User, Author and Professional Indexing in Context: An Exploration of Tagging Practices on CiteULike

Le contexte de l'indexation des utilisateurs, des créateurs et des intermédiaires: une exploration des pratiques d'étiquetage social sure CiteULike

Margaret E.I. Kipp
University of Wisconsin Milwaukee
kipp@uwm.edu

Abstract: This paper examines the context of online indexing from the viewpoint of three different groups: users, authors, and intermediaries. User, author and intermediary keywords were collected from journal articles tagged on CiteULike and analysed. Descriptive statistics, informetric measures, and thesaural term comparison shows that there are important differences in the context of keywords from the three groups.

Cet article examine le contexte de l'indexation en ligne du point de vu de trois groupes : les usagers, les auteurs, et les intermédiaires. Les mots-clés des usagers, des auteurs et des intermédiaires tirés des articles de périodiques ont été étiquetés avec CiteULike et analysés. Les statistiques descriptives et la comparaison des termes thésauraux montrent qu'il existe des différences importantes entre le contexte des mots-clés provenant des trois groupes.

Introduction

Searching a large document space, such as a journal article database, for information is a difficult problem: the sheer size of the space prohibits holistic scanning, while the ambiguities inherent in natural languages make search strategies complex and unreliable. This problem is only exacerbated by the increasing use of digital databases consolidating masses of data. The substantial increase in access to information afforded by the Internet has only strengthened the

importance of being able to, at once, distinguish between similar documents and locate relevant documents. These issues of navigability, resource discovery, and relevance, under the guise of information retrieval and information seeking, have been of importance to the field of library and information science since its inception.

Indexing documents with a subject headings list attempts to resolve this problem; the controlled vocabulary rationalises natural languages by removing ambiguities and consolidating similar items. Many thesauri embed their terms in solidly-designed classification systems, providing useful cross-references that help to reduce the difficulty inherent in searching large spaces for information (ISO 2788:1986; ISO 5964). A key feature of all controlled vocabularies is the entry vocabulary (non preferred terms) which leads to the preferred (or authorised) terms.

While the creation of generic hierarchical classification systems or subject specific taxonomies has a long history, the design of these classification systems has largely been left to professional indexers. Because of the increasing amount and specialisation of information being collected and user requests for greater specificity in indexing terms, these systems can be too generic for user needs. And, while full text search can provide this fine grained access to supplement controlled vocabularies, this access tends to be at the expense of precision due to the use of differing terminology.

The rise of collaborative tagging systems suggests an alternative method for creating classification systems. In fact, such social bookmarking sites are being touted as a potential solution to the problems of scale inherent in the application of any controlled vocabulary to a large document set (Mathes 2004; Hammond et al. 2005; Morville 2005). It has also been suggested that user tags, combined with topic maps and tag clusters, may have the potential to provide the benefits of a controlled vocabulary, which controls for terminological differences,

while still allowing the use of natural language vocabulary (Shirky 2005).

This paper reports on the results of an exploratory study of CiteULike (a social bookmarking service). It examines 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 professional indexers and author keywords assigned by authors to their own journal articles.

Social Bookmarking Tools

CiteULike (http://citeulike.org/) is a social bookmarking service specialised for use by academics who wish to bookmark academic articles for later retrieval. It was created by Richard Cameron in November 2004 (Cameron).

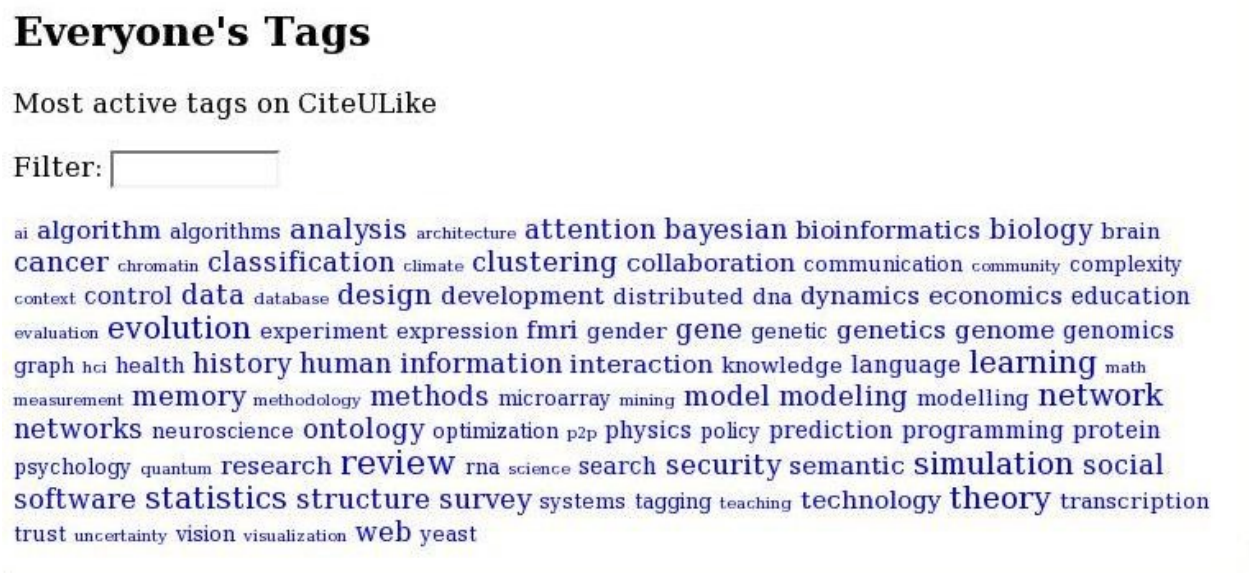

Figure 1: CiteULike Tag Cloud (without CSS style).

Similar to the more commonly known del.icio.us, CiteULike allows users to assign any number of tags to the articles in their library. Users may search by tag (Figure 1) to relocate articles in their own library, as well as in the libraries of other users.

Since CiteULike tags are often associated with journal articles (as opposed to web sites or books), it is possible to collect author keywords and descriptors for many of the articles. Thus, a comparison can be made between user tags, author keywords and professional indexer descriptors attached to a single article.

Related Studies

In order to discover if tags can truly provide a useful replacement or enhancement for controlled vocabularies, it is important to examine whether or not they appear to provide a similar contextual dimension, in terms of coverage of concepts and application of indexing terms to these concepts, to the existing classification systems. While users untrained in indexing are unlikely to produce a complex hierarchical structure on their own, it is possible to examine the tags they do assign to see how they compare to the descriptors assigned by a trained indexer. As well, there is an additional group involved in the creation of this metadata surrounding journals: authors.

