**Margaret E. I. Kipp**
**University of Western Ontario, London, Ontario, Canada**

# Tagging for health information organisation and retrieval

**Abstract:** This paper examines the tagging practices evident on CiteULike, a research oriented social bookmarking site for journal articles. Articles selected for this study were health information and medicine related. Tagging practices were examined using standard informetric measures for analysis of bibliographic information and analysis of term use. Additionally, tags were compared to descriptors assigned to the same article.

## 1. Introduction

Professionals in information related fields are faced with a constantly increasing volume of literature being published worldwide. As a result, they tend to limit themselves to a few tried and true journals that pertain directly to their specialised areas of research. However, interesting papers may be published outside of their usual journals. As in other fields, medical professionals seek to capture the benefits of a world in which papers can be located via keyword or free text search in digital libraries or on the web. However, materials on the web increase in volume as fast, or faster, than traditional paper journals. Many medical professionals seek methods for finding material directly pertaining to their work that has not yet been indexed in on-line databases and systems that do not require extensive training to search. Google and otherx search engines can provide a multitude of results (Tang and Ng, 2006a; 2006b), but a health information professional needs to ensure that only accurate information is retrieved (Wentz, 2006). Social bookmarking, a phenomenon where users tag items for their own user, offers a potential way to locate new and relevant information. CiteULike (http://www.citeulike.org), a social bookmarking service specialised for academics, allows academic articles to be tagged with useful keywords for later retrieval.

Information organisation has been an important issue since the invention of writing. The subsequent recording of information created the first sets of documents that needed to be stored for later retrieval. While early writing and information access was restricted to the small group of educated citizens; mass education and mass production have created an increasing amount of information with a resultant interest in locating and using that information. As Eisenstein notes in "The Printing Revolution in Early Modern Europe" the development of the movable type printing press in the 1500s caused an upsurge in the amount of printed material and provided a pressing need to increase the capacity of organisational systems for documents (Eisenstein, 1983). Information creation has continued apace and methods for storing and transmitting material electronically, especially via the Internet, have only increased the average user's thirst for access to information. This rapid expansion of both information and access to it is rapidly outpacing attempts to enhance organisation and retrieval and creates a fresh need for new methods of information organisation.

One such shift in the pace and volume of information production is occurring now as academics and researchers increasingly turn to the web to locate articles, often in preprint archives. The increasing existence of open access archives and web archives of conference proceedings has increased the availability of research materials prior to journal publication. Information organisation is intended to reduce the difficulty inherent in searching large document spaces for information. A classification system using terms and keywords appropriate to the context of the intended user can help make the difference between a usable document space and a space which is difficult to navigate and find the information sought. Library style hierarchical classification systems and subject specific taxonomies have a long history, but the development of these systems has largely been left to professional indexers. Despite the utility of such systems for organising material, users often find these systems do not provide the access they seek. Full text search can provide fine grained access to materials by allowing users to search using natural language, but this access comes at the expense of precision due to the use of differing terminology. User tagging and folksonomies created in a distributed fashion through social bookmarking sites have been touted as a potential solution to these problems (Hammond et al., 2005; Kipp and Campbell, 2006; Voss, 2007), but only if user tagging provides a similar or better search context.

## 2. Social Bookmarking Tools

Social bookmarking tools have become increasingly popular since del.icio.us was first released to the public by Joshua Schachter. Additional social bookmarking tools have since been developed including CiteULike, which is specialised for academics. CiteULike was created in November 2004 by Richard Cameron after a frustrating experience with standard bibliographic tools. (Cameron, n.d.) He originally planned CiteULike as a web based bibliographic tool which could be used to store citations.

Social bookmarking services take advantage of the availability of many academic papers through on-line bibliographic databases and open access publications. Since this information is available on the web, it can be collected by a web applet and transferred directly to the social bookmarking system. Users are encouraged to enter all their articles in the database and provide useful tags and notes to go with them. Each article may be entered by many users and each user may enter many articles. Each entry is referred to as a post and a post consists of bibliographic information about the article as well as user assigned tags and notes.

