2015

Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching

Tina Gross
St. Cloud State University, tmgross@stcloudstate.edu

Arlene G. Taylor
University of Pittsburgh - Main Campus, ataylor@sis.pitt.edu

Daniel N. Joudrey
Simmons College, joudrey@simmons.edu

Recommended Citation
Tina Gross, Arlene G. Taylor & Daniel N. Joudrey (2015) Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching, Cataloging & Classification Quarterly, 53:1, 1-39, DOI: 10.1080/01639374.2014.917447

Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching

Abstract. In their 2005 study, Gross and Taylor found that more than a third of records retrieved by keyword searches would be lost without subject headings. A review of the literature since then shows that numerous studies, in various disciplines, have found that a quarter to a third of records returned in a keyword search would be lost without controlled vocabulary. Other writers, though, have continued to suggest that controlled vocabulary be discontinued. Addressing criticisms of the Gross/Taylor study, this study replicates the search process in the same online catalog, but after the addition of automated enriched metadata such as tables of contents and summaries. The proportion of results that would be lost remains high.

Introduction

Over the last three decades, it has been acknowledged that online public access catalogs are difficult for patrons to use.[1] Part of this difficulty is related to the complexity of subject searching in the catalog.[2] Part of it stems from patrons becoming more accustomed to Google-like keyword searching. It has been suggested that because a large percentage of patrons start their information seeking by using keyword searches, libraries should discontinue using and maintaining controlled subject vocabularies. Such suggestions have not been viewed favorably by some in the library and information professions, including the Library of Congress Policy and Standards Division (formerly the LC Cataloging Policy and Support Office).[3]

The Working Group on the Future of Bibliographic Control, convened by the Library of Congress to examine current cataloging practices and present findings and recommendations to LC, supported the continued use of Library of Congress Subject headings (LCSH) and other controlled vocabularies in its 2008 report:

> Although there is much speculation that improvements in machine-searching capabilities and the growth of databases eliminate the need for authoritative forms of names, series, titles, and subject concepts, both public testimony and available evidence strongly suggest that this is not the case. While such mechanisms as keyword searching provide extremely useful additions to the arsenal of searching capabilities available to users, they are not a satisfactory substitute for controlled vocabularies. Indeed, many machine-searching techniques rely on the existence of authoritative headings even if they do not explicitly display them.[4]

Despite the objections raised to suggestions that subject headings be abandoned and the ostensible reprieve for LCSH, the future of controlled vocabularies at times still seems precarious.

In response to assertions about the lack of importance of controlled vocabulary in the catalog, Tina Gross and Arlene G. Taylor published a study in 2005 to determine the role that LCSH played in results retrieved through keyword searching. They noted “that some keyword searches retrieve records in which one or more sought-after word(s) is found only in a subject string in a subject-heading field.”[5] This research investigated how often this might occur. They found that “if subject headings were to be removed from or no longer included in catalog records, users performing keyword searches would miss more than one third of the hits they currently retrieve. On average, 35.9 percent of hits would not be found.”[6]

The results were persuasive, but some argued the study might have dramatically underestimated the proportion of hits that would be lost in the absence of subject headings because of the decision to limit search results to English. The authors assumed the proportion to be higher when foreign language materials are included because "the vast majority of bibliographic records for foreign language materials with English language subject headings could only contain many of the English language search terms from the sample in their subject headings," but the study did not actually look at results including languages other than English.

Others dismissed the study's results, suggesting that the addition of tables of contents (TOCs) and summary notes in catalog records could minimize the need for controlled vocabulary. In "The Changing Nature of the Catalog," a 2006 report commissioned by the Library of Congress, Karen Calhoun actually cites the 2005 Gross and Taylor study in the same step of the report's

"ten-step planning process" in which she recommends that libraries "abandon the attempt to do comprehensive subject analysis manually with LCSH in favor of subject keywords" and "urge LC to dismantle LCSH."[7] The corresponding footnote implies that because "automated enriched metadata such as TOCs can supply additional keywords for searching,"[8] the results of the Gross and Taylor study could be safely ignored.