Mathes (2004) notes that there are three common groups involved in the assignment of keywords to documents. These groups are authors, professional indexers and users (Mathes 2004). A search of the literature reveals that author keywords have received relatively little attention. And, while professional indexers have been indexing documents for some time, large scale user created collections of tagged documents are relatively new.

Like the hierarchical thesauri created by professional indexers to organise knowledge formally, the new user created folksonomies allow the user to navigate from one topic to another using related links (related terms in a thesaurus). However, relationships in the world of folksonomies include relationships that would never appear in a thesaurus, including the identity of the user (or users) who used the tag (Morville 2005, 137). This phenomenon adds a new

contextual dimension to the act of organising information that is not present in professional indexer assigned keywords, but has been noted by authors studying personal information management (Kwasnik 1991).

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. Additionally, a comparison can be made at the level of the assigned metadata itself. Voorbij (1998) studied the correspondence between, on the one hand, words in the titles of monographs in the humanities and social sciences and, on the other hand, the librarian assigned descriptors existing in the online public access catalogue of the National Library of the Netherlands. His study used a seven point scale of comparison between the title keywords and these descriptors, comparing the descriptors to the title words selected by the author. Voorbij 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 led to failure to find some articles.

Kipp (2005) compared the three user groups involved in indexing (user, author and professional indexer) using a set of articles tagged on CiteULike. Many user terms were found to be related to the author terms. Users terms were also related to the professional indexer terms but were not part of the formal thesauri and, thus, were not formally linked to the professional indexer terms in these thesauri. Other terms were identical to thesaurus terms or part of the entry

vocabulary of the thesaurus itself (Kipp 2005). The results of this study suggested that there was overlap between the three user groups, in some cases potentially sufficient to act as a crosswalk between them, however limitations in the available data suggested that further study using data collected from another field of study would be beneficial.

A few more recent studies have examined tagging as a form of indexing, generating comparisons between tagging and controlled vocabularies on academic social bookmarking tools (Lin et al. 2006; Kipp 2007a; Kipp 2007b; Bruce 2008; Good and Tennis 2008; Good and Tennis 2009; Trant 2009). In addition, a few studies have examined tagging in comparison to the author keywords assigned to some journal articles (Kipp 2007a; Kipp 2007b; Heckner et al. 2008). These studies have shown agreement with the results from Kipp (2005), showing differences between user and professional indexer terminology, but have not in general examined author keywords or compared all three types of terms. Consequently, this study proposes to examine the question of convergence and divergence between tags, keywords and descriptors by continued exploration of the tagging phenomenon as it is growing at CiteULike.

This study, therefore, posed the following research questions:

- To what extent do term usage patterns of user tags, author keywords and professional indexer descriptors suggest that professional indexers are merely engaging in essentially the same activities as authors and users, but merely at a more rigorous, thorough and consistent level?
- To what extent do term usage patterns suggest that authors and users are engaging in a fundamentally different activity, one that cannot be usefully compared or linked to the activities of professional indexers?

These research questions encapsulate the intent of this analysis of the three different user

groups involved in applying aboutness terms to articles, and their differing contexts and term usage.

Methodology

This study examined three forms of index term creation originating from three different groups: users, authors and professional indexers. Data for the study was collected from CiteULike.

Selection of Articles

The selection of articles followed a specific pattern. First, journals that would potentially have all three indexing terms and were from the correct field were selected. Second, journals were located in CiteULike by journal name and user tags were located and collected. Third, author keywords and professional indexer descriptors were collected for all articles located on CiteULike.

Articles for the study were selected from scholarly journals in the field of Library and Information Science which request that authors submit keywords for their articles. These journals were located manually from journal websites and direct examination of sample articles. To ensure that the majority of articles from each of these journals that had been tagged in CiteULike were returned, a search was performed on all common variations and abbreviations of the journal names. CiteULike was chosen for this study as it provides a facility for searching by journal name, something which is not available in similar tools such as Connotea.

Data collected from CiteULike (Figure 2) included the article title, authors, source (i.e. journal name, volume number, etc.), publication date, abstract (where available), userids of users who posted the article and any tags associated with the article. The post and author data for this preliminary study is combined into one set so there are no separate user tag lists and thus no

duplicate tags.