On CiteULike, a post consists of standard bibliographic details such as article or book title, author names, journal names and issue information in addition to the tags and notes provided by the user. Users are not required to provide notes or tags, but their use is recommended. Just as in del.icio.us, CiteULike users are encouraged to take advantage of the network effects of the collaborative venture, but no effort is made to suggest tags to the user. Once entered, it is possible to view the tags used by other users and the existing tag list can be modified.

Currently, articles listed on CiteULike are most likely to be from medicine and the biological sciences, but this is not an inherent feature of the system, only of the early adopters or most frequent users.

## 3. Related Studies

Bowker and Star (1999) suggest that classification is a basic practice of all humans. (Bowker and Star 1999) Traditional classification methods have tended to rely on trained indexers, cataloguers or taxonomists to organise and describe information. While other groups have been involved in creating keywords or index terms (for example, journal article authors who are asked to provide a certain number of keywords with their submitted articles), these keywords generally have a small circulation and are not widely used. Such small scale indexing is common but generally covers a narrow range of topics and is specific to the article. Additionally, such keywords are often derived from the work itself and may or may not have wide circulation outside a small subset of the field. Collaborative tagging systems such as CiteULike allow users to publicly participate in the classification of journal articles.

To discover if tags can truly provide a useful replacement or enhancement for controlled vocabularies, it is important to examine whether or not they provide a similar contextual dimension to the existing classification systems. While it seems unlikely that untrained users will produce a full featured classification system similar to the traditional library systems, it is possible to examine the tags they do assign to see how they compare to the descriptors assigned by a trained indexer and to keywords assigned by authors.

Adam Mathes (2004) notes that there are three major groups that are commonly involved in the classification of documents. These groups are authors, intermediaries and users. (Mathes, 2004) While intermediary index terms (often subject headings) have been widely promulgated, author keywords and user terminology have tended to be relatively local. In fact, author keywords have received relatively little attention in the literature. (Kipp, 2006; Ansari, 2005; Voorbij, 1998) While intermediaries have been indexing documents for some time, the development of large scale user created collections of tagged documents is new.

This leads one to ask if user categories are indeed different from subject headings or author keywords and if so, how they differ? Are there differences in context, type, or some other semantic relationship? If so, it could be quite important to examine the differences between these categories and the reasons that they do not appear in traditional classification systems. Perhaps these categories are considered to be too short term, too user centric or too subjective to be included? Terms such as @toread and cool after all, do not describe the aboutness of a document and would seem to be of little use in the organisation and retrieval of information. Yet, they are an important part of the phenomenon of tagging. (Kipp, 2007) These short term and highly specific tags suggest important differences between user classification systems and author or intermediary classification systems.

Descriptive statistics can be used to make a basic comparison of the indexing practices of each of the three groups involved in the classification of journal articles (users of a document, authors of a document, and intermediaries or indexers of a document). Additionally, a comparison can be made at the level of the assigned metadata itself. Tags can be examined to see how well they fit the aboutness of the document and to see how closely they match the existing descriptors and author keywords already assigned to the documents.

A few studies have made comparisons of different types of keywords. Voorbij (1998)

studied the correspondence between words in the titles of monographs in the humanities and social sciences and librarian assigned descriptors existing in the online public access catalogue of the National Library of the Netherlands. His study used the different relationships in a thesaurus as an indication of closeness of match, beginning with an exact (or almost exact) match, continuing to synonyms, narrower terms, broader terms, related terms, relationships not formally in the thesaurus, and terms which did not appear in the title at all. (Voorbij, 1998, 468) A similar study by Ansari (2005) examined the degree of exact and partial match between title keywords and the assigned descriptors of medical theses in Farsi. She found that the degree of match was greater than 70 per cent. (Ansari, 2005, 414) Both studies suggest that title keyword searching alone and controlled vocabulary searching alone lead to failure to find some articles. However, there is very little research in this area. Consequently, this study continues to examine the question of convergence between tags, keywords and descriptors by exploring the tagging phenomenon as it is growing at CiteULike.

