Advancements in Integration between Controlled Vocabularies and Natural Language: An Annotated Bibliography

Valerie Florez
INFO 522-901: Information Access & Resources
December 5, 2010

Introduction and Scope

The digital movement enables end users to conduct their own information retrieval searches, but these searches require a fusion between controlled vocabulary and natural language. As a result, this bibliography reviews the area of research exploring integration of the two. The included papers and articles present a variety of research efforts that target specific user groups as well as different methods of approaching the integration, including logic-based decision trees and mathematical models. A few articles focus on medical user groups, whose circumstances vary considerably from scholarly and general users. Most of the articles originated in the United States, with a small number from Canada and Europe. This topic has been relevant for decades but recently gained momentum due to the digital movement, and the publication dates reflect that: spanning from 1991 to 2010 with most of the articles published in 2000 or later. Cross-language research is excluded from this review; it continues to grow in importance but more due to globalization than the digital movement.

Description

Controlled vocabularies were the predominant indexing method prior to electronic databases because of the limited ability to support free-text searching. Since the dawn of the computer age new ways to index and access documents now exist, and "[free-text searching] is a very powerful retrieval method" (Martin, 1991, p. 22). Natural-language searches appear in almost every nook and cranny of modern information retrieval, especially in the medical, scholarly, and general information realms. The primary goal of integrating indexing terms and free text is to enhance access and increase information discovery (Mendes, Quiñonez-Skinner, & Skaggs, 2009, p. 30). "The natural language interface serves as a uniform frontend in which users can state their information needs, which can then be translated into the protocols required by various databases and knowledge bases" (Johnson, Aguirre, Peng, & Cimino, 1993, p. 294).

Summary of Findings

Most of the research on integrating controlled vocabulary and natural language is semi-quantitative: some comparisons are based on qualitative scales; some comparisons are presented using statistical analysis, but the variables being compared are qualitative in nature; and other comparisons are made via the judgment of subject-matter experts. The main exceptions to this

quasi-qualitative trend are the mapping models that translate free-text queries; most of these are grounded in mathematics. This general area of research has recently turned its focus to social tagging. Prior to the invention of tags, most research focused on the aforementioned mapping models as well as the interplay between free-text searching and controlled vocabularies.

The majority of the research around free-text searching involves the creation of enhanced thesauri. This area of research dates back the farthest because as soon as computer systems became popular, solutions to integrate natural language and controlled vocabulary became viable (Martin, 1991). It was not long after this that mathematical-based models were created to alleviate the problem in a quantitative manner; Chen (1994) determined that "collaboration" between free text and controlled vocabulary can be achieved using a vector space model (VSM) or similar algorithms to create "concept spaces," which are akin to living thesauri. He found that the concept spaces reminded users of terms they had forgotten as well as introduced users to novel terms of which they were unaware but found to be relevant. The research of Knapp, Cohen, and Juedes (1998) focuses on whether to "integrate humanities and social science terms and their free-text synonyms" (p. 412). Their results confirmed that since the subject areas share significant overlap, integration improved the search results. This approach can be translated to other subject areas that also share significant topical overlap.

Once social tagging was introduced to the masses by Delicious (http://www.delicious.com) in 2003, research shifted to study how user-generated input could enhance cataloging efforts. Kipp (2005) compared tags generated by users, authors, and catalogers for a collection of scholarly articles and determined that tags complement cataloging by increasing the number of terms that the user can search on in order to retrieve related information. Lawson (2009) found that catalogers can benefit directly from social tagging by using the tags as a launching pad for improved term selection during subject heading assignment, resulting in "enriched bibliographic records" (p. 580). For a different perspective, Adler's (2009) research focused on award-winning books related to transgender topics. She determined that, in general, controlled vocabulary does not keep pace with social tagging, and thus limits discovery by searchers. This is especially true for topics that are emerging or outside of social norms.

In addition to tagging documents, tagging images has become popular. Yoon (2009) quantitatively analyzed tag names and frequencies against image attributes for the Flickr

database in order to propose an enhanced thesaurus method specific to image collections. She found that such a method differs from those for text-based catalogs, like an online public access catalog (OPAC).

There are many "content creators" (Mendes et al., 2009, p. 32) that add metadata to catalogs, including publishers and catalog vendors. With social tagging, users are now crossing over to become a new type of content creator. Mendes et al. (2009) studied the implementation of LibraryThing for Libraries (LTFL), a system that supplements a library's OPAC with user-generated tags and algorithm-based recommendations. They found in the academic library studied that the social tags enhanced access to information via the user interface, especially for fictional books.

An upcoming area of research under the umbrella of integration focuses on using controlled vocabulary to add structure to user-generated content; this grew out of research around adding structure to natural-language keyword and index searches, like that of Micco and Popp (1994a; 1994b). They found that keyword searches were often inadequate or over-abundant with results, so they designed a system to create "clusters" that led a user from a free-text search term to an indexed term and finally to the records under that indexed term.