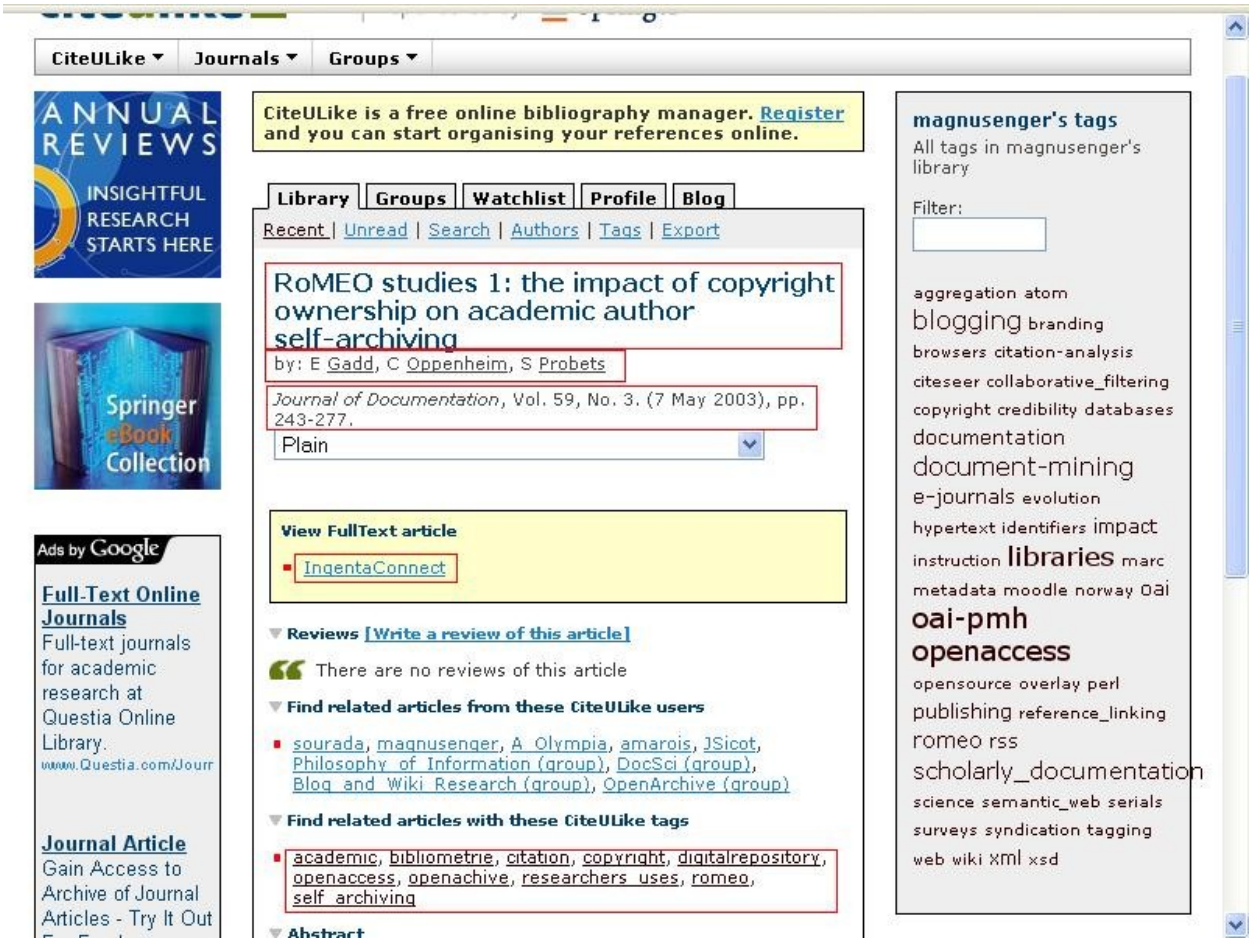

Figure 2: Sample CiteULike post with collected data highlighted.

Author keywords (Figure 3) were collected from online journal databases using the digital object identifier or DOI (http://www.doi.org/) collected from CiteULike.

Figure 3: Sample article metadata with author keywords and DOI highlighted.

Professional indexer terms, in the form of descriptors, for this study were located manually in INSPEC and Library Literature using exact title match (Figure 4). Each of these systems provides professional indexer assigned controlled vocabulary subject headers for searchers. Therefore, each article in this study was represented by three sets of indexing terms.

Figure 4: Sample INSPEC data for an article with descriptors.

A number of measures of analysis were used including:

- Descriptive statistics (including number of posts per user, number of tags per user, number of tags per article),
- Informetrics methods (especially user vocabulary length and an examination of trends in number of index terms used by professional indexers, authors and taggers),
- Term comparison,
- Thesaural comparison.

Term comparison involved direct examination of terms used by each group and

categorisation of terms which did not seem to be directly subject related. Included in this category were methodological terms, geographical terms, proper names and any other term which was not an obvious subject term.

For the thesaural comparison, user tags, author keywords, and professional indexer assigned descriptors were compared based on a 7 point scale, similar to that used by Voorbij (1998) in a study of title keywords. While Voorbij examined descriptor correspondence to title keywords, this study examines the correspondence between all three sets of tags using a structured thesaurus (INSPEC and Library Literature for this pilot study) to generate similarity comparisons. Where possible, comparisons have been done across all three sets of terms, but where the term (or any related term) is lacking from one set, the other two sets were compared against the 7 categories. Comparisons using this 7 category system were done by the author.

The following are the categories as modified.

1. Same - the descriptors and keywords are the same or almost the same (e.g. plurals, spelling variations, acronyms and multiword terms split into facets)
2. Synonym - the descriptors and keywords are synonyms (corresponds to USED FOR in a thesaurus)
3. Broader Term - the keywords or tags are broader terms of the descriptors in the thesaurus
4. Narrower Term - the keywords or tags are narrower terms of the descriptors (like Broader Term, this indicates that the user or author term is in the thesaurus as a broader or narrower term of the associated indexer term)
5. Related Term - the keywords or tags are related terms of the descriptors
6. Related Not In Thesaurus (Related) - there is a relationship (conceptual, etc) but it is not obvious to which category it belongs or it is not formally in the thesaurus

7. Not Related - the keywords and tags have no apparent relationship to the descriptors, also used if the descriptors are not represented at all in the keyword and tag lists

Selection of Field of Study and Journals

For this pilot study, journals were selected from the field of Library and Information Science in order to take advantage of the author's domain knowledge. Journals included in this pilot study include the *Journal of Documentation, Information Processing and Management* and *the Journal of the American Society for Information Science and Technology*. (See Table 1 for the full list.) Journals were initially selected based on prominence in the field of library and information science (measured by Journal Impact Factor), but this selection was expanded to include information science journals with author keywords that were indexed in INSPEC or Library Literature since not all journals have author keywords.

Journal	**Article Count**	**Number of Posts**
Journal of the American Society for Information Science and Technology	68	121
Journal of Documentation	17	39
Information, Communication and Society	6	15
Information Processing and Management	49	80
International Journal of Geographical Information Science	6	8
Information and Organization	4	10
The Information Society	15	24
Total	**165**	**297**

Table 1: Journals with author assigned keywords.

Descriptors were located for articles using INSPEC (Institution of Engineering and Technology, Hertfordshire, UK) or Library Literature (H.W. Wilson Company, New York). Both INSPEC and Library Literature provide professional indexer assigned controlled vocabulary subject headers for searchers and both databases index articles from the field of information science. These online databases were selected for this study as they both index large numbers of

library and information science articles for users working in fields such as information science, library science, information organisation, information retrieval and knowledge management.

Data for the initial study was collected from CiteULike on January 10, 2006, via a python script (citeulike.py). To ensure that all articles from the chosen journals were returned, an exhaustive search of CiteULike was performed examining all common variations of the names of journals in the study, as well as their abbreviations. Using this method, a total of 205 article entries were collected from citeulike.org. Each had been tagged by users of CiteULike with at least one tag. These results were parsed to exclude articles which had not yet been tagged by users as CiteULike also provides access to articles from selected journals which have not yet been tagged. This assists in the location of new material. In this initial study, tags were collected for each article without association to specific users so it was not possible to report data on the number of times each tag had been used per article.

All articles were then located in a publisher journal database (e.g. Wiley InterScience or Emerald) by their DOI or in rare cases by exact title match. Articles for which author keywords could not be located were tagged for review and discarded if descriptors were also not found. This data was also collected using a python script.

Descriptors were included from both sources--INSPEC and Library Literature-- where available. The sets were combined and analysed as a set of terms in the same manner as tags from multiple users were combined to describe an article. Duplicate terms were eliminated. Due to differences in the composition of the indexing terms from INSPEC and Library Literature, precoordinate subject headings from Library Literature were split to generate postcoordinate headings (e.g. Databases – Evaluation became Databases and Evaluation as they would be in INSPEC's thesaurus) and terms matching the INSPEC postcoordinate headings were eliminated

to remove duplicates. This method was adopted since it proved impossible to collect sufficient descriptors from any one source (see Table 2). As there is overlap between the journals indexed by Library Literature and INSPEC and the subjects covered in the two databases it is not unreasonable to expect matches between descriptors used in each databases. Further study using a data set which is consistently indexed in a journal article database would be beneficial to support the findings from this preliminary study.