This study builds on a previous study (Kipp, 2006) examining the emerging phenomenon of social bookmarking in comparison to existing structures for organising and classifying information. Articles from library and information science, tagged on CiteULike, were examined for contextual differences in keyword usage between users of social bookmarking sites, authors and indexers. This study found similarities in terminology use and some intriguing differences in context. Users tended to use tags such as 'toread' and 'todo' to indicate their interest in further use or study of an item. Additionally, geographic index terms were generally descriptors and only rarely used as tags. Many tags were related to terms in the formal thesaurus from which the descriptors were located, but were not formally in the thesaurus. This difference was often due to the use of new or emerging terminology, but also to the use of related but different terminology from different areas of a field. (Kipp, 2006)

This study examines the following questions: a) to what extent term usage patterns of user tags and intermediary descriptors suggest a similar (or differing) context between users and intermediaries? b) how do tags assigned to health related articles reveal clues to the information context of the taggers? c) What differences are apparent in tag use between academic and professional journals?

## 4. Methodology

This study builds on previous work (Kipp, 2006) which examined three forms of index term creation originating from three different groups: users of a document, authors of a document and intermediaries or indexers of a document. In Kipp (2006) it was found that while users often did use terms which were directly from the thesaurus used to assign descriptors to the articles, terms were also often similar or related terms which were not formally linked in the thesaurus. The most prominent example was the use of information retrieval versus information seeking (related but distinct areas of research). Additionally, users tended to include personal information management terminology such as 'toread' in their tag sets, but were less likely to include geographic information. (Kipp, 2006) While the findings from the preliminary study showed that there were differences in the way users, authors and intermediaries classified documents, the size of the data set--165 articles--made it difficult to generalise these findings to

larger data sets from other fields. A larger data set, from a different field, which showed similar patterns of term usage and thesaural matches would strengthen conclusions from the earlier study.

Tag data for the current study was collected from CiteULike between January 12, 2007 and January 24, 2007 via a python script. Author keywords and descriptors were collected from on-line journal databases and Pubmed respectively using additional python scripts.

This study examines the use of tags in a health information context by examining articles posted on CiteULike from three medicine or biology related journals: Journal of the American Medical Association - JAMA, Proteins, and Journal of Molecular Biology. Each of these three journals is indexed in Pubmed. Proteins and the Journal of Molecular Biology also have author assigned keywords. All articles from these selected journals, which have been tagged on CiteULike by at least one user, were collected. To ensure that all articles from these journals were collected, the python script was designed to collect under all common variants of their names (e.g. J. Mol. Biol. for Journal of Molecular Biology). These results were parsed to exclude currently untagged articles. (To aid in the location of new articles, CiteULike also provides listings for articles from selected journals that have not yet been tagged.) Posts for 1299 unique articles were retrieved from CiteULike; author keywords (where available) were located using journal websites; and, MeSH subject headings were collected from Pubmed.

Data collected included title, journal name, volume, issue, page numbers, author names, abstract where available, and URLs providing access to the article or its abstract. URLs were collected for each article and automatically separated into categories as potential sources of keywords or descriptors. Digital Object Identifiers (DOIs - http://www.doi.org/) were selected by preference as a source of author keywords for journal articles and Pubmed URLs were used to locate descriptors (in this case MeSH indexing terms).