Examination of the issues raised by these criticisms is warranted. Furthermore, dismissals of the study's evidence—based not on criticism of the methodology, but apparently based on viewing the obsolescence of subject headings as a foregone conclusion—raised other questions. Does the available evidence support or contradict this widespread view? What does the body of research say on the matter of whether keyword searching is adequate without the presence of subject headings?

The current study is a follow-up to the 2005 Gross and Taylor research. It looks at the same issues as the earlier study with three major differences. First, it begins with an exhaustive literature review that aims to provide a definitive summary of the past two decades of research on the topic of keywords versus subject headings. Second, the study's research was conducted in the same catalog as the earlier study, but the searching was performed *after* tables of contents had been added to enrich the database. The third difference is that the study looks at search results that included materials in all languages, not just English language materials.

Literature Review

For several decades, research has been carried out on the topic of keywords versus subject headings (or controlled vocabulary). However, no one since Jennifer Rowley in 1994[9] has looked at all this research as a whole with the purpose of determining if there is established

theory as to whether keyword searching is satisfactory without controlled vocabulary. The first research on the topic compared titles with subject headings to determine how many words they had in common. In 1964 Donald Kraft, researching keyword-in-context (KWIC) indexing of titles, wrote: "Interpretation of data revealed, among other things, that 64.4% of the title entries contained as keywords one or more of the … subject heading words under which they were indexed,"[10] which means that just over one third of the titles did not have a match to a subject heading word. Carolyn Frost, comparing title words with LCSH in 1989, found that, "For 27% of the sample, there were no words from the title which matched any part of the subject heading."[11]

In 1992 Barbara Keller looked at bibliographic records for Master's theses and compared the first word of a LCSH heading with words in the title to find how often there would be a match. She found an overlap of 64%, which means that 36% did not match. [12] In a study reported in 1998, Henk J. Voorbij wanted to learn whether the presence of controlled terms led to better results than searching by uncontrolled terms. He asked librarians to judge whether descriptors in a record were the same or almost the same as the title words. He then asked whether addition of the descriptors to the records resulted in enhancements that were "slight" or "considerable." His results showed that 37 percent of the records were considerably enhanced by a subject descriptor.[13]

In 2003, Elaine Nowick and Margaret Mering compared keyword queries with *Library of Congress Subject Headings*, *Water Resources Abstracts Thesaurus*, and *Aqualine Thesaurus 2* and found that "[b]etween 30 percent and 40 percent of the free-text queries were exact matches to a term in one of the controlled vocabularies."[14] Gross and Taylor, as mentioned above, found that 35.9% of hits in keyword searches do not have the keywords anywhere in the records except in the subject headings.[15] In a 2010 study comparing LCSH to keywords in book titles, Caimei

Lu, Jung-ran Park, and Xiaohua Hu found that " [O]nly a minority of books have LCSH terms appearing in the book titles. This is because subject experts intentionally avoid repeating the title in subject terms."[16] These studies have consistently shown that human-supplied controlled vocabulary has added around one third or more of the words that make keyword searching successful.

Prevalence of Keyword Searching

Even though research continues to show the importance of controlled vocabulary, keyword searching has become the most often used, and, in fact, the preferred, method of conducting a search in any online system. OCLC's 2009 evidence-based study of what constitutes "quality" in catalog data states that "[k]eyword searching is king, but an advanced search option (supporting fielded searching) and facets help end users refine searches, navigate, browse and manage large results sets. End users want to be able to do a simple Google-like search and get results that exactly match what they expect to find."[17] The researchers added that "[e]nd users … expect the catalog to 'know' what they are looking for based on the terms they type in the search box. Additionally, if the words they use in their searches have multiple meanings depending on the context, they still expect their searches to return appropriate materials on exactly what they want."[18] However, as Kayo Denda writes, "[t]he relevance and usefulness of controlled vocabularies … in emerging interdisciplinary fields and the suitability of conventional library tools for organizing and accessing digital information are in question."[19]

Recent literature on controlled vocabulary versus keyword searching seems to fall into two groups:

- Successful keyword searching relies on controlled vocabulary as part of a system.