Journals/Descriptor Sources	**Library Literature**	**INSPEC**	**Both**	**Totals**
Journal of the American Society for Information Science and Technology	29	13	26	68
Journal of Documentation	5	3	9	17
Information, Communication and Society	0	6	0	6
Information Processing and Management	2	26	21	49
International Journal of Geographical Information Science	0	6	0	6
Information and Organization	0	4	0	4
The Information Society	0	15	0	15
Totals	36	74	56	165

Table 2: Sources of Descriptors for the study by Journal and Descriptor Source. Each number represents the number of articles for each journal which was indexed in Library Lit, INSPEC or both.

Exact title match was used to locate articles and associated descriptors manually in the databases since this data could not be collected automatically.

Entries for which author keywords or professional indexer descriptors could not be found

(a total of 40 articles) were excluded manually leaving 165 entries. Thus, each article selected for this study had 3 sets of keywords assigned by three different classes of metadata creators.

Once collected, data from all 165 journal articles were analysed through descriptive statistics, term comparison and thesaural comparison based on a modified version of Voorbij's (1998) 7 point scale. While Voorbij examined descriptor correspondence to title keywords, this study examines the correspondence (similarities and differences) between all three sets of tags using the structured thesauri provided by INSPEC and Library Literature to generate similarity comparisons. Where possible, comparisons have been done across all three sets of terms, but where the term (or any related term) is lacking from one set, the other two sets were compared against the seven categories.

Data analysis was begun with an initial sample of 10 entries. These entries were examined to determine if additional categories would be necessary. Then, the remaining 165 entries were examined to see if there was any evidence of differences in context between user, author and professional indexer metadata as demonstrated by descriptive statistics and term usage.

Results

Authors, Users and Journals

Bibliographic data for a total of 165 articles in information science tagged by at least one user was collected from CiteULike. Some articles had been posted by multiple users resulting in a total of 297 posts.

There were a total of 125 unique user names present in the data. Due to the use of user selected user names and the fact that it is possible to sign up for an account using different email addresses, it is not possible to ensure that these are indeed 125 distinct people.

Each user name was associated with at least one post in the data set. One user (Table 3) had posted 22 articles out of the 165 collected. Most users posted significantly fewer articles (Maximum 22, Minimum 1, Median 1).

Username	Number of Posted Articles
cyrille	22
qaramazov	12
lschiff	11
Enro	11
treatb	10

Table 3: Top 5 Taggers.

A similar drop off can be seen in the data set when examined based on the number of users who have posted a link to a specific article (Table 4). In this case, the maximum number of users per article was 13, the minimum 1, and the median 1.

Number of Users per Article	Article Title
13	Indexing by Latent Semantic Analysis
8	Serendipity and information seeking: an empirical study
8	Indexing and access for digital libraries and the internet: Human database, and domain factors
6	Real life, real users, and real needs: a study and analysis of user queries on the web
6	Information and digital literacies: a review of concepts

Table 4: Number of users who posted a link to a specific article.

Number of users per article	Number of articles
1	109
2	27
3	9
4	11
5	2
6	4
8	2
13	1

Table 5: Number of articles with X users.

In fact, the number of users who posted more than one article dropped off quite quickly (66% or 109 articles were posted only once, median was 1 user per article, see Table 5). This matches findings from citation analysis which show that a few articles tend to be highly cited while many others are infrequently cited (Price 1976).

The majority of the articles in the data set had between 1-3 authors (92.1%), a total of 157 articles, with a maximum of nine authors on one paper. Articles in the data set were tagged by between 1-13 users, with 136 articles (82.5%) having been tagged by 1-2 users.

Tags, Keywords and Descriptors

In the full data set, there were 529 tags, 775 author keywords and 727 professional indexer descriptors. The largest number of tags provided by users for a single article was 21, by authors: 10, and by professional indexers: 12. Over 60% of tagged articles had between 1-3 tags, 4-6 author keywords and 3-5 professional indexer descriptors assigned (Table 6). Despite the potential for a large number of tags assigned by different users, articles did not tend to have a substantially larger number of tags. The two exceptions had 13 and 21 tags and had been tagged by 8 and 13 users respectively. This relatively small number of user assigned tags, compared to the number of keywords assigned by authors and professional indexers, may be due to the small

volume of highly tagged articles in the sample set and the absence of an indexing policy. The majority of articles had been tagged by 1-2 users, although a few articles had been tagged by as many as 13 users.[1]

Number of Index Terms (Tags, Keywords or Descriptors) assigned to an article	Tags	Keywords	Descriptors
1	45 (27.3%)	3 (1.8%)	6 (3.6%)
2	40 (24.2%)	13 (7.9%)	19 (11.5%)
3	29 (17.6%)	26 (15.8%)	40 (24.2%)
4	16 (9.7%)	41 (24.8%)	34 (20.6%)
5	13 (7.9%)	31 (18.8%)	27 (16.4%)
6	5 (3.0%)	27 (16.4%)	11 (6.7%)
7	6 (3.6%)	12 (7.3%)	9 (5.5%)
8	2 (1.2%)	8 (4.8%)	11 (6.7%)
9	4 (2.4%)	1 (0.6%)	7 (4.2%)
10	3 (1.8%)	3 (1.8%)	0
11	0	0	0
12	0	0	1 (0.6%)
13	1 (0.6%)	0	0
21	1 (0.6%)	0	0

Table 6: Number of tags, keywords and descriptors applied to individual articles. Each number in the table represents the total number of articles with 1, 2, 3... 21 index terms assigned (number of index terms is the total number of unique terms). Lines between 14 and 20 have been omitted because they are all 0.

Many tags, keywords and descriptors were found throughout the data set. The most popular tag was 'no-tag', used 18 times followed by 'information' used 15 times. The most

1 Tag data collected for this paper was collected by article and not by user so there are no duplicate tags in the data set for a single article. A tag used more than once by different users will still only appear once in the set. Similar tags (e.g. variations in spelling) are not combined as these are treated separately by the tagging system and are thus examined separately in this preliminary study.

popular keyword was 'information retrieval' used 29 times and the most popular descriptor was also 'information retrieval' used 60 times (Table 7). It is worth noting that many articles tagged information retrieval in the INSPEC database were tagged information seeking by users and even authors so the term information retrieval is not as heavily used in CiteULike.