All articles were then located in Pubmed and on publicly available abstract pages from on-line journal database sites using the URLs collected from CiteULike. Where possible, pubmed URLs and DOI URLs were used directly, otherwise a series of scripts was used to locate pubmed URLs given the DOI, the DOI given the pubmed ID or, in extreme cases, Google Scholar was used to locate articles using the article title and other bibliographic information. A total of 19 items could not be located on Pubmed, via a DOI (all had at least a DOI or a Pubmed ID) or on Google Scholar. These 19 were excluded from the following study. This resulted in a total of 1280 articles for analysis. Since many articles were tagged by more than one user, this resulted in a total of 1802 posts with tag lists for analysis.

| Journal Name | Number of Articles | Number of Posts |
|---|---|---|
| **Journal of Molecular Biology** | 649 | 931 |
| **Proteins** | 434 | 657 |
| **JAMA** | 197 | 214 |
| **Total** | 1280 | 1802 |

**Table 1. Journals with author assigned keywords**

In the end, each article selected for this study had at least 2 sets of keywords assigned by up to three different classes of metadata creators (JAMA does not have author keywords). The data was stored in a MySQL database and preliminary informetrics analysis was done using SQL scripts as suggested by Wolfram (2005). Descriptive statistics and basic informetric data were collected to provide a good picture of the scope of the collected data. Additionally, a sample of highly tagged articles was selected to have its tags and descriptors examined for term usage.

All retrieved articles were analysed using standard informetric techniques to examine the use of user assigned tags in the context of the articles themselves and their Pubmed assigned MeSH index terms. Additionally, data was analysed for term usage and categorised to see what contextual clues users have left behind in their tag use. Data will be compared to previous studies (Kipp, 2006; Kipp and Campbell, 2006) of term usage in social bookmarking tools.

## 5. Results
### 5.1. Users, Tags and Descriptors: Totals

Metadata for a total of 1280 articles was collected from CiteULike. This data set included articles tagged by at least one user from the journals: Proteins, Journal of Molecular Biology and Journal of the American Medical Association (JAMA). Since many articles had been tagged by more than one person, this data set contains a total of 1802 posts.

There were a total of 314 unique users in the data set. Each user posted at least 1 article (max. 94, median 2). A number of articles were tagged by as many as 14 users, but many were tagged by only 2-4 users. A total of 94 of the 314 users tagged only one article. Due to the use of user selected user names and the fact that it is possible to sign up for multiple user names with multiple e-mail addresses, it is not possible to ensure that all 314 unique user names represent 314 unique users.

A total of 1449 unique tags were used in the data set (4289 with duplicates). Some articles were heavily tagged by users, but most had 2 tags (max. 29, min. 1, median 2). Descriptors were more heavily assigned to articles with 2746 unique descriptors in the data set (total 14507). Articles had, on average, 10 descriptors assigned (max. 40, min. 2). This is to be expected as previous studies have shown that users tend to use only 1-3 tags per article (Kipp 2006).

A few tags were extremely popular, occurring frequently throughout the sample. The most popular tags were: protein_structure (140), no-tag (134), and protein (114). Separated by journal, the most popular tags were: docking (Proteins, 85), no-tag (JAMA, 20), and protein_structure (J Mol Biol, 52). No-tag is a system assigned tag which indicates the user did not assign a tag. Thus, it seems that users tagging articles from JAMA do not always assign a tag and may simply be bookmarking their articles

Unsurprisingly, many of the top 10 tags are related to key concepts in biology such as proteins, evolution and DNA or RNA. Tags emphasise the interest in structures in biology covering such issues as protein configurations (protein_structure, folding, protein_folding). The term bioinformatics, while still not an accepted descriptor, was one of the top 10 tags used to denote articles covering the use of computer simulation and modelling to explore concepts in biology.
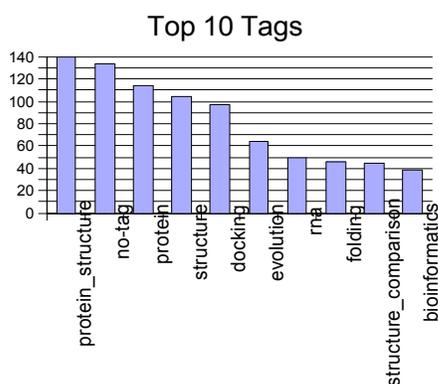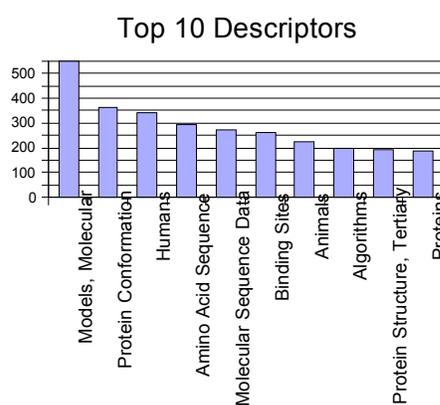