With the increasing popularity of social tags, Heymann & Garcia-Molina (2006) took an algorithm-based approach to address the issue of organizing user-generated tags to increase their value to other users. The result was a program to automatically generate a hierarchical index based on user-generated tags; it was most effective on bodies of user tags that had a certain level of similarity and overlap. In the United Kingdom similar research was conducted by Golub, Moon, Tudhope, Jones, Matthews, Puzoń, and Nielsen (2009), but they focused on supplying recommendations from the Dewey Decimal Classification (DCC) system during the tag creation process via an innovative user interface known as EnTag (Enhancing Tagging for Discovery). Their findings were along the line of Chen's (1994): users found the suggestions relevant and helpful. Additionally, they found that through EnTag users gave more thought to tag names, which are often applied hastily–misspellings and all. In a sister study, Matthews, Jones, Puzoń, Moon, Tudhope, Golub, and Nielsen (2010) focused on how authors, instead of users, responded to the EnTag user interface when adding keywords to their scholarly articles for cataloging. Comparing the two studies, the researchers determined that users are more likely to enter tags spontaneously while authors are more likely to consult a thesaurus or list of terms.

Medical information retrieval is usually considered separately due to the uniqueness of medical terminology and the existence of standards like the Unified Medical Language System (UMLS). This separation extends to the arena of free text and controlled vocabulary integration. Research in this area tends to be mathematical and fastidious due to the accuracy needs of the field. Several researchers have focused on mapping natural-language searches to medical indexes, often based on UMLS. One approach concentrates on optimizing the user interface: Johnson, Aguirre, Peng, and Cimino (1993) developed A QUery Analyzer (AQUA), a successful natural language processing program that considers the lexicon up front and uses semantics to determine how the word will be mapped to the UMLS Metathesaurus. During development they determined that UMLS has some weaknesses with respect to verbs that hinder natural language mapping. Similar but more recent research by Gault, Shultz, and Davies (2002) focused on mapping to the Medical Subject Headings (MeSH), which have a more narrow/limited scope than UMLS. They concluded that alphabetical mapping was the least successful and enhanced mapping, like cross-referencing with UMLS in addition to MeSH to setup a synonym-mapping structure, was the most successful. Medical information database vendors can optimize user interfaces based on these findings.

Another approach to improving mapping from natural language to medical-specific controlled vocabulary is through mathematical mapping models. Using VSM similar to Chen (1994), Wenlei and Wesley (2006) determined that a phrase-based VSM is more successful at returning relevant results from free-text medical searches than the more common stem-based VSM. They also found that certain limitations exist with this model due to the elaborate calculations necessary and the sizeable document databases. J. Sun and Y. Sun (2006) investigated how to automatically map various local, semi-standardized medical information databases to the gold-standard UMLS using the Lexical INtegrator of Concepts (LINC). While LINC's novel DNA-structure mapping approach improved mapping accuracy, the system could not be completely automated and still required a review of mappings by a medical expert.

Research on the topic of natural language and controlled vocabulary integration has progressed steadily over the past few decades, but the extraordinary explosion of the Internet and online information access into almost every aspect of modern life warrants a surge in this type of research. End users (general, academic, and medical) have much more information at their fingertips and require tools to assist them with finding what is relevant and accurate. Based on

the findings of this bibliography, future research will probably focus heavily on the inclusion of new types of metadata, including user content, into existing information databases as well as the creation of new information systems using innovative cataloging. Since these areas are grounded in information technology, new possibilities will develop as digital and computational advances continue.

Bibliography

Adler, M. (2009). Transcending library catalogs: A comparative study of controlled terms in Library of Congress subject headings and user-generated tags in LibraryThing for transgender books. *Journal of Web Librarianship, 3*(4), 309-331.

Abstract: "Perhaps the greatest power of folksonomies, especially when set against controlled vocabularies like the Library of Congress Subject Headings, lies in their capacity to empower user communities to name their own resources in their own terms. This article analyzes the potential and limitations of both folksonomies and controlled vocabularies for transgender materials by analyzing the subject headings in WorldCat records and the user-generated tags in LibraryThing for books with transgender themes. A close examination of the subject headings and tags for twenty books on transgender topics reveals a disconnect between the language used by people who own these books and the terms authorized by the Library of Congress and assigned by catalogers to describe and organize transgender-themed books. The terms most commonly assigned by users are far less common or non-existent in WorldCat. The folksonomies also provide spaces for a multiplicity of representations, including a range of gender expressions, whereas these entities are often absent from Library of Congress Subject Headings and WorldCat…"

Annotation: This article highlights how cataloging acts as a representation of social norms instead of recent diversification. Such an approach reduces discovery by searchers because the controlled vocabulary lacks robust cross-referencing. This is particularly detrimental to groups not adequately represented by the Library of Congress Subject Headings (LCSH), like the transgendered as discussed in this article. One approach to offsetting this bias is using folksonomies and social tags in conjunction with LCSH to present emerging concepts.

Search Strategy: This was my first search, so I used Dialog and chose the INFOSCI OneSearch category to ensure comprehensive coverage of Library and Information Science databases. I chose a keyword approach because I was new to the topic and wanted to cast a wide net. To narrow down the keyword results, I ranked them by descriptor and then focused on the most applicable descriptor (Information Retrieval). I accessed the full text via an ILLiad request.

Database: ERIC [Dialog]

Method of Searching: Keyword searching

Search String: SS librar? AND (control?()vocabular? OR natural()language? OR metatag? OR tag?)
RANK DE
VIEW 10/3,ab/1-20

Chen, H. (1994). Collaborative systems: Solving the vocabulary problem. *Computer*, *27*(5), 58-66.