Tags	**Frequency**
no-tag	18
information	15
information_seeking_behavior	10
internet	9
ir	8

Table 7: Most commonly used tags.

A total of 173 tags were used only once in the data set and 52 were only used twice. Out of a total of 529 tags, 33% were unique.

Keywords	**Frequency**
information retrieval	29
relevance	10
end user searching	9
online searching	8
information seeking	8

Table 8: Most commonly used keywords.

Out of 775 author keywords, 438 (57%) were used only once and 63 were only used twice (Table 8). Thus author keywords appear to be more diverse and less frequently reused than tags or descriptors.

Descriptor	Frequency
Information retrieval	60
Evaluation	53
Internet	40
Use studies	20
Citation analysis	19

Table 9: Most commonly used descriptors.

Of 727 descriptors, 202 (28%) were used only once and 60 were used only twice (Table 9). These results suggest that users and professional indexers may have a higher agreement amongst themselves on which terms to use when indexing articles than authors.

Given the differences in term usage by the three indexing groups, the question arises as to whether there is a relationship between the number of authors and the number of author keywords assigned, or the number of users and the number of tags assigned.

The correlation value obtained when comparing authors versus keywords did not show a significant relationship. This is reasonable as journals request a certain number of keywords per article and thus the number of keywords is not likely to be related to the number of authors. The correlation value for users versus unique tags did show a significant relationship with an R^2 value of 0.654 ($p < 0.05$). This suggests that there is a significant positive correlation between the number of users and the total number of unique tags assigned to an article. The regression equation for the relationship between users and tags is Number of Tags = 1.344 * Number of Users + 0.781. However, it is worth noting that while this result is significant for this data set it is not possible to extrapolate this to the entire data set of articles tagged on CiteULike since it is not a random sample.

An interesting measure for examining term usage in tagging is the measure of user vocabulary length, most often used to analyse search query logs (Wolfram 2005). This measure

represents all the tags used by a specific user in a specific context (for example, all tags used by a user for a particular article). User vocabulary length could not be calculated for all the data because the tag sets were not collected by user but by article; thus tag sets are a combination of the tags used by all users and contain no duplicates. However, an analysis of articles tagged by only one user shows that user vocabulary length was 6 at maximum and 1 at minimum (Table 10). A total of 50 articles tagged by one user were examined.

User vocabulary length	Number of users
6	2
5	2
4	4
3	11
2	12
1	19

Table 10: Number of users with a specific vocabulary length. Only items tagged by one user is considered.

This is higher than the average number of tags assigned by users in other studies (Kipp 2005; Kipp and Campbell 2006; Golder and Huberman 2006) but only 8 of the 50 users used 4 or more tags. The other 42 (84%) used between 1 and 3 tags.

Term Comparison

Acronyms and abbreviations were extremely common in user tags, as were spelling variations. User tag lists tended to contain both spelling variants and plurals of the author keywords and professional indexer descriptors. For example, "communities-of-practice" and "communities_of_practice" were used as tags for the same article, as were "information_seeking_behavior" and "information-seeking-behaviour".

Occasionally, users have provided helpful spelling variations and both long forms and

abbreviations in their tag sets. This situation, though, occurs most frequently when one user tags with abbreviations and another user uses long forms, and, similarly, for spelling variations or plurals. As expected, this phenomenon did not occur in the author keywords or descriptors.

This linkage of terms, which are then all displayed on the articles page, could be extremely useful. INSPEC provides a similar service with its controlled and uncontrolled terms, where the controlled terms will tend to contain the full form of the term and the uncontrolled terms will contain the acronym. For example, the term "GIS" is used by both users and authors, while INSPEC provides "Geographic Information Systems" in its controlled terms and "GIS" in the uncontrolled terms. This apparent duplication would be extremely useful to newcomers to the field or interdisciplinary researchers.

Thesaural Relations

Using the modified version of Voorbij's scale, it was found that the most common relationship discovered in the groups of user, author and professional indexer keywords examined was category 6 or Related Not In Thesaurus (see Table 11). This form of relationship occurred in 133 of 165 articles or 80.6%. The next most common relationship was the Same relationship, where the terms were identical or distinguished only by punctuation or plural forms. This relationship occurred in 103 of 165 articles or 62.4%. Following this was Related Term in 82 articles, Narrow Term and Broader Term combined in 55 articles and Synonym in 46 articles. Not Related terms occurred in 138 of 165 articles or 84% of cases. On average 3.5 Not Related terms occurred per article.

	Same	**Synonym**	**NT/BT**	**RT**	**Related**	**Not Related**
0	62	119	110	83	32	28
1	63	34	39	45	33	0

	Same	Synonym	NT/BT	RT	Related	Not Related
2	26	10	13	18	35	36
3	11	2	2	12	34	30
4	3	0	1	5	14	30
5	0	0	0	1	6	14
6	0	0	0	0	2	15
7	0	0	0	0	7	3
8	0	0	0	0	1	5
9	0	0	0	0	1	4
10	0	0	0	1	0	1
Total Matches (1-10)	103	46	55	82	133	138
Sum by Frequency of Matches (1-10)	160	60	75	152	369	549

Table 11: Frequency of occurrence of the thesaural comparison categories. The left column represents the number of articles with 0, 1,2 ... 10 matches of that type. Each number in the table represents the total number of matches (either binary or trinary) between the three sets of index terms. Note that the sum of matches represents the sum of all matches not the sum of the frequencies. This value is calculated by adding the totals multiplied by the frequency.

In total, there were 549 Not Related terms and 816 matches in the thesaural comparisons. Related Term (RT in a thesaurus) at 152 matches and Same (identical to the descriptor) at 160 matches were the most common of the thesaural comparisons, but combined were less than the 369 matches for the 6th category-- Related Not In Thesaurus. This, and the high number of non matches, suggests that while users often use terminology which is somewhat like that used in a thesaurus, they tend not to use the exact terminology of the thesaurus to describe their work. This tends to reinforce the idea that tagging could be very useful in providing an entry vocabulary to the traditional controlled vocabulary, allowing users the benefits of both systems.

Though thesaural relations were less common, many matches did fall into the Same or

Related Term categories, and some 30% of articles had Narrow Term/Broader Term or Synonym matches as well.

These relationships were less common than the final two non thesaural categories, covering the Related Not In Thesaurus and Not Related categories respectively. In total, the thesaural relations accounted for 447 matches out of 816 total matches or 55% of all matches. This includes the equivalence category (Same), Synonyms, Broader Terms, Narrower Terms and Related Terms.