- Controlled vocabulary should be abandoned in favor of keywords.

Relying on Controlled Vocabulary in Keyword Searching

In 2000 Lois Mai Chan stated: "When the searcher's keywords are mapped to a controlled vocabulary, the power of synonym and homograph control [can] be invoked and the variants of the searcher's terms [can] be called up…. [B]uilt-in related controlled terms [can] also be brought up to suggest alternative search terms and to help users focus their searches more effectively. In this sense, controlled vocabulary is used as a query-expansion device."[20] On the other hand, she pointed out, "[s]ubject categorization defines narrower domains within which term searching can be carried out more efficiently and enables the retrieval of more relevant results."[21]

Rebecca Donlan, and Rachel Cooke, in a 2005 article about library licensing of texts through Google Scholar observe that, "Federated search engines depend upon keyword searching, which in turn is only as good as the subject headings used in the databases that are included. All databases are not equal in this respect. Libraries must continue to support quality subject access in the databases to which we subscribe, and librarians must be able to explain why subject analysis is worth the cost…."[22] Donlan and Cooke go on to emphasize the importance of controlled vocabularies: "We need to be able to explain and defend the added value of subject thesauri in the databases for which we pay a considerable percentage of our materials budgets. Otherwise, we cannot blame our funding agencies for thinking that Google is 'just as good.' The irony, of course, is that eventually, Google will not be 'just as good' as those expensive proprietary databases if we stop paying for them."[23]

Jeffrey Garrett, reporting in 2007 on an experiment at Northwestern University Library to add subject headings to online records for the Eighteenth Century Collections Online (ECCO), writes, "users today find what they are looking for by using subject headings not as verbatim search expressions, but as sources for frequently unique keyword material."[24] After citing Gross and Taylor, Garrett states: "The fact is that the assignment of descriptive language in the subject heading fields frequently attaches important terms and concepts to a bibliographic record that the record will not otherwise contain."[25]

An interesting simile is presented by Sue Ann Gardner in her 2008 discussion about how the emerging information environment is impacting cataloging issues. After quoting from Nancy Fallgren's 2007 paper that says, "traditional bibliographic access points of author, title, and subject now constitute a small proportion of the data that can be retrieved with full text keyword searching,"[26] Gardner observes: "Declaring that the traditional access points constitute a small proportion of the data/metadata is like dismissing diamonds because they constitute just a small proportion of the slurry in which they are found. They may represent but a fraction, but they are precious bits."[27]

Oksana Zavalina reports in her 2010 dissertation the results of a study of aggregations of digital collections to determine how collection-level bibliographic records compare with item-level records and to determine how subject access affects success in searching collection level records. Using an adaptation of Gross and Taylor's methodology, she found that subject metadata "provides a significant source of matches to user search terms, with at least one retrieved collection record having a match to a user search term in this field in 50% of searches, and 27% of searches retrieving one or more records with a match exclusively in this field."[28] She also found that "if only the free-text *Description* field is used in collection metadata records, almost

half (41%) of the collections would not be retrieved in response to subject-specific collection searches in aggregation."[29]