Abstract: "Vocabulary differences have created difficulties for on-line information retrieval systems and are even more of a problem in computer-supported cooperative work (CSCW), where collaborators with different backgrounds engage in the exchange of ideas and information. We have investigated two questions related to the vocabulary problem in CSCW. First, what are the nature and characteristics of the vocabulary problem in collaboration, and are they different from those observed in information retrieval or in human-computer interactions research? Second, how can computer technologies and information systems be designed to help alleviate the vocabulary problem and foster seamless collaboration? We examine the vocabulary problem in CSCW and suggest a robust algorithmic solution to the problem."

Annotation: The use of advanced mathematics, including VSM, along with advances in computer processing capabilities make it possible for larger domains, even entire collections, to utilize a process similar to the one described in this article to create a user-influenced index of terms. There are various algorithms for creating these quasi-thesauri, and the end product is intended to boost discovery by improving the partnership between database designers and end users.

Search Strategy: This discovery was an additional round of the footnote chasing (footnote chasing of a previously-chased footnote). This article expands the discussion around organizing social vocabularies. I was able to search for the article title and journal name via Drexel Libraries online to obtain the full text.

Database: N/A

Method of Searching: Footnote chasing

Search String: Referenced in:
Heymann, P., & Garcia-Molina, H. (2006, April 24). *Collaborative creation of communal hierarchical taxonomies in social tagging systems* (InfoLab Technical Report 2006-10). Stanford, CA: Stanford University. Retrieved from http://heymann.stanford.edu/taghierarchy.html.

Gault, L.V., Shultz, M., & Davies, K.J. (2002). Variations in Medical Subject Headings (MeSH) mapping: From the natural language of patron terms to the controlled vocabulary of mapped lists. *Journal of the Medical Library Association*, *90*(2), 173–180.

Abstract: "**Objectives:** This study compared the mapping of natural language patron terms to the Medical Subject Headings (MeSH) across six MeSH interfaces for the MEDLINE database. **Methods:** Test data were obtained from search requests submitted by patrons to the Library of the Health Sciences, University of Illinois at Chicago, over a nine-month period. Search request statements were parsed into separate terms or phrases. Using print sources from the National Library of Medicine, Each parsed patron term was assigned corresponding MeSH terms. Each patron term was entered into each of the selected interfaces to determine how effectively they mapped to MeSH. Data were collected for mapping success, accessibility of MeSH term within mapped list, and total number of MeSH choices within each list.
Results: The selected MEDLINE interfaces do not map the same patron term in the same way, nor do they consistently lead to what is considered the appropriate MeSH term.
Conclusions: If searchers utilize the MEDLINE database to its fullest potential by mapping to MeSH, the results of the mapping will vary between interfaces. This variance may ultimately impact the search results. These differences should be considered when choosing a MEDLINE interface and when instructing end users."

Annotation: Considering the problem of natural-language searches but focusing on the MeSH and UMLS, Gault et al. detail how database vendors can apply different enhancements to expand term mapping. Considering both UMLS and MeSH when configuring the synonym-mapping structure is one of the more successful approaches. MEDLINE database vendors could use this research to improve their technology, incorporating some of their competitors' successful features (*e.g.*, combination of UMLS followed by an alphabetical search).

Search Strategy: This was my second search, and it began as a search for synonyms as opposed to articles. However, on the first page of results I noticed several scholarly items. I accessed this article through PubMed.

Database: PubMed [via http://www.Google.com]

Method of Searching: Keyword searching

Search String: ["natural language" AND "subject heading"]

Golub, K., Moon, J., Tudhope, D., Jones, C., Matthews, B., Puzoń, B., & Nielsen, M.L., (2009). EnTag: Enhancing social tagging for discovery. *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, 163-172.

Abstract: "The EnTag (Enhanced Tagging for Discovery) project investigated the effect on indexing and retrieval when using only social tagging versus when using social tagging in combination with suggestions from a controlled vocabulary. Two different contexts were

explored: tagging by readers of a digital collection and tagging by authors in an institutional repository; also two different controlled vocabularies were examined, Dewey Decimal Classification and ACM Computing Classification Scheme. For each context a separate demonstrator was developed and a user study conducted. The results showed the importance of controlled vocabulary suggestions for both indexing and retrieval: to help produce ideas of tags to use, to make it easier to find focus for the tagging, as well as to ensure consistency and increase the number of access points in retrieval. The value and usefulness of the suggestions proved to be dependent on the quality of the suggestions, both in terms of conceptual relevance to the user and in appropriateness of the terminology. The participants themselves could also see the advantages of controlled vocabulary terms for retrieval if the terms used were from an authoritative source."

Annotation: The EnTag research project approaches the integration between controlled vocabulary and natural language from the opposite direction of most of the research on this topic: using controlled vocabulary to provide structure to user-generated tags. Offering suggestions from the Dewey Decimal Classification schema improved the quality of tags because users thought critically about their choices before adding them. Unfortunately, context was sometimes lost in translation during these suggestions because the user interface did not show the relationship between a suggested term and its upstream and downstream relatives.

Search Strategy: This is a sister publication to the article by Matthews et al. Upon finishing the Matthews et al. article, I reviewed the footnotes for any leads and found this conference paper. I obtained the full text from ACM Digital Library.