While Voorbij's initial study examined matches between two items (binary comparisons), this study examined matches between three items (trinary comparisons) where possible as well. Binary comparisons were more common than trinary comparisons. In total there were 618 binary matches versus 198 trinary matches (Table 12). The most common trinary relationship was related but not in the thesaurus, as might be expected. This was also the most common binary relationship.

	Binary Matches	**Trinary Matches**	**Total Matches**
Same	145	15	160
Synonym	44	16	60
Narrower or Broader Term	53	22	75
Related Term	98	54	152
Related	278	91	369

Table 12: Comparison of binary versus trinary matches.

The number of comparisons per article was somewhat dependent on the length of the term lists for tags, keywords and descriptors. An article with a higher number of tags, keywords

and descriptors would have a higher chance of having a larger number of matches and would also likely have more non matches.

	Binary Matches	Trinary Matches
Same	4	2
Synonym	2	3
Narrower or Broader Term	3	4
Related Term	10	3
Related	7	5
Not Related		10

Table 13: Maximum number of occurrences of each match per article.

The maximum number of occurrences of specific matches show, again, that binary matches are generally more common than trinary matches (Table 13).

The maximum number of matches of any kind per article was 15, the minimum 1 and the median 5.

While trinary matches involved an index term from each of the three user categories, binary matches only involved terms from two of three categories. One question that is worth asking is whether one form of binary match was more common than the others.

	User/professional indexer		Author/professional indexer		Author/User	
	Raw	**Percent**	**Raw**	**Percent**	**Raw**	**Percent**
Same	14	0.16	62	0.19	69	0.33
Synonym	13	0.15	16	0.05	15	0.07
Narrower or Broader Term	12	0.14	25	0.08	16	0.08
Related Term	13	0.15	62	0.19	22	0.11
Related	35	0.4	157	0.49	86	0.41
Totals	87	1	322	1	208	1

Table 14: Comparison of number of binary matches between user/professional indexer, author/professional indexer and author/user.

While author/professional indexer matches were most common overall, when normalised it proved to be author/user matches in the related category that were the most common of the thesaural matches (Table 14). User/professional indexer matches were more likely to be thesaural matches while author/professional indexer matches were less likely to be thesaural matches. One potential limitation of this study is that it is impossible to ensure that items tagged by only one person have not been tagged by the article author. Since author/users matches are the most common category of thesaural matches, there remains a possibility that users tagging articles may in some cases actually be the authors of the articles in question. This becomes an issue since authors may have an incentive to promote their articles on CiteULike, an issue which would not occur in a traditional journal database. However, it remains impossible to match a CiteULike user name to the name of an author of an article.

Related Tags

Many relationships fell into the 6th category (45%) -- related but with some ambiguity in

the relationship. This category included relationships that were ambiguous or difficult to fit into categories 1-5, as well as relationships that were not formally listed in the thesaurus but suggested by user tags, author keywords, or INSPEC's uncontrolled terms. Common relationships included: the relationship between an object and its field of study, the relationship between two fields of study which examine different aspects of the same phenomenon, and the use of a methodology or form of inquiry in a new environment.

One of the most common examples of differing terminology choice was the use of "information seeking" and "information retrieval" to refer to the same articles. While these two areas of research examine different aspects of the same phenomenon (finding information), they are considered separately in information science literature. In INSPEC's thesaurus, "information seeking" is not a descriptor, but it is often used in the uncontrolled terms since these terms are taken from the document itself, including the title and abstract (Institution of Electrical Engineers, 18). Since it is not a controlled term, "information seeking" related articles tend to be tagged as "information retrieval" in INSPEC, while authors and users are more likely to tag them as "information seeking." Although Library Literature, the other source of professional indexer descriptors, does make the distinction between "information seeking" and "information retrieval" not all articles in the study were indexed in this database.

Another example of a non thesaural relationship between terms is the relationship between "knowledge" and "knowledge management." Authors and users frequently use the term "knowledge" in their keywords and tags while the professional indexer descriptor "knowledge management" is used by INSPEC. This relationship is not one of equivalence, narrower or broader term, but there is a relationship between the two as knowledge management is the field of study concerned with the organisation and processing of organisational knowledge so that it

can be located and reused.

An example of the use of a methodology or form of inquiry in a new environment is the use of the terms "link analysis" and "citation analysis" to describe the study of the relationships between web hyperlinks. While citation analysis has a long history in library and information science, and the term citation analysis is an INSPEC descriptor, link analysis or hyperlink analysis is a relatively newer field examining a similar phenomenon (references to other articles or sites) in a new environment. Combining the terms "citation analysis" and "Internet" or "web" would serve the same function as the term "link analysis" but the combined term allows users to be more specific without adding terms. This inclusion of newer terms in the user tags can happen faster than in a traditional thesaurus, as one of the goals of a thesaurus is to reproduce the accepted state of knowledge in a field, which leaves the leading edge of the field time to determine standard terminology that will eventually be added to the thesaurus.

Unrelated Tags

Tags, keywords and descriptors falling into the 7th category (Not Related) tended to fall into six basic types: time and task management tags, geographic descriptors, specific details and qualifiers, generalities, emergent vocabulary and other. Since the author of this paper does not want to presume that the thesaurus is inherently superior in its indexing, descriptors that did not match any terms used by the author or users were also placed in this category.

Time and task management terms

The most common time and task management tags were "todo" (7), "new" (7), "print" (4), and "maybe" (3). Tags such as "todo," "maybe" and "new" suggest that users wish to be reminded of the item but have not yet read or not yet decided what to do with it. This appears to be the electronic equivalent of a stack of articles to be read. This type of tag is not represented in

either author keywords or professional indexer descriptors because it is not thought to have value to anyone outside the individual assigning the tag. These tags also tend to have a short lifespan and so would require frequent updating of entries in a database or OPAC. Additionally, they tend to be user or small group specific. However, Amazon has shown that such tags can have value. Wishlists and recommender systems ("people who bought this book also bought these other things") can help people to find new and interesting items by following the purchasing and viewing trails of people who read and enjoy similar material. This suggests that scholars might well find a todo or toread tag useful if they find another scholar who is reading similar material, as suggested by the creator of CiteULike (Cameron). It is worth noting here that a specific toread tag did not turn up in the sample, but this information is encoded in the stars located in the article entries and is requested separately on the article entry form using a scale ranging from "Top priority" to "I don't really want to read this" (http://citeulike.livejournal.com/6890.html).

Another time management tag located in the unrelated category was "lis510" which looks like a course code. This is another example of a time or space sensitive tag which would presumably be of little use to anyone not teaching or taking the course. However, this tag could be extremely useful in an academic library where users could then search the catalogue for books and articles the professor has marked for the course.

