Margaret E. I. Kipp
Faculty of Information and Media Studies, University of Western Ontario
mkipp@uwo.ca

Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator and Intermediary Keywords

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 and thesaural term comparison shows that there are important differences in the context of keywords from the three groups.

Résumé: Ce papier examine le context de la création de marquage du point de vu de trois groupes: les usagers, les auteurs, et les intermédaires. Le marquage des usagers, auteurs, et intermédaires venant des articles de revue qui furent taggés sur citeulike fut collectioné et analysé. Les statistic descriptives et la comparaison thésaural démontrent qu'il y a des différences importantes entre le contexte du marquage des trois groupes.

1. Introduction

Searching a large document space for information is a difficult problem due to the sheer size of the space, as well as the ambiguities inherent in natural languages. 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, findability, 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.

Classification and indexing via a hierarchical classification system or thesaurus are common methods of attempting to resolve this problem by using controlled vocabularies to rationalise natural languages by removing ambiguities and consolidating similar items. A solidly designed classification system using terms and keywords appropriate to the context of the intended user can help to reduce the difficulty inherent in searching large document spaces for information.

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 intermediaries. Because of the increasing amount and specialisation of information being collected and user requests for more fine grained access, these systems can be too generic for user needs.

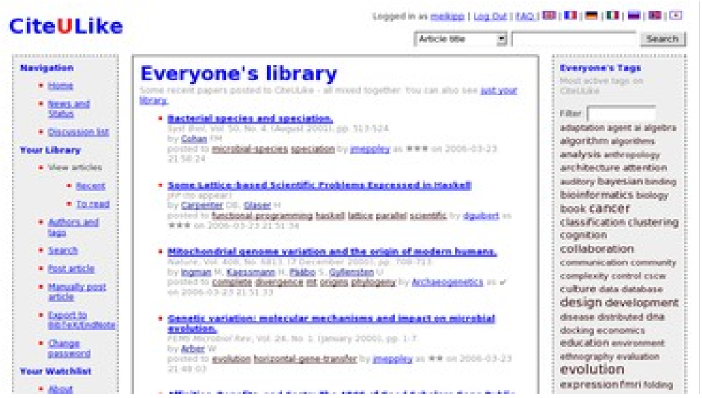
And, while full text search can provide this fine grained access to supplement controlled vocabularies, this access tends to be at the expense of precision due to the use of differing terminology.

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 intermediary indexers and author keywords assigned by authors to their own journal articles.

2. Citeulike

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. (http://www.citeulike.org/faq/all.adp)



Everyone's library

Some recent papers posted to CiteULike - all mixed together. You can also see just your library.

- Bacterial species and speciation.
- Some Lattice-based Scientific Problems Expressed in Haskell
- Mitochondrial genome variation and the origin of modern humans.
- Genetic variation: molecular mechanisms and impact on microbial evolution.

Illustration 1: Citeulike's main page

Similar to the more commonly known del.icio.us, citeulike allows users to assign an arbitrary number of tags to the articles in their library. Users may search by tag 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, 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 intermediary descriptors attached to a single article.

3. 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 to the existing classification systems. While untrained users 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, intermediaries and users. (Mathes 2004) A search of the literature reveals that author keywords have received relatively little attention. And, while intermediaries have been indexing documents for some time, large scale user created collections of tagged documents are relatively new.

Like the hierarchical thesauri created by intermediaries 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 intermediary assigned keywords.

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 lead to failure to find some articles. However, there is very little research in this area. Consequently, this study proposes to examine the question of convergence between tags, keywords and descriptors by exploring the tagging phenomenon as it is growing at citeulike.

This study, therefore, posed the following research question:

- To what extent do term usage patterns of user tags, author keywords and intermediary descriptors suggest a similar context between users, authors and intermediaries? To what extent to they show a differing context?

4. Methodology

This study examined three forms of index term creation originating from three different groups: users, authors and intermediaries. Data for the study was collected from citeulike on January 10, 2006, via a python script (citeulike.py). Articles selected for the study were chosen from scholarly journals whose instructions for authors request author keywords. These journals were discovered manually by examining sample articles and journal webpages. Journals included in this pilot study are all in the field of information science including 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.) To ensure that all articles from each of these 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, to aid in the location of new material.)