Abandoning Controlled Vocabulary

In the last few years, there have been several calls for abandoning traditional controlled vocabulary in favor of relying on free-text searching of bibliographic records. Members of the 2005 Bibliographic Services Task Force of the University of California (UC) Libraries agreed that controlled vocabularies are still valuable for name, uniform title, date, and place; but not all task force members agreed that the current controlled vocabularies are effective for topical subjects. Different points of view during their discussions included both: (1) "[U]sing controlled vocabularies such as LCSH and MeSH for topical subjects is no longer as necessary or valuable. Given our limited cataloging resources, we should apply subject analysis only to material that is not self-discoverable through textual searching"[30]; and (2) "Even with full text searching and enhanced metadata, topical subject headings still provide a valuable collocation service when searching large collections, particularly in multiple languages."[31] The Task Force finally made a recommendation to "Consider using controlled vocabularies only for name, uniform title, date, and place, and abandoning the use of controlled vocabularies [LCSH, MESH, etc] for topical subjects in bibliographic records. Consider whether automated enriched metadata such as TOC, indexes can become surrogates for subject headings and classification for retrieval."[32]

Deanna Marcum in a discussion of how her audience should think about cataloging in the Age of Google, argues that, "now, digital full-length texts are available. And thousands if not millions more of them are in prospect. Potentially, people will be able to search every word from a book's dust jacket to its back-of-the-book index. The need for intermediate-level descriptions

[apparently meaning metadata records including all controlled vocabulary access points] will come under serious scrutiny."[33]

Karen Calhoun, reporting on her structured interviews for her 2006 report to LC on the changing nature of the catalog, states that interviewees did not like LCSH.[34] Calhoun argues that, according to the UC report, "automated enriched metadata such as TOCs can supply additional keywords for searching"[35]; thus, her recommendation: "Abandon the attempt to do comprehensive subject analysis manually with LCSH in favor of subject keywords; urge LC to dismantle LCSH."[36]

Following these reports, LC set up its Working Group on the Future of Bibliographic Control, which worked for more than a year before issuing its report in 2008. One recommendation in this report is: "Optimize LCSH for Use and Reuse."[37] The working group recommended recognizing the flaws in LCSH and working to overcome them:

> Subject analysis is a core function of cataloging, and Library of Congress Subject Headings have great value in providing controlled subject access to works. … While it is recognized as a powerful tool for collocating topical information, LSCH suffers, however, from a structure that is cumbersome from both administrative and automation points of view. Many of the perceived flaws of LCSH are inherent in any subject vocabulary that must encompass the entire range of intellectual creation, rather than a more discrete subject area.[38]

Controlled Vocabulary is Needed for Scholarly Research

A view expressed in much of the literature is that keyword searching is fine for finding a quick answer for a brief, uncritical question; but more is needed for scholarly research. Ingrid Hsieh-Yee wrote in 1998: "For a quick, cursory search, keyword searching is promising even on the Web; but for more in-depth or extensive searches, the limitations of keyword searching, such as the lack of control over synonyms and the need for context to make the words more specific, will result in many irrelevant items for the searcher to wade through."[39] Daniel N. Joudrey, in a 2006 review of the aforementioned reports by Calhoun and by the Bibliographic Services Task Force of the University of California Libraries observes: "Neither [report] discriminates between the related (but distinct) processes of simple information seeking and in-depth scholarly research. It is alarming that they place so much emphasis on the needs of casual information seekers and [give] so little attention to the needs of scholars."[40]

In a detailed description in 2006 of how the Keystone Library Network achieved authority control across its membership, Michael Weber, Stephanie Steely, and Marilou Hinchcliff, speaking of variants such as spelling, language, etc., observe: "[O]ne of the major problems resulting from a lack of proper authority control [is that] in order to obtain complete results, the user needs to have knowledge of cross references and must search on each and every alternative."[41] These are concerns shared by Thomas Mann, who has written extensively about the necessity for using controlled vocabulary for scholars."[42] In 2008, X. Liu, K. Maly, M. Zubair, Q. Hong, and C. Xu address their approach to language issues in *Arc*, an OAI compliant federated digital library. Among other challenging issues listed are these: "how to build a rich unified search interface when there is a lack of controlled vocabulary, and how to federate collections in different languages."[43] In a 2008 case study of a multilingual knowledge

management system for a large organization, Daniel O'Leary asserts: "Multilingual systems have begun to find use in a large number of settings, including government, medical systems and libraries. … [S]ome of the most important technical issues in multilingual systems are ontologies, since they help facilitate communication, structure and search about knowledge issues."[44]