Database: N/A

Method of Searching: Footnote chasing

Search String: Referenced in:
Matthews, B., Jones, C., Puzoń, B., Moon, J., Tudhope, D., Golub, K., & Nielsen, M.L. (2010). An evaluation of enhancing social tagging with a knowledge organization system. *Aslib Proceedings*, *62*(4-5), 447-465.

Heymann, P., & Garcia-Molina, H. (2006, April 24). *Collaborative creation of communal hierarchical taxonomies in social tagging systems* (InfoLab Technical Report 2006-10). Stanford, CA: Stanford University. Retrieved from http://heymann.stanford.edu/taghierarchy.html.

Abstract: "Collaborative tagging systems—systems where many casual users annotate objects with free-form strings (tags) of their choosing—have recently emerged as a powerful way to label and organize large collections of data. During our recent investigation into these types of systems, we discovered a simple but remarkably effective algorithm for converting a large corpus of tags annotating objects in a tagging system into a navigable hierarchical taxonomy of

tags. We first discuss the algorithm and then present a preliminary model to explain why it is so effective in these types of systems."

Annotation: Heavily influenced by mathematical principles, this research develops an iterative process that determines how to position a specific tag within a system's tag hierarchy based on rankings and comparisons. Similar to a small body of related research, this indexing provides structure and promotes thoughtful tag usage. The authors did not include detailed results; instead, they presented a modeling of the results, so it was difficult to gauge effectiveness.

Search Strategy: Since Yoon's article was relevant, I reviewed the references for further reading related to creating structure around user-supplied descriptors. This article's title seemed promising, and the URL to the full text was provided as part of the reference.

Database: N/A

Method of Searching: Footnote chasing

Search String: Referenced in:
Yoon, J. (2009). Towards a user-oriented thesaurus for non-domain-specific image collections. *Information Processing & Management, 45*(4), 452-468.

Johnson, S.B., Aguirre, A., Peng, P., & Cimino, J. (1993). Interpreting natural language queries using the UMLS. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 294-298.

Abstract: "This paper describes AQUA (A QUery Analyzer), the natural language front end of a prototype information retrieval system. AQUA translates a user's natural language query into a representation in the Conceptual Graph formalism. The graph is then used by subsequent components to search various resources such as databases of the medical literature. The focus of the parsing method is on semantics rather than syntax, with semantic restrictions being provided by the UMLS Semantic Net. The intent of the approach is to provide a method that can be emulated easily in applications that require simple natural language interfaces."

Annotation: This symposium paper discusses how the anatomy of a sentence or query can be likened to and modeled by mathematical equations. Examples of user queries and grammatical mappings help explain the theory behind the model. The working model considers the lexicon first and then uses semantics to determine how to compare/map the words to UMLS. The results indicate that the lexicon can successfully interact with UMLS to create natural language mappings for user queries.

Search Strategy: This was my second search, which began as a search for synonyms but returned several scholarly results. After finding a relevant article, I expanded my browsing in PubMed by clicking on articles with

relevant titles in the 'Related citations' section. This article was a related citation of a related citation of the first PubMed article returned in the search.

Database: PubMed [via http://www.Google.com]

Method of Searching: Browsing

Search String: Related to:
Zieman, Y.L., & Bleich, H.L. (1997). Conceptual mapping of user's queries to medical subject headings. *Proceedings of the AMIA Annual Fall Symposium,* 519-522.
Related to:
Gault, L.V., Shultz, M., & Davies, K.J. (2002). Variations in Medical Subject Headings (MeSH) mapping: From the natural language of patron terms to the controlled vocabulary of mapped lists. *Journal of the Medical Library Association, 90*(2), 173–180.

Kipp, M.E.I. (2005). Complementary or discrete contexts in online indexing: A comparison of user, creator, and intermediary keywords. *Canadian Journal of Information and Library Science, 29*(4), 419-436.

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."

Annotation: Comparing tags assigned by users, keywords assigned by authors, and index terms assigned by catalogers determines the interplay between these three primary information-retrieval stakeholders. Inspection of often-disregarded 'personal' user-supplied tags reveals that they add a level of interpretation to scholarly publications that can be utilized by other searchers to enhance information retrieval success. While this article experiences similar contextual losses as described in other research reports, it concludes that authors employ a consistent tag application process whereas users are more haphazard and less consistent.

Search Strategy: This was a repeat of one of my initial searches but in a new database that I learned about in class. I chose a keyword approach because I wanted to keep consistency between the original search and this later search. I retrieved the full-text version of the article from EBSCOhost via Drexel Libraries online.

Database: E-LIS (http://eprints.rclis.org/)

Method of Searching: Keyword searching

Search String: [Full Text matches ""CONTROL VOCABULARY" "NATURAL LANGUAGE"" AND Subjects matches "I. Information treatment for information services" AND Refereed]

Knapp, S.D., Cohen, L.B., & Juedes, D.R. (1998). A natural language thesaurus for the humanities: The need for a database search aid. *Library Quarterly, 68*(4), 406-430.