Figure 2. Top 10 tags in the study



Figure 1. Top 10 descriptors in the study

Descriptors were more heavily reused than tags. The most popular descriptors were: 'Models, Molecular', Protein Conformation, and Humans. By journal, the most popular descriptors were: 'Models, Molecular' (Proteins, 252), 'Models, Molecular' (J. Mol. Biol., 385), and Humans (JAMA, 137). Again, terminology use of descriptors was unsurprising, covering important concepts in biology such as molecular modelling (Models, Molecular), proteins and biological structures. Other common descriptors described specific user groups studied and methodologies used in the studies.

### 5.2. Users, Tags and Descriptors: By Article or Journal

User vocabulary length, a measure of how many tags each user used, tended to be short. A few users used many different tags, but most used only a few common tags to describe their posted articles. Per article, the highest number of unique tags used by a user was 18 (min. 1, median 2). In total, the highest number of unique tags used by a single user was 66 (min. 1, median 4). Although most of the users who had a high user vocabulary were heavy posters (having posted more than 25 articles) the user with the highest user vocabulary had only posted 9 articles.

| User | Total | Max/Article | Min/Article | Median/Article | Articles Posted |
|------|-------|-------------|-------------|----------------|-----------------|
| 322  | 66    | 13          | 2           | 8              | 9               |

| User | Total | Max/Article | Min/Article | Median/Article | Articles Posted |
|------|-------|-------------|-------------|----------------|-----------------|
| 1143 | 62 | 8 | 1 | 3 | 94 |
| 1005 | 60 | 8 | 1 | 3 | 65 |
| 3357 | 54 | 6 | 1 | 3 | 34 |
| 1698 | 50 | 9 | 1 | 2 | 34 |

**Table 2. User vocabulary length**

JAMA is a professional journal and the questions arises as to whether there is a difference between academic and professional journals in terms of tag and descriptor usage. In this study, it is obvious that there is a different focus in both the tags and the descriptors used to describe the articles from JAMA versus the other two journals. Popular tags and descriptors used for JAMA articles focus on methodologies and user groups involved in the various studies (family-studies, Male, Female, Aged) while popular tags and descriptors for Journal of Molecular Biology and Proteins focus on elements of biology such as proteins and RNA (protein_structure, docking, Protein Conformation).

| J. of Molecular Biology | | JAMA | | Proteins | |
|-------------------------|-----------|------|-----------|----------|-----------|
| Tag | Frequency | Tag | Frequency | Tag | Frequency |
| protein_structure | 52 | no-tag | 20 | docking | 85 |
| no-tag | 52 | cardiology | 10 | protein_structure | 52 |
| structure | 36 | family-studies | 9 | protein | 51 |
| protein | 36 | mghlcspub | 8 | no-tag | 38 |
| rna | 35 | review | 7 | structure | 34 |

**Table 3. Most popular tags by journal**

| J. of Molecular Biology | | JAMA | | Protein | |
|-------------------------|-----------|------|-----------|---------|-----------|
| Descriptor | Frequency | Descriptor | Frequency | Descriptor | Frequency |
| Models, Molecular | 385 | Humans | 137 | Models, Molecular | 252 |
| Protein Conformation | 268 | Female | 66 | Protein Conformation | 223 |
| Molecular Sequence Data | 212 | Male | 58 | Proteins | 201 |
| Amino Acid Sequence | 204 | Middle Aged | 50 | Algorithms | 151 |

| J. of Molecular Biology | | JAMA | | Protein | |
|---|---|---|---|---|---|
| **Descriptor** | **Frequency** | **Descriptor** | **Frequency** | **Descriptor** | **Frequency** |
| Proteins | 169 | Aged | 41 | Protein Folding | 117 |