And apparently, not only do scholars miss much of the relevant information if the system has been designed only for quick retrievals, but also, scholars benefit from a controlled vocabulary network if it is there, even when they do not realize it is there. Ying-hsang Liu studied different kinds of users of a database containing MeSH vocabulary. Liu reported, "experimental results strongly suggest that searchers with substantial domain knowledge can benefit from the use of MeSH terms in terms of the precision measure, even though their perception of the usefulness of MeSH terms did not agree with search performance."[45]

Users' Difficulty with Subject Searching

With so much evidence that scholars need more than keyword searching, why are some authors recommending that controlled vocabulary be abandoned? Several researchers have pointed out that many patrons cannot do subject searching successfully. For example Marcia Bates, in 2003, observed: "People have a lot of no-match or poor-match hits when searching for subject, and have learned to use keyword searching as a substitute …. Yet they still like to do subject searching online."[46]

Some writers believe that vocabulary control is 'so last century.' Clay Shirky, in a blog posting about ontologies in 2005, asserts that categorization belongs to a world where things are placed on shelves, not the digital world: "The categorization scheme is a response to physical

constraints on storage, and to people's inability to keep the location of more than a few hundred things in their mind at once."[47] He writes about how categorizing in advance forces the cataloger to do mind-reading of what users want and to predict what they will want in the future: "Whenever users are allowed to label or tag things, someone always says 'Hey, I know! Let's make a thesaurus, so that if you tag something 'Mac' and I tag it 'Apple' and somebody else tags it 'OSX', we all end up looking at the same thing!'"[48] But, says Shirky, "You can't do it. You can't collapse these categorizations without some signal loss. The problem is, because the cataloguers assume their classification should have force on the world, they underestimate the difficulty of understanding what users are thinking, and they overestimate the amount to which users will agree, either with one another or with the catalogers, about the best way to categorize."[49]

Other authors write about the negative reaction of users to LCSH and traditional subject access. Calhoun writes: "Interviewees had a lot to say about LCSH and library tradition for providing subject access. Opinions ranged from the strongly critical to an attitude akin to quiet resignation. There were no strong endorsements for LCSH."[50] Karen Antell and Jie Huang, in their 2008 study using transaction log analysis and user interviews state: "Overall, the research from both transaction log analysis and user-response studies shows that subject searching is difficult for patrons, unlikely to be very successful, and becoming less frequent as patrons's behavior is shaped by keyword search engines such as Google."[51]

Reasons against Relying on Keyword Searching

However, even though most users cannot negotiate subject-heading searches successfully, many authors are not ready to abandon controlled vocabularies. Chan points out that when the question

of whether there is still a need for controlled vocabulary is directed "to information professionals who have appreciated the power of controlled vocabulary, the answer has always been a confident 'yes.' To others, the affirmative answer became clear only when searching began to be bogged down in the sheer size of retrieved results. Controlled vocabulary offers the benefits of consistency, accuracy, and control … which are often lacking in the free-text approach."[52] Antell and Huang state that "reference librarians are aware that patrons doing keyword searches in online catalogs do not find the best results. In fact they frequently retrieve unhelpful result sets of zero, or they retrieve far too many results to be useful."[53] Athena Salaba, in a 2009 study of end-user understanding of indexing language, reports that "[p]articipant statements suggest that they perceive that even though subjects represent a broader area than keywords, results from a subject search are more relevant to their query than the results of a keyword search, which retrieves a narrower area and more irrelevant results."[54]

Garrett points out, in his aforementioned report of an experiment with adding subject headings to ECCO, that certain historical collections would have many non-findable items if it were not for controlled vocabulary: "For a number of reasons, some having to do with changes in the lexicon, some with a century-specific perceived need for circumlocution, words such as 'hygiene' and 'prostitution' occurred far less frequently in the eighteenth century than they do today—not to mention the often disastrous effects of pre-1800 orthography on modern-day keyword searches."[55]