Abstract: "Database searching presents special difficulties for humanists because many subjects may be covered, many synonyms may be used to describe a single concept, and terms may vary in precision. Databases may be searched by using controlled vocabularies, free-text (natural language) terms, or a combination of both. A significant cause of recall failure in a free-text search is the inability of the searcher to think of all the terms an author may have used. The current study was undertaken to determine the potential value to humanists of a thesaurus integrating free-text terms from the humanities and social sciences. In the first part of the study, a sample of common-noun subject headings from the "Humanities Index" was analyzed to determine how many have at least quasi-synonymous terms. The subject headings were compared to terms in "The Contemporary Thesaurus of Social Science Terms and Synonyms: A Guide for Natural Language Computer Searching" to determine the overlap of terminology between the humanities and social sciences. The results indicate a high degree of overlap, suggesting that a thesaurus integrating terms from the humanities and the social sciences would be of value to scholars in both disciplines. Results also demonstrate that a high proportion of common-noun subject headings have at least quasi-synonymous terms useful for searching. In the second part of the study, searches for humanities scholars were conducted on controlled-vocabulary databases, using both controlled vocabulary and free-text terms to determine whether the latter retrieved additional relevant records not retrieved by the controlled vocabulary. The results indicate that combining both approaches yields more relevant items and higher recall than either method alone…"

Annotation: Knapp et al. make the case in favor of combining humanities and social science thesauri to support the interplay between the two areas of study. The research further reveals that natural language searches are much more common in humanities information retrieval, so a merge should focus on natural-language thesauri as opposed to controlled-vocabulary thesauri. An interesting bias exists in this research because the authors only examine the results from humanities researchers but extend the benefits to social science researchers as well.

Search Strategy: After reviewing one of the articles from the third search string tweak during my first search (in the INFOSCI OneSearch category of Dialog) and determining that it was not relevant enough, I reviewed its references and found this article. I was able to search for the article title and request the full text via ILLiad.

Database: N/A

Method of Searching: Footnote chasing

Search String: Referenced in:
East, J.W. (2007). Subject retrieval from full-text databases in the humanities. *Portal: Libraries and the Academy, 7*(2), 227-241.

Lawson, K.G. (2009). Mining social tagging data for enhanced subject access for readers and researchers. *Journal of Academic Librarianship, 35*(6), 574-582.

Abstract: "Social tagging enables librarians to partner with users to provide enhanced subject access. This paper quantifies and compares LC subject headings from each of 31 different subject divisions with user tags from Amazon.com and LibraryThing assigned to the same titles. The intersection and integration of these schemas is described and evaluated."

Annotation: This research took an approach similar to other studies by comparing cataloger-assigned index terms to user-provided social tags. The scope is much more comprehensive than other studies, though, so the results serve a larger audience. This research offers a unique recommendation for the integration between natural language and controlled vocabulary: The cataloger should consult social tagging to find additional relevant subject headings that may also be more familiar to the user.

Search Strategy: This was my first search, so I used Dialog and chose the INFOSCI OneSearch category to ensure comprehensive coverage of Library and Information Science databases. I chose a keyword approach because I was new to the topic and wanted to cast a wide net. To narrow down the keyword results, I ranked them by descriptor and then focused on the most applicable descriptor (Information Retrieval).

Database: ERIC [Dialog]

Method of Searching: Keyword searching

Search String: SS librar? AND (control?()vocabular? OR natural()language? OR metatag? OR tag?)
RANK DE
VIEW 10/3,ab/1-20

Martin, M.M. (1991). Subject indexing in the new "Ethnographic Bibliography of North America." *Behavioral & Social Sciences Librarian, 11*(1), 13-26.

Abstract: "This paper is concerned with practical and theoretical problems I encountered in creating the Subject Thesaurus for the supplement to the fourth edition of the *Ethnographic Bibliography of North America* and for the forthcoming comprehensive fifth edition of the bibliography. There are several issues involved. The key practical problem is that the bibliography will appear in two formats – print and CD-ROM. Each of these formats requires a

different kind of searching and thus different kinds of indexes. The use of printed indexes requires an index with terms that uniquely identify things while computer searching with software that permits the use of Boolean operators in the search strategy is better served by an index with terms representing classes that can be used alone or in combination. A more theoretical issue involves the question of indexing using natural-language versus index-language terms. This latter issue is related to that of the relative merits of searching in a free-text mode versus searching with the aid of a controlled vocabulary.

I bring these matters to the attention of librarians for two reasons. First, the effective use of the Subject Index of the bibliography will be directly related to understanding the nature of the index and the issues and problems encountered in its construction. Second, it is quite interesting that very practical problems reflecting "classic" theoretical issues surfaced almost immediately in this project. In other words, my experience demonstrates that theory is not dead but is very relevant to everyday practical problems…"

Annotation: This research presents the classic approach of weighing the needs of the user and their potential searches when determining controlled vocabulary. Martin discusses the individualities of catalogers and how that plays a role in the weighting, which is a point not often highlighted in this realm of research. She also displays a heavy preference for post-coordination versus pre-coordination for indexes as the digital movement marches forward.

Search Strategy: This was my third search string during my initial search session; I was still exploring Dialog and the INFOSCI OneSearch category to ensure comprehensive coverage of Library and Information Science databases. I chose a keyword approach due to my newness with the topic, but I focused on joining two main terms to determine the volume and types of information available for the combined concept. I accessed the full text via an ILLiad request.