Geographic and Personal Terms

Geographic tags, as previously indicated, were found mainly in the descriptors. This suggests that professional indexers are more likely to consider the geographic locations associated with the article to be relevant to the subject of the article. In the case of a copyright related article tagged as "copyright, openaccess, romeo", the addition of the descriptor "Great Britain" would be extremely useful to a user searching for copyright related articles since

copyright law varies greatly depending on country of origin. However, it is quite understandable that the users tagging this article did not consider this to be as important as the tags they actually used since this would, presumably, already be known to them. Another example of this phenomenon was a study of library students in Turkey in which the descriptor "Turkey" was not included in either the author or user tags. Only four examples of geographic tags were found in user or author keywords, two referring to Internet policy in developing countries ("brasil") and another two referring to the location of the authors of the article ("Berkeley"). Interestingly, these user tags were assigned where the descriptors failed to cover geographic location.

Specific Details and Qualifiers

Another category of unrelated terms consists of specific details of the systems or user groups studied, qualifiers and methodologies. Surprisingly, the majority of these terms occurred only in the professional indexer descriptors and did not appear in user or author keywords. Examples of these keywords included "College and university students," the specific group studied in the article, "medical information systems," the specific type of information system used in the information seeking study, and "surveys," representing the specific investigative method used in the tagged article.

The lack of such identifiers in many user and author tagged studies suggests that, for example, both users and authors appear more interested in indicating that the article is about information seeking rather than about information seeking in a specific environment. Interestingly, the type of specific qualifiers used by users tended to refer to specific parts of the content of the article. For example, the term "web-graph" for a webometrics study was used to indicate that the article contains an application of graph theory to the topology of web links, while "pubmed-mining" indicated an article involving data mining from Pubmed and Medline.

One additional area where users added specific tags was for the names of the authors of the paper. This was uncommon and only occurred 3 times in the data set.

Generalities

Comparable to the Specifics category, another category of unrelated items was Generalities. This category consisted of extremely general terms that could apply to almost any article in a field. Examples of this included the terms: computers, libraries/library and information. This is not wholly unexpected as tagging systems lack a predesigned hierarchical thesaurus to provide access to broader or narrower terms. As a result, users of tagging systems have to provide any terms they consider relevant, including terms that might be considered too general to provide good distinction from other articles in the field.

Emergent Vocabulary

Emergent vocabulary was another category found in the unrelated tags. Two prime examples of this phenomenon relate to the topic of this paper. The terms "folksonomy" and "tagging" have been used in this data set to tag articles related to online cataloguing efforts. While the term tagging is not new, its use in this context is somewhat new, replacing the term labelling. The term folksonomy was introduced recently into the vocabulary by Thomas Vander Wal to indicate a collaboratively developed taxonomy (Vander Wal 2007).

Other

The most commonly used tag in this category was "no-tag", which occurred 18 times in the data set. This turned out to be a system created default tag assigned to entries when the user has not assigned a tag. As such, it does not provide any useful information about the contextual aboutness of the document for the user, although it does show interest in the document. It occurs in combination with other tags when multiple users have tagged the same document or if the

original user neglects to remove it when editing the entry to add tags. This tag functions rather strictly as a bookmark and is one way for users to identify an article without having to commit to a specific category of aboutness or interest in the article.

Also in the Other category were two foreign language tags: "etsint_prosessit" and "Relevansvurdering". The term etsint_prosessit appears to be Finnish for search processing or query processing (via AltaVista Babelfish). The article in question was also tagged as "searchprocessing" by another user. Relevansvurdering appears to be Norwegian, with vurdering referring to an appraisal, appraisement, assessment, evaluation, judgement, or judgment. If relevans is relevance then this also matches a tag given by another user. Non English keywords were extremely rare in this data set. There were only three and two were duplicates of etsint_prosessit.

This suggests that currently many users of large scale social bookmarking systems such as del.icio.us or CiteULike are English speaking or use English as a language of correspondence.

Discussion

This study demonstrates that there are differences between the user, author and professional indexer views of the concept space of the articles analysed. While professional indexers considered geographic location to be an important part of the description of the aboutness of an article, authors and users tended to assume it was somewhat less important than the other contexts of the articles. In many cases this may be true. For example, the difference between an information retrieval study performed in the United Kingdom and one performed in the United States is probably not significant due solely to the difference in geographic location.

A comparison of the use of single word and multiword indexing terms could be of interest, but is somewhat hampered by the requirement that a CiteULike tag be a single word.

Many users have chosen to use hyphens or underscores to allow the use of multiword tags in a single word and others have simply removed the spaces from multiword groupings. The frequency of occurrence of such multiword groupings is generally due to the lack of a single term in English to denote the subject, but may also be related to familiarity with traditional multiword library subject headings as opposed to faceted classification systems. In faceted classification systems, core concepts are assigned separately to an item and can be combined in an ad hoc fashion to fully describe the aboutness of a document. Many tag sets presented examples of both a reliance on traditional multiword subject headings and an attempt to build a faceted classification system.

Users considered time management information to be important as a tag for articles. They wanted to encode information about their desire to read the article into the tags for easy access. This is seen in the use of tags such as "todo" and "maybe", as well as in the use of the "toread" interface provided by CiteULike when entering articles into the system. These terms suggest that users may be interested in codifying relationships that are outside the boundaries set by traditional thesaural categories, relationships which may fit Vannevar Bush's associative trails in the memex (Bush 1945).

Many user terms were found to be related to the author and professional indexer terms but were not part of the formal thesauri used by the professional indexers and, thus, were not formally linked to the professional indexer terms in these thesauri. In some cases, this was due to the use of broad terms which were not included in the thesaurus such as information, knowledge or computers. In many cases, this was due to the use of newer terminology or to differences in approach to a problem (information seeking versus information retrieval).

Users were much more likely to have provided a word which was a synonym, or actually

used in the thesaurus, rather than a strict NT/BT, RT relationship. Many user terms fell into the Related category meaning they might qualify as an entry vocabulary to the stricter controlled vocabulary or provide evidence of the use of the article in fields of study not envisioned by the author or original indexer. However, care by the indexer to provide sufficient coverage of the article can help to alleviate the problem; INSPECs uncontrolled tags are useful this way.

Conclusions and Future Work

Although categorisation, description and classification are ubiquitous human activities with deep roots in both cognition and culture, large document collections have traditionally relied on a professionalised version of these activities. Professional intermediaries in the form of cataloguers and indexers classify and describe the documents according to strict standards of term consistency (describing everything in the same way, to enhance recall) and adherence to a set of policy decisions, set out in the texts of cataloguing standards, classification systems, and controlled vocabularies. This consistency of terms is usually accompanied with an entry vocabulary, in the form of lead-in terms, that guides users from terms they might use to the ones used within the system. Controlled vocabularies; however, require training to use and are expensive to apply.