Jeffrey Beall describes the ways in which keyword-based full-text searching can fail. He lists the following as issues or problems with keyword searching: synonyms, variant spellings, word forms, different languages, obsolete terms, disciplinary differences, homonyms, uncontrolled personal names, false cognates, inability to employ facets, clustering, inability to sort, spamming,

aboutness issues, figurative language, word lists, abstract topics, search term not in database, search term unknown, non-textual resources, and paired topics that are difficult to search (e.g., "Art and mental illness").[56] An addition by Mann is that keyword searching "cannot segregate the appearance of the right words in conceptual contexts apart from the appearance of the same words in the wrong contexts."[57]

Cost Moved to Users

Several researchers discuss the problem of moving the cost of providing controlled vocabulary to users when controlled vocabulary is not maintained. Chan says that "[e]ven in the age of automatic indexing and with the ease in keyword searching, controlled vocabulary has much to offer in improving retrieval results and in alleviating the burden of synonym and homograph control placed on the user."[58] George Macgregor and Emma McCulloch, discussing a 2005 blog post by Ian Davis,[59] write: "He has argued that any economies achieved in indexing or classifying resources are simply moved onto the price of resource discovery for users, since the lack of collocation increases the number of locations that users have to explore before satisfying their information need. Davis states that the historical purpose of controlled vocabularies has not altered and notes that high costs have always been incurred by a very small number of information professionals in order to reduce the discovery costs for a large number of users."[60]

Mann[61] and William Badke[62] also give examples of how difficult it is for users who must rely on keyword searching. And Yee asks, "Is it too much to ask for our colleagues in the profession, at least, to understand and acknowledge the value of human intervention for information organization, expensive though it is?"[63]

Controlled Vocabulary Needed for Non-textual Resources

Some types of information resources require *at least* manually assigned keywords, if not controlled vocabulary. One of the UC Task Force's recommendations is: "In allocating resources to descriptive and subject metadata creation, consider giving preference to those items that are completely undiscoverable without it, such as images, music, numeric databases, etc."[64] Donna Slawsky, writing in 2007 about a collection of visual assets states: "[W]e have found that people use different words to express similar ideas, concepts and even things. Therefore, ambiguity is inevitable. This ambiguity makes a controlled vocabulary in the form of a thesaurus essential to any image-retrieval system."[65] The LC Working Group on the Future of Bibliographic Control addressed both non-textual works and non-English works: "As keyword searching becomes increasingly prevalent, non-textual works and works in languages other than English are at risk of becoming less accessible, or even inaccessible."[66] Cosmin Munteanu, reporting in 2009 on a project to provide metadata for Webcast lectures, writes: "[A] set of keywords relevant to each lecture was manually extracted from the slides by the teaching assistant associated with the course. While several automatic, both supervised and unsupervised, keyword extraction algorithms exist, they do not produce entirely accurate results.…"[67]

Controlled Vocabulary in Particular Fields of Study

Numerous studies in particular fields outside the realm of libraries recently have demonstrated the need for controlled vocabulary when searching databases in those fields. In addition to business management, which is addressed below, articles were found in thirteen other subject areas that indicate that controlled vocabulary should be used when searching databases in these disciplines. These subject areas are listed here in order of date of article: Water quality[68],

Physics[69], Medical theses,[70] Women's studies,[71] Bioinformatics,[72] Genomics,[73] Tissue engineering,[74] Medicine,[75] Neuroscience,[76] Biomedicine,[77] Veterinary Medicine,[78] Astronomy,[79] and Clinical Nursing.[80]