All articles were then located in an online journal database by their digital object identifier (http://www.doi.org/) 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. 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 intermediary assigned controlled vocabulary subject headers for searchers and both databases index articles from the field of information science. Exact title match was used to locate descriptors. Where database descriptors were available, but author keywords were not, author keywords were replaced by significant words from the title of the article. There were 10 such articles in the data set. Entries for which author keywords and database descriptors could not be found 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.

Journal	**Article Count**
Journal of the American Society for Information Science and Technology	68
Journal of Documentation	17
Information, Communication and Society	6
Information Processing and Management	49
International Journal of Geographical Information Science	6
Information and Organization	4
The Information Society	15

Table 1: Journals with author assigned keywords

Once selected, all 165 journal articles were subjected to two forms of analysis: descriptive statistics and term comparison. User tags, author keywords, and intermediary assigned descriptors were compared based on a seven point scale, similar to that used by Voorbij (1998). 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. 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
4. Narrower Term - the keywords or tags are narrower terms of the descriptors
5. Related Term - the keywords or tags are related terms of the descriptors
6. 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

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 rest of the 165 entries were examined to see if there was any evidence of differences in context between user, author and intermediary metadata as demonstrated by descriptive statistics and term usage.

5. Results

5.1. Descriptive Statistics

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.

In the full data set, there were 529 tags, 775 author keywords and 727 intermediary descriptors. The largest number of tags provided by users for a single article was 21, by authors: 10, and by intermediaries: 12. Over 60% of tagged articles had between 1-3 tags, 4-6 author keywords and 3-5 intermediary descriptors assigned. 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 had13 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 intermediaries, may be due to the small volume of highly tagged articles in the sample set. The majority of articles had been tagged by 1-2 users, although a few articles had been tagged by as many as 13 users.

	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 2: Number of tags, keywords and descriptors applied to individual

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 tags did show a significant relationship with an R^2 value of 0.645 ($p < 0.05$). This suggests that there is a significant positive correlation between the number of users and the number of tags assigned to an article. The regression equation for the relationship between users and tags is Number of Tags = 1.342 * Number of Users + 0.790. 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.

Using the modified version of Voorbij's scale, it was found that the most common relationship discovered in the groups of user, author and intermediary keywords examined was category 6 or related but not formally in the thesaurus. 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 79 articles, Narrow Term and Broader Term combined in 58 articles and Synonym in 47 articles. Not Related terms occurred in 157 of 165 articles or 95% of cases. On average 3.5 not related terms occurred per article.

	Same	***Synonym***	***NT/BT***	***RT***	***Related***	***Not Related***
0	62 (37.6%)	118 (71.5%)	107 (64.8%)	86 (52.1%)	32 (19.4%)	8 (4.8%)
1	64 (38.8%)	37 (22.4%)	43 (26.1%)	46 (27.9%)	35 (21.2%)	22 (13.3%)
2	29 (17.6%)	8 (4.8%)	13 (7.9%)	18 (10.9%)	41 (24.8%)	27 (16.4%)
3	7 (4.2%)	2 (1.2%)	1 (0.6%)	12 (7.3%)	27 (16.4%)	36 (21.8%)
4	3 (1.8%)	0	1 (0.6%)	1 (0.6%)	17 (10.3%)	27 (16.4%)
5	0	0	0	1 (0.6%)	8 (4.8%)	18 (10.9%)
6	0	0	0	0	1 (0.6%)	14 (8.5%)
7	0	0	0	1 (0.6%)	4 (2.4%)	3 (1.8%)
8	0	0	0	0	0	5 (3.0%)
9	0	0	0	0	0	4 (2.4%)
10	0	0	0	0	0	1 (0.65)

	Same	***Synonym***	***NT/BT***	***RT***	***Related***	***Not Related***
Matches (1-10)	103	47	58	79	133	157
Total: All Matches	155	59	76	134	340	573

Table 3: Frequency of occurrence of the thesaural comparison categories

In total, there were 573 not related terms and 764 matches in the thesaural comparisons. Related Term (RT in a thesaurus) 134 matches and Same (identical to the descriptor) at 155 matches were the most common of the thesaural comparisons, but combined were less than the 340 matches for the 6th category--Related, but not in the 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.