**Table 4. Most popular descriptors by journal**

The tag 'no-tag', which is a system assigned tag for articles which have not been tagged, is prominent in the popular tag lists for all three journals. In two cases (JAMA and Journal of Molecular Biology) it is the most popular tag by number and for Proteins it is the 4th most popular tag. This suggests that users often forego tagging in favour of simply storing the article of interest. Since it is still possible to search the collection by author, title and journal name, it is possible that these users do not consider the need to think up useful tags to be worth the time required to do so.

When examined by journal on a per article basis, the highest number of descriptors is 40 for a JAMA article and the lowest 30 for a Proteins article. For tags, the highest number is 29 for a Proteins article and 19 for an article from Journal of Molecular Biology. Of the articles with the highest number of descriptors, 6 of the 10 are JAMA articles. For tags, only 1 of the 10 is a JAMA article. Thus, it seems that users posting JAMA articles do tend to use fewer tags, however, many JAMA articles are highly tagged.

| Journal | Tags | Descriptors |
|---|---|---|
| **JAMA** | 20 | 40 |
| **J. Molecular Biology** | 19 | 36 |
| **Proteins** | 29 | 30 |

**Table 5. Maximum number of index terms per article**

Even with multiple users posting, there are fewer tags assigned than descriptors, but this might not remain true as the number of unique tags assigned will likely grow with the number of users.

## 5.3. Term Use

Comparison of tag lists and descriptor lists shows many of the same similarities and differences as the previous study (Kipp, 2006). Many user terms were found to be related to the intermediary terms but were not part of the formal thesaurus used by the intermediaries and, thus, not formally linked to the intermediary terms. In some cases this was due to faceting of terms for example 'diet' and 'fat' used separately in the tag lists where they were linked as 'dietary fats' in the thesaurus.

| **Tags:** |
|---|
| user1: chd, diet, fat, food, health, heartdisease, lipid, review |

> user2: coronary, diet, disease, heart
> **Descriptors:**
> Coronary Arteriosclerosis, Diet, Dietary Carbohydrates, Dietary Fats, Dietary Fiber, Folic Acid, Humans, Life Style, Lipoproteins

**Figure3. Tags and Descriptors for Article 90: Optimal diets for prevention of coronary heart disease**

Terms such as 'human', 'animal', and 'family-studies' showed that users tagging biology related articles are extremely interested in methodology and user groups associated with articles. This is distinct from the previous study where such terms were more common in the descriptors unless they described extremely specific kinds of methodologies, such as 'pubmed-mining' for data-mining of Pubmed Geographic tags were more common in this study as well with tags such as 'ottawaanklerules' and 'ottawaknee' to specify location. (Kipp, 2006)

## 6. Conclusions

Results from this study show that the results from the initial study do hold true over a different field with a larger data set, this suggests that the results will hold true over the entire set of articles in CiteULike. This study examined the relationship of collaborative tagging to classical classification and indexing by comparing the tags assigned to academic journal articles by users of the CiteULike bookmarking system to library descriptors assigned by intermediary indexers.

Tag and descriptor use in this study showed patterns similar to the tag and descriptor user in the previous study (Kipp, 2006). Users tended to use 1-3 tags on average for each article they posted, although individual users used as many as 18 tags for a single article. Descriptors were more plentiful with a high of 40 descriptors used for a single article. A high user vocabulary length was often related to a high number of articles posted, but this was not a given. A few users with high overall user vocabularies had posted fewer than 10 articles.

Examining the data by separating the professional journal, JAMA, from the other two journals showed that users did indeed use different tags for JAMA articles. Additionally, JAMA articles were not in general as heavily tagged as other articles. Descriptor use did not show any particular pattern unique to JAMA or the other journals in this study.