Gregory Schymik, Robert St. Louis, and Karen Corral, in a 2009 conference paper, present an explanation of why full-text search alone in enterprise search systems[†] cannot give efficient results, and they demonstrate "the order of magnitude improvements that can be obtained through the incorporation of subject indexes into the search process…."[81] They cite Google for "data indicating that knowledge workers are wasting almost half of their time as a direct result of failed searches."[82] They argue that "by obliterating the more traditional approach to archive management, corporations have introduced tools destined to dissatisfy their users."[83] They assert that, "adding contextual information to the search will decrease the number of irrelevant documents without decreasing the number of relevant documents in the result set.… If searchers, particularly in the enterprise context, are presented a smaller result set, they are more likely to take the time to review the results and not give up on the search."[84] And finally, they declare that "[o]ur findings support the earlier findings of Voorbij (1998) and Gross and Taylor (2005) that the addition of subject metadata search can improve search results.… Our results also show that incorporating metadata into the search process is very likely (.975) to result in a tenfold

---

[†] *From Wikipedia 7/21/11: "'Enterprise Search' is used to describe the software of search information within an enterprise (though the search function and its results may still be public). Enterprise search can be contrasted with web search, which applies search technology to documents on the open web, and desktop search, which applies search technology to the content on a single computer.… Enterprise search systems index data and documents from a variety of sources such as: file systems, intranets, document management systems, e-mail, and databases. Many enterprise search systems integrate structured and unstructured data in their collections."*

improvement in search for 97.95% of searches. This is very strong evidence that the use of subject metadata should be incorporated into the search process."[85]

In a separate 2009 conference paper Schymik further elucidates the enterprise search problem:

> Enterprise search is a popular, but frequently unsuccessful, mechanism for transferring knowledge amongst knowledge workers inside individual firms. According to data presented during a recent Google webinar on the release of a new version of their enterprise search appliance, knowledge workers are wasting almost half of their time as a direct result of poor search capabilities.... They also spend another 25% of their time conducting what they define to be successful searches for information, leaving only about one quarter of a knowledge worker's time being spent on truly value added activity. Middle managers further noted that often times, the information they do find is wrong.... This data makes it no surprise that 86% of enterprise searchers are unsatisfied with their enterprise search capabilities….[86]

In order to be able to justify the up-front cost of determining and entering the data required to significantly improve enterprise searches, Karen Corral, David Schuff, Robert St. Louis, and Ozgur Turetken present a model for estimating the total cost to a company of relying on keyword searches versus relying on a subject category approach: "Our analysis of the model shows that a surprisingly small number of searches are required to justify the cost associated with encoding the metadata necessary to support a dimensional [i.e., subject categories] search engine. The results imply that it is cost effective for almost any business organization to implement a dimensional search strategy."[87] The authors go on to say that having predefined subject

information "eliminates the ambiguity of words (which causes so many of the problems for keyword search) through the use of pre-defined categories (dimensions) to define documents as well as finite sets of possible values for each category. It has been demonstrated that dimensional search reduces the number of irrelevant documents returned in the result set.… From our model we were able to determine the break-even point, in terms of the number of searches, at which dimensional search becomes more cost effective than keyword search. That is, we were able to determine the number of searches an organization must do in order to justify the up-front cost of determining and entering the metadata that is required to support dimensional search."[88] Finally, the authors declare that "[f]or a firm with 1,000 employees and 100,000 documents in the document store, an average of only 25 searches per employee (25,000 searches) would be required to justify the cost of encoding the metadata required to support dimensional searches. This provides convincing evidence that organizations should strongly consider implementing dimensional document stores."[89]

In 2010 Corral, Schuff, Schymik, and St. Louis reported an experiment that measured the impact of adding subject metadata to keyword-based full-text searches. They concluded: "Our extremely encouraging results suggest that the traditional library process of indexing the contents of the library against a controlled vocabulary of subjects, authors, and titles might need to be rejuvenated in the context of enterprise search."[90]

Solutions Offered

The literature suggests a few solutions for resolving the keyword searching versus controlled vocabulary dilemmas. The most prominent are:

- make use of both keyword searching and controlled vocabulary

- make use of tagging done by users
- use user search terms to augment controlled vocabularies
- create tools specifically designed to help untrained users to make use of controlled vocabulary
- automatically add tables of contents, summaries, or other metadata that can supply additional words for keyword searching

Both Controlled Vocabulary and Keyword Searching

Numerous authors suggest that controlled vocabulary can be used to augment keyword searching to give users a more satisfactory result. Over a decade ago, Chan observed: "Controlled vocabulary most likely will not replace keyword searching, but it can be used to supplement and complement keyword searching to enhance retrieval results."[91] Several reports of research back up Chan's suggestion: Nowick and Mering[92]; Elizabeth Jenuwine and Judith Floyd[93]; Mohammad Reza Davarpanah and Mohammad Iranshahi[94]; Weber, Steely, and Hinchcliff[95]; and Pamela Morgan.[96]

Several other authors write about their observations concerning the complementary nature of controlled vocabulary and keyword searching. After a complaint in the Los Angeles Times in 2009 about failure of a keyword search in a library catalog, Judith Herman wrote a letter to the editor, saying: "If she had clicked 'Browse Catalog,' then selected 'Subject Browse' from the menu, she would have found the subject heading [for the topic sought]…. Unfortunately, cutbacks at the Library of Congress threaten the future of subject headings and so threaten us all

with the loss of information that keywords will never find."[97] Also in 2009, Gilles Hubert and Josiane Mothe assert, "Combining the two modes [searching with keywords or with descriptors] allows users to select categories they clearly identify as related to their information needs and to complement their queries with keywords for which they do not identify corresponding categories."[98] Sevim McCutcheon, in a 2009 article comparing keyword searching and controlled vocabulary, says, "My view from the catalog librarian's perspective is that the two main tools of information retrieval, keyword and controlled vocabulary, in fact complement one another."[99]

Jack Hang-tat Leong argues in 2010 that the somewhat separate areas of metadata schemas and bibliographic control are converging.[100] He sees them as engaging in kind of a spiral dance as they work around each other to use natural language at times and controlled vocabulary at times to provide subject access. He says: "This convergence will lead to the triumph of the hybrid approach, a combination of the human approach of controlled vocabulary and the automation approach of algorithmic generation of metadata, in providing subject access."[101]

User Tagging Systems

Another suggested solution to the keyword versus controlled vocabulary dilemma is to make use of collaborative tagging systems. Tags and "folksonomy" – the collection of tags used within one platform – have many of the same issues that are found with keyword searching, and tagging has the additional issue of tags that are personal (e.g., 'to read'), are silly, or are purposely misleading. Folksonomies, though, are touted because of the perception that no formal thesaurus can keep up with user needs.

A number of articles address the tagging phenomenon, comparing it to traditional indexing.[102] In a thorough analysis published in 2006, Macgregor and McCulloch write: "Collaborative tagging

has emerged as a means of organising information resources on the Web and is contradictory to the ethos of controlled vocabularies."[103] They say at another point: "The emergence of 'collaborative tagging' is therefore considered by some as a useful way in which to supersede the subject indexing role of the information professional…."[104] They observe that, in 2006, "[n]o control is exerted in collaborative tagging systems over synonyms or near-synonyms, homonyms and homographs, and the numerous lexical anomalies that can emerge in an uncontrolled environment. The probability of noise in a user's result set is therefore very high."[105]

Peter Rolla compares LibraryThing's user tags and LCSH and suggests that while user tags can enhance subject access to library collections, they cannot replace the valuable functions of a controlled vocabulary like LCSH. He writes, "If libraries do allow users to contribute tags to their catalogs, they will need to figure out how to deal with some of the inherent problems encountered in folksonomies."[106] Jo Bates and Jennifer Rowley examine LibraryThing from a British perspective and find it dominated by United States taggers, which has an impact on the tagging of ethnic minority resources. They observe: "