5.2. 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 intermediary 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".

Some 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 the other 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.

5.2.1. Thesaural Relations

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 and not related categories respectively. In total, the thesaural relations accounted for 464 matches out of 764 total matches or 55.5% of all matches. This includes the equivalence category, synonyms, broader terms, narrower terms and related terms.

A comparison of the use of single word and multi word 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.

5.2.2. Related Tags

Many relationships fell into the 6th category (44.5%) -- 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 tended to be tagged as "information retrieval" in INSPEC, while authors and users would most likely tag them as "information seeking." Although Library Literature, the other source of intermediary 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 used the term "knowledge" in their keywords and tags while the intermediary descriptor "knowledge management" would be used by INSPEC. This relationship is not 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 it would 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.

5.2.3. 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.

5.2.3.1. Time and task management tags

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 intermediary 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 the recommender system ("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. (http://www.citeulike.org/faq/all.adp) 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.

5.2.3.2. Geographic Tags

Geographic tags, as previously indicated, were found mainly in the descriptors. This suggests that intermediaries 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.

5.2.3.3. Specifics

Another category of unrelated terms comprises specific details of the systems or user groups studied, qualifiers and methodologies. Surprisingly, the majority of these terms occurred only in the intermediary 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.

5.2.3.4. 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. Users of tagging systems then 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.

5.2.3.5. 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. (http://en.wikipedia.org/wiki/Folksonomy)

5.2.3.6. 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.

6. Conclusions

This study demonstrates that there are differences between the user, author and intermediary views of the concept space of the articles analysed. While intermediaries 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.

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.

Many user terms were found to be related to the author and intermediary terms but were not part of the formal thesauri used by the intermediaries and, thus, were not formally linked to the intermediary 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.

This study has implications for the design of systems for accessing, indexing and searching document spaces. The popularity of Google has demonstrated that users prefer to be able to search for items in a more natural way using one interface to locate items of a varied nature; however, controlled vocabulary usage can be expensive. (Campbell and Fast 2004) Additionally, the evidence of unusual connections between articles, as evinced by such time management tags as "todo," "print," and "new", as well as project specific tags such as "lis510", suggests that this 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)

Thus, user tagging, with its lower apparent cost of production, 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 as well as other associations which are new and novel.

Acknowledgments

Preliminary findings from this study were reported at the Information Architecture Summit 2006 in Vancouver, March 23-27 (sponsored by the American Society for Information Science and Technology). This study was partially supported by the Social Sciences and Humanities Research Council of Canada.

References

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

Bush, Vannevar. 1945. As We May Think. The Atlantic Monthly, July 1945. Retrieved April 10, 2006, from http://ccat.sas.upenn.edu/~jod/texts/vannevar.bush.html

Campbell, D. Grant, and Karl V. Fast. 2004. Panizzi, Lubetzky, and Google: How the Modern Web Environment is Reinventing the Theory of Cataloguing. *Canadian Journal of Information and Library Science, 28*(3), 25-38.

Hammond, Tony, et al. 2005. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine, 11*(4). Retrieved January 11, 2006, from http://www.dlib.org/dlib/april05/hammond/04hammond.html

Institution of Electrical Engineers. no date. Inspec on Engineering Village 2. Retrieved January 11, 2006, from http://www.iee.org/publish/support/inspec/document/UserG/EV2UG.pdf

Mathes, Adam. 2004. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. *Adammathes.com*. Retrieved January 11, 2006, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

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

Shirky, Clay. 2005. Ontology is Overrated: Categories, Links, and Tags. *Clay Shirky's writings about the internet*. Retrieved January 11, 2006, from http://shirky.com/writings/ontology_overrated.html

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*(4), 466-76.