Database: ERIC [Dialog]

Method of Searching: Keyword searching

Search String: SS control?()vocabular? AND natural()language?
TYPE s7/3,ab/1-20

Matthews, B., Jones, C., Puzoń, B., Moon, J., Tudhope, D., Golub, K., & Nielsen, M.L. (2010). An evaluation of enhancing social tagging with a knowledge organization system. *Aslib Proceedings*, *62*(4-5), 447-465.

Abstract: "Traditional subject indexing and classification are considered infeasible in many digital collections. Automated means and social tagging are often suggested as the two possible solutions. Both, however, have disadvantages and, depending on the purpose of use or context, require additional manual input. This study investigates ways of enhancing social tagging via knowledge organization systems, with a view to improving the quality of tags for increased information discovery and retrieval performance. Benefits of using both social tags and

controlled terms are also explored, including enriching knowledge organization systems with new concepts.
Keywords: folksonomy; social tagging; knowledge organization system; controlled vocabulary"

Annotation: This report discusses a different branch of the EnTag project than its previously-released sister report by Golub et al. This branch of the study utilizes a small number of participants and focuses on authors as opposed to end users. The results include a detailed discussion of observations and behaviors, which the researchers analyze to expand the thinking around possibilities for social tagging.

Search Strategy: This was a repeat of one of my initial searches strings but in a new database of which its relevancy I felt comfortable. I chose a keyword approach because I wanted to keep consistency between the original search and this later search. I obtained the full text from Google Scholar using a title search.

Database: Web of Science

Method of Searching: Keyword searching

Search String: [Topic=("control* vocabular*" AND ("natural language*" OR metatag* OR tag* OR "free text*")). Timespan=All Years. Databases=SCI-EXPANDED, SSCI, A&HCI.]

Mendes, L.H., Quiñonez-Skinner, J., & Skaggs, D. (2009). Subjecting the catalog to tagging. *Library Hi Tech, 27*(1), 30-41.

Abstract: "*Purpose* – The purpose of this paper is to present the implementation of LibraryThing for Libraries (LTFL) in an academic library and analysis of usage of LTFL data and their potential for resource discovery in the catalog.
Design/methodology/approach – The paper reviews the literature on social tagging and incorporation of third-party user-generated metadata into the library catalog. It provides an assessment based on the analysis of total absolute usage figures and frequency of use of LTFL data.
Findings – Based on the data available, usage of LTFL data in the catalog is low, but several possible contributing factors are identified.
Originality/value – The paper contributes to the literature on the implementation of LTFL in an academic library and provides usage statistics on LTFL data. It also provides directions for future research about tagging in the catalog.
Keywords Online catalogues, Tagging, Subject heading lists, Libraries
Paper type Technical paper"

Annotation: This research differentiates itself from others by including a critical examination of a real-world application, installing a community-generated catalog in parallel with an existing OPAC. Mendes et al. elevate the status of the user from a passive participate to an active content

creator. Even though the research and installation are focused on an academic library, the results offer a promising template for non-academic institutions to measure the effectiveness of a collaborative OPAC.

Search Strategy: One of the articles from my initial search mentioned LibraryThing for Libraries. A keyword search for "LibraryThing for Libraries" in INSPEC [Dialog] returned one result, but it was too short to include in this bibliography. When retrieving that article's full text from Emerald Insight, I also browsed the titles of the other articles in the same journal issue. I saw this article's title and retrieved the full text for review.

Database: N/A

Method of Searching: Browsing

Search String: Published in the same special issue of *Library Hi Tech* (Next generation OPACs) as:
Westcott, J., Chappell, A., & Lebel, C. (2009). LibraryThing for libraries at Claremont. *Library Hi Tech, 27*(1). 78-81.

Micco, M., & Popp, R. (1994a). *Developing an information infrastructure to support information retrieval: Towards a theory of clustering based in classification*. Indiana, PA: Indiana University of Pennsylvania.

Abstract: "Techniques for building a world-wide information infrastructure by reverse engineering existing databases to link them in a hierarchical system of subject clusters to create an integrated database are explored. The controlled vocabulary of the Library of Congress Subject Headings is used to ensure consistency and group similar items. Each database becomes a system object, and each package within the database is assigned a subject cluster based on its content. An expert system matches the user profile to the information package best suited to need and locates the appropriate database. This is supplemented by a machine-generated natural language mapping scheme to lead the user into the clusters of interest. For the prototype, an object-oriented hypermedia user interface was developed, using MARC records. Packages are grouped into subject clusters consisting of the classification number and the first subject heading/keyword assigned. Use of a hierarchical classification number (Dewey number) makes it possible to broaden or narrow a search at will. It is anticipated that the system will be useful to searchers and will also provide a basis for automated indexing. Fifteen computer prototype screens are presented as illustrations."

Annotation: This research explores the role of natural language mapping as part of a comprehensive approach to increasing discovery during information retrieval. Micco and Popp use a stepwise method to travel from the controlled vocabulary of subject headings to the free text of user searches. Due to its thorough presentation and relatively early publication date, this paper serves as an anchor point for others on this topic.

Search Strategy: This was my third search string during my initial search session; I was still exploring Dialog and the INFOSCI OneSearch category to ensure comprehensive review of Library and Information Science databases. I chose a keyword approach due to my newness with the topic, but I focused on joining two main terms to determine the volume and types of information available for the combined concept. To narrow down the keyword results, I ranked them by descriptor and then focused on the most applicable descriptor (Information Retrieval).