Collaborative tagging systems such as CiteULike allow users to participate in the classification of journal articles. These systems provide an intriguing conduit between professional classification and the innate, ubiquitous categorisation activities common to all humans. Adam Mathes and others suggest that user classification systems would allow librarians to see what vocabulary users actually use to describe concepts and that this could then be incorporated into the system as entry vocabulary to the standard thesaurus subject headings (Mathes 2004; Hammond et al. 2005; Morville 2005) or allow items which had previously been

outside the mandate of a library or indexing service to be categorised.

This study indicates that some of the differences between user tagging and professional indexing are mere differences of wording that can be bridged through algorithms using truncation or stemming. In other cases there are similar principles of aboutness and indexing practice, but with vocabulary that differs from the professional vocabulary, or shows variations in indexing exhaustivity. Many tagging categories have been considered too short term to be relevant, but as Shirky points out, East Germany was a short term category which was used in many library catalogues (Shirky 2005), and which should continue to be used in order to provide access to material from the era of East Germany.

However, time and task-related tags and affective tags indicate principles of indexing that are significantly different from those traditionally used in libraries (Kipp 2007a) where the goal has been to provide general not personal access. These short term and highly personal tags suggest important differences between user classification systems and author or intermediary classification systems which could have implications for system design.

Findings from this study suggest that it would be worthwhile to examine a data set which has been indexed in a consistent fashion in one journal article database using a single controlled vocabulary in order to correct for any possible complications caused by the use of descriptors from different sources. Additionally, further work should examine articles from other disciplines to correct for any bias due to elements specific to a single discipline.

References

Ansari, Mariam. 2005. Matching between assigned descriptors and title keywords in medical

theses. *Library Review* 54, no.7:410-4.

Bruce, Robert. 2008. Descriptor and folksonomy concurrence in education related scholarly research. *Webology* 5, no. 3. http://www.webology.ir/2008/v5n3/a59.html (accessed September 30, 2009).

Bush, Vannevar. 1945. As we may think. *The Atlantic Monthly*, July 1945 http://ccat.sas.upenn.edu/~jod/texts/vannevar.bush.html (accessed September 30, 2009).

Cameron, Richard. no date. CiteULike: Frequently asked questions. http://www.citeulike.org/faq/faq.adp (accessed September 30, 2009).

Golder, Scott, and Bernardo A. Huberman. 2006. The structure of collaborative tagging systems. *Journal of Information Science* 32, no.2:198-209. http://www.hpl.hp.com/research/idl/papers/tags/ (accessed September 30, 2009).

Good, Benjamin M., and Joseph T. Tennis. 2009. Term based comparison metrics for controlled and uncontrolled indexing languages. *Information Research* 14(1), paper 395. http://informationr.net/ir/14-1/paper395.html (accessed September 30, 2009).

Good, Benjamin M. and Tennis, Joseph T. 2008. Evidence of term-structure differences among folksonomies and controlled indexing languages. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, Columbus, Ohio, USA, October 24-

29, 2008. http://www.asis.org/Conferences/AM08/posters/78.html (accessed September 30, 2009).

Hammond, Tony, Timo Hannay, Ben Lund, and Joanna Scott. 2005. Social bookmarking tools (I): A general review. *D-Lib Magazine* 11, no. 4. http://www.dlib.org/dlib/april05/hammond/04hammond.html (accessed September 30, 2009).

Heckner, Markus, Susanne Mühlbacher, and Christian Wolff. 2008. Tagging tagging. Analysing user keywords in scientific bibliography management systems. *Journal of Digital Information* 9, no. 2. http://journals.tdl.org/jodi/article/view/246 (accessed September 30, 2009).

Institution of Electrical Engineers. no date. Inspec on Engineering Village 2.

International Organization for Standardization. 1986. *ISO 2788: Guidelines for the establishment and development of monolingual thesauri.* London: British Standards Institution.

International Organization for Standardization. 1985. *ISO 5964: Guidelines for the establishment and development of multilingual thesauri.* London: British Standards Institution.

Kipp, Margaret E.I. 2005. Complementary or discrete contexts in on-line indexing: A comparison of user, creator and intermediary keywords. *Canadian Journal of Information and Library Science 29*, no. 4:419-436. http://eprints.rclis.org/8379/ (accessed September 30, 2009).

Kipp, Margaret E.I. 2007a. Tagging for health information organization and retrieval. *North American Symposium on Knowledge Organization (NASKO), Toronto, Ontario, Canada, June 14-15, 2007.* http://eprints.rclis.org/11412/ (accessed September 30, 2009).

Kipp, Margaret E.I. 2007b. Tagging Practices on Research Oriented Social Bookmarking Sites. *Proceedings of the 35th conference of the Canadian Association for Information Science*, Montreal, QC, May 10-12, 2007. http://www.cais-acsi.ca/proceedings/2007/kipp_2007.pdf (accessed September 30, 2009).

Kipp, Margaret E.I. and Campbell, D. Grant. 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/08315/ (accessed September 30, 2009).

Kwasnik, Barbara H. 1991. The importance of factors that are not document attributes in the organisation of personal documents. *Journal of Documentation* 47, no. 4:389-398.

Lin, Xia, Joan E. Beaudoin, Yen Bui, and Kaushal Desai. 2006. Exploring characteristics of social classification. *Proceedings of the 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*, 1-19.

Mathes, Adam. 2004. Folksonomies - Cooperative classification and communication through shared metadata. *Adammathes.com.* http://www.adammathes.com/academic/computer-

mediatedcommunication/folksonomies.html (accessed September 30, 2009).

Morville, Peter. 2005. *Ambient findability.* Sebastopol, CA: O'Reilly.

Price, Derek De Solla. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* 27, no. 5:292-306.

Shirky, Clay. 2005. Ontology is overrated: categories, links, and tags. *Clay Shirky's Writings about the Internet: Economics & Culture, Media & Community.* http://shirky.com/writings/ontology_overrated.html (accessed September 30, 2009).

Trant, Jennifer. 2009. Tagging, folksonomy and art museums: Results of steve.museum's research. Technical Report. http://dlist.sir.arizona.edu/2593/ (accessed September 30, 2009).

Vander Wal, Thomas. 2007. Folksonomy coinage and definition. *Vanderwal.net.* http://vanderwal.net/folksonomy.html (accessed September 30, 2009).

Voorbij, Henk 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, no. 4:466-76.

Wolfram, Dietmar. 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 (accessed September 30, 2009).