehmann, J.: Qood grid: A meta-ontology-based framework for ontology evaluation and selection. In: Proc. of the 4th International Workshop on Evaluation of Ontologies for the Web (EON 2006), Edinburgh, Scotland (2006)
10. Golder, S.A., Huberman, B.A.: The structure of collaborative tagging systems (2005) CoRR abs/cs/0508082
11. Gonçalves, M.A., Moreira, B.L., Fox, E.A., Watson, L.T.: What is a good digital library? In: A quality model for digital libraries. Inf. Process Manage, vol. 43(5) (2007)
12. Gonçalves, M.A., Fox, E.A., Watson, L.T., Kipp, N.A.: Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. ACM Trans. Inf. Syst. 22(2) (2004)
13. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In: WWW 2007: Proceedings of the 16th international conference on World Wide Web. ACM Press, New York (2007)
14. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Int. Conf. on Computational Linguistics, Nantes, France (1992)

15. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
16. `http://www.nlm.nih.gov/pubs/factsheets/mesh.html` (last accessed on 25.03.2009)
17. `http://www.nlm.nih.gov/pubs/factsheets/medline.html` (last accessed on 25.03.2009)
18. Khoo, M., Pagano, J., Washington, A., Recker, M., Palmer, B., Donahue, R.A.: Using web metrics to analyze digital libraries. In: JCDL (2008)
19. Krestel, R., Chen, L.: The art of tagging: Measuring the quality of tags. In: Domingue, J., Anutariya, C. (eds.) ASWC 2008. LNCS, vol. 5367, pp. 257–271. Springer, Heidelberg (2008)
20. Kruk, S.R., Woroniecki, T., Gzella, A., Dabrowski, M.: JeromeDL - a Semantic Digital Library. In: Semantic Web Challenge (2007)
21. Kruk, S.R., Kruk, E., Stankiewicz, K.: Evaluation of Semantic and Social Technologies for Digital Libraries. In: Semantic Digital Libraries. Springer, Heidelberg (2009)
22. Li, Y., Bandar, Z.A., Mclean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering 15(4) (2003)
23. Lozano-Tello, A., Gómez-Pérez, A.: OntoMetric: A method to choose the appropriate ontology. Journal of Database Management, Special Issue on Ontological analysis, Evaluation, and Engineering of Business Systems Analysis Methods 15(2) (2004)
24. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19(1) (1989)
25. Razikin, K., Goh, D.H.-L., Chua, A.Y.K., Lee, C.S.: Can social tags help you find what you want? In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 50–61. Springer, Heidelberg (2008)
26. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In: Proc. of Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Berkeley, CA, USA. ACM, New York (1999)
27. Saracevic, T.: Digital library evaluation: toward evolution concepts. Library Trends 49(2) (2000)
28. Tartir, S., Aroinar, I.B., Moore, M., Sheth, A.P., Aleman-Meza, B.: OntoQA: Metric-based ontology analysis. In: Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge sources (2005)
29. Vrandečić, D., Sure, Y.: How to design better ontology metrics. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 311–325. Springer, Heidelberg (2007)