Database: ERIC [Dialog]

Method of Searching: Keyword searching

Search String:
SS control?()vocabular? AND natural()language?
RANK DE
VIEW 2/3,ab/1-20

Micco, M., & Popp, R. (1994b). Improving library subject access (ILSA): A theory of clustering based in classification. *Library Hi Tech, 12*(1), 55-66.

Abstract: "The ILSA prototype was developed using an object-oriented multimedia user interface on six NeXT workstations with two databases: the first with 100,000 MARC records and the second with 20,000 additional records enhanced with table of contents data. The items are grouped into subject clusters consisting of the classification number and the first subject heading assigned. Every other distinct keyword in the MARC record is linked to the subject cluster in an automated natural language mapping scheme, which leads the user from the term entered to the controlled vocabulary of the subject clusters in which the term appeared. The use of a hierarchical classification number (Dewey) makes it possible to broaden or narrow a search at will."

Annotation: Micco and Popp continue the discussion of their clustering prototype in this article by delving into the details of the software architecture. The program aligns the clusters with the Dewey Decimal Classification system to provide a user-friendly structure that allows easy search refinement. The user interface contains very forward-thinking options considering the time frame of its design, including an interactive library map as well as the ability to submit real-time requests. Regrettably, the article lacks detailed user-testing results, only implying that user testing was conducted.

Search Strategy: This was my third search string during my initial search session; I was still exploring Dialog and the INFOSCI OneSearch category to ensure comprehensive coverage of Library and Information Science databases. I chose a keyword approach due to my newness with the topic, but I focused on joining two main terms to determine the volume and types of information available for the combined concept.

To narrow down the keyword results, I ranked them by descriptor and then focused on the most applicable descriptor (Information Retrieval). This article is closely related to the other Micco and Popp record, but it contains more technical details. I accessed the full text via an ILLiad request.

Database: ERIC [Dialog]

Method of Searching: Keyword searching

Search String: SS control?()vocabular? AND natural()language?
RANK DE
VIEW 2/3,ab/1-20

Sun, J.Y., & Sun, Y. (2006). A system for automated lexical mapping. *Journal of the American Medical Informatics Association, 13*(3), 334-343.

Abstract: "*Objective:* To automate the mapping of disparate databases to standardized medical vocabularies.
Background: Merging of clinical systems and medical databases, or aggregation of information from disparate databases, frequently requires a process whereby vocabularies are compared and similar concepts are mapped.
Design: Using a normalization phase followed by a novel alignment stage inspired by DNA sequence alignment methods, automated lexical mapping can map terms from various databases to standard vocabularies such as the UMLS (Unified Medical Language System) and LOINC (Logical Observation Identifier Names and Codes).
Measurements: This automated lexical mapping was evaluated using three real-world laboratory databases from different health care institutions. The authors report the sensitivity, specificity, percentage correct (true positives plus true negatives divided by total number of terms), and true positive and true negative rates as measures of system performance.
Results: The alignment algorithm was able to map 57% to 78% (average of 63% over all runs and databases) of equivalent concepts through lexical mapping alone. True positive rates ranged from 18% to 70%; true negative rates ranged from 5% to 52%.
Conclusion: Lexical mapping can facilitate the integration of data from diverse sources and decrease the time and cost required for manual mapping and integration of clinical systems and medical databases."

Annotation: This paper represents a growing body of research concerning controlled vocabularies and natural language searches in the medical arena. The proposed mapping model experiences some of the same issues as other such automation attempts, like relevancy and synonymy; but it shows promise as a way to correlate a non-standardized, clinical list of terms to one of the standard lists, like UMLS, due to its innovative matching method. This method is especially helpful for abbreviations, which are common in medical terminology.

Search Strategy: While searching in PubMed, I compared the MeSH terms of two articles selected for this bibliography and combined the most relevant terms into one search statement. This article was returned as one of the search results.

Database: PubMed

Method of Searching: Keyword searching

Search String: ["Information Storage and Retrieval"[All Fields] AND "Vocabulary, Controlled"[All Fields] AND "Unified Medical Language System"[All Fields]]

Wenlei, M., & Wesley, W.C. (2006). The phrase-based vector space model for automatic retrieval of free-text medical documents. *Data & Knowledge Engineering, 61*(1), 76–92.

Abstract: "*Objective:* To develop a document indexing scheme that improves the retrieval effectiveness for free-text medical documents.
Design: The phrase-based vector space model (VSM) uses multi-word phrases as indexing terms. Each phrase consists of a concept in the unified medical language system (UMLS) and its corresponding component word stems. The similarity between concepts are defined by their relations in a hypernym hierarchy derived from UMLS. After defining the similarity between two phrases by their stem overlaps and the similarity between the concepts they represent, we define the similarity between two documents as the cosine of the angle between their corresponding phrase vectors. This paper reports the development and the validation of the phrase-based VSM.
Measurement: We compare the retrieval effectiveness of different vector space models using two standard test collections, OHSUMED and Medlars. OHSUMED contains 105 queries and 14,430 documents, and Medlars contains 30 queries and 1033 documents. Each document in the test collections is judged by human experts to be either relevant or non-relevant to each query. The retrieval effectiveness is measured by precision and recall.
Results: The phrase-based VSM is significantly more effective than the current gold standard—the stem-based VSM. Such significant retrieval effectiveness improvements are observed in both the exhaustive search and cluster-based document retrievals.
Conclusion: The phrase-based VSM is a better indexing scheme than the stem-based VSM. Medical document retrieval using the phrase-based VSM is significantly more effective than that using the stem-based VSM."