The articles tagged in this study were all biology or medicine related. While many trends in tagging held true between this study and the previous study (Kipp, 2006), tag users in the current study were more likely to provide geographic tags and use methodology related terms to describe articles as expected.

The popularity of Google suggests that users prefer to be able to search for items in a more natural way using natural language vocabulary and a simple interface. However, users may also express frustration at being unable to locate items or narrow their search results from a huge search set (e.g. 300 000 hits on Google). Controlled vocabularies help to narrow a search set to a manageable size, but controlled vocabulary usage can be expensive. User tagging, which has a lower cost of creation, may help to provide a consensus vocabulary for searching while narrowing the field somewhat from a full natural language vocabulary.

The differing terminology use in tag lists suggests that tagging may be a working example of Vannevar Bush's associative trails. He argued that associative trails better represented how users actually work with their documents: by association rather than by categorisation. (Bush, 1945) This suggests that user tagging could provide additional access points to traditional controlled vocabularies and provide users with the associative classifications necessary to tie documents and articles to time and task relationships, which users find useful, as well as other associations which are new and novel.

## References

Ansari, M. (2005). Matching Between Assigned Descriptors and Title Keywords in Medical Theses. *Library Review* 54(7): 410-4.

Bowker, G.C.; Star, S.L. (1999). *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.

Bush, V. (1945). As We May Think. *The Atlantic Monthly,* July 1945. http://ccat.sas.upenn.edu/~jod/texts/vannevar.bush.html

Cameron, R. (n.d.). CiteULike.org: Frequently Asked Questions. http://www.citeulike.org/faq/all.adp

Eisenstein, E.L. (1983). The Printing Revolution in Early Modern Europe. Cambridge: Cambridge University Press.

Hammond, T., Hannay, T., Lund, B., Scott, J. (2005). Social Bookmarking Tools (I): A General Review. *D-Lib Magazine* 11(4). http://www.dlib.org/dlib/april05/hammond/04hammond.html

Kipp, M.E.I. (2006). Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords. *Canadian Journal of Information and Library Science.* http://dlist.sir.arizona.edu/1533/. (in press)

Kipp, M.E.I. and Campbell, D.G. (2006). Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. *Proceedings of the 2006 Annual Meeting of the American Society for Information Science and Technology, Austin, November 3-8, 2006*. http://eprints.rclis.org/archive/00008315/

Kipp, M.E.I. (2007). @toread and Cool: Tagging for Time, Task and Emotion. 8th Information Architecture Summit, March 22-27, Las Vegas, Nevada.

Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Adammathes.com. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Tang, H. and Ng, J.H.K. (2006a). Googling for a diagnosis -- use of Google as a diagnostic aid: internet based study. *British Medical Journal* 333 (2 Dec. 2006): 1143-1145.

Tang, H. and Ng, J.H.K. (2006b). Authors' reply to responses on: Googling for a diagnosis -- use of Google as a diagnostic aid: internet based study. *British Medical Journal* 333 (2 Dec. 2006): 1143-1145. British Medical Journal 333 (16 Dec. 2006): 1270.

Voorbij, H.J. (1998). Title Keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences. *Journal of Documentation*,

54(4), 466-76.

Wentz, R. (2006). Is Google like 10 000 Monkeys?: Commentary on Tang, Hangwi and Ng Jennifer Hwee Kwoon. 2006. Googling for a diagnosis -- use of Google as a diagnostic aid: internet based study. British Medical Journal 333 (2 Dec. 2006): 1143-1145. *British Medical Journal* 333 (16 Dec. 2006): 1270.

Wolfram, D. (2005). Applications of SQL for Informetric Data Processing. Proceedings of the 33rd conference of the Canadian Association for Information Science, London, June 2-4, 2005. http://cais-acsi.ca/proceedings/2005/wolfram_2005.pdf

Voss, J. (2007). *Tagging, Folksonomy & Co-Renaissance of Manual Indexing?* http://arxiv.org/abs/cs.IR/0701072v2.