Annotation: This research focuses on the highly-mathematical VSM but limits its scope to medical databases and language. This confined approach allows the authors to investigate the VSM on a finer level of granularity than previous research and to tune their model to address some of the specifics unique to the medical lexicon. This concentration, however, dilutes the usefulness of the research to arenas other than the medical community.

Search Strategy: After using the ACM Digital Library database to retrieve the full text of another article, I discovered that it covered publications not covered in the other databases that I had searched. Therefore, I repeated a slight variation of one of my initial searches in this database. I chose a keyword approach and limited the modifications to the keywords because I wanted to keep consistency between the original search and this later search. I was able to obtain the full text via a link from ACM Digital Library to Science Direct.

Database: ACM Digital Library

Method of Searching: Keyword searching

Search String: ["Controlled Vocabularies" ("Natural Language" OR "Free Text")]

Yoon, J. (2009). Towards a user-oriented thesaurus for non-domain-specific image collections. *Information Processing & Management, 45*(4), 452-468.

Abstract: "This study explored how user-supplied tags can be applied to designing a thesaurus that reflects the unique features of image documents. Tags from the popular image-sharing Web site Flickr were examined in terms of two central components of a thesaurus—selected concepts and their semantic relations—as well as the features of image documents. Shatford's facet category and Rosch et al.'s basic-level theory were adopted for examining concepts to be included in a thesaurus. The results suggested that the best approach to Color and Generic category descriptors is to focus on basic-level terms and to include frequently used superordinate- and subordinate-level terms. In the Abstract category, it was difficult to specify a set of abstract terms that can be used consistently and dominantly, so it was suggested to enhance browsability using hierarchical and associative relations. Study results also indicate a need for greater inclusion of Specific category terms, which were shown to be an important tool in establishing related tags. Regarding semantic relations, the study indicated that in the identification of related terms, it is important that descriptors not be limited only to the category in which a main entry belongs but broadened to include terms from other categories as well. Although future studies are needed to ensure the effectiveness of this user-oriented approach, this study yielded promising results, demonstrating that user-supplied tags can be a helpful tool in selecting concepts to be included in a thesaurus and in identifying semantic relations among the selected concepts. It is hoped that the results of this study will provide a practical guideline for designing a thesaurus for image documents that takes into account both the unique features of these documents and the unique information-seeking behaviors of general users."

Annotation: This research quantitatively analyzes the application of social tags in an image database. Such databases are currently a very popular method of sharing and obtaining images, which makes this research timely and relevant. Focusing on general users allows the author to expand on the previous research findings of others, which center primarily on subject-matter experts for creating image thesauri.

Search Strategy: After reviewing one of the articles from my first search (in the INFOSCI OneSearch category of Dialog) and determining that it was not relevant enough, I reviewed its references and found this article. I was able obtain the full text by searching for the article title and journal name via Drexel Libraries online.

Database: N/A

Method of Searching: Footnote chasing

Search String: Referenced in:
Chung, E., & Yoon, J. (2009). Categorical and specificity differences between user-supplied tags and search query terms for images: An analysis of Flickr tags and Web image search queries. *Information Research, 14*(3), paper 408. Retrieved from http://informationr.net/ir/14-3/paper408.html.

Conclusion and Personal Statement

Originally I considered this annotated bibliography project an Iron Man triathlon and I a mere 5K jogger. In all seriousness, the enormity of this project was manageable due to the amazing databases and broad range of access to scholarly literature, in electronic format and through ILLiad. I believe that many current students do not appreciate this compared to what it was like in "the old days" when you expected to visit the stacks to locate a print copy of the journal in order to photocopy the article of interest.

At the beginning of this assignment I was overwhelmed by the amount of information in the various databases (even considering Dialog alone). However, after testing a variety of search terms and synonyms across databases, I noticed that some articles were appearing in multiple result sets. These reappearances instilled confidence in me that I was covering the topic thoroughly and that I was ready to move to the next stage, synthesizing and coalescing. All the while, my fundamental search skills were improving as I learned new techniques from class lectures, readings, and especially the real-world assignments. It is interesting to ponder how my approach would differ if I had not started this bibliography until after completing the INFO 522 course work: what mistakes might I avoid; would my findings change; would my approach be more streamlined?

With an undergraduate degree in Chemical Engineering, most of my research experience (hands-on or literature reviews) was based on lab work and chemical processes. Prior to this project, many aspects of LIS research were new to me; in fact, this was my first experience synthesizing qualitative results. Furthermore, I was unaware of the large body of research specific to medical databases and of the substantial controlled vocabularies specific to the medical language. I now understand the emergence of the Healthcare Informatics discipline.

As a result of this experience, I feel equipped to conduct a similar level of research on another LIS topic and possibly even on a topic outside of LIS.

I certify that:

- This assignment is entirely my own work.
- I have not quoted the words of any other person from a printed source or website without indicating what has been quoted and providing an appropriate citation.
- I have not submitted this assignment to satisfy the requirements of any other course.

Signature Valerie Florez
Date 12/05/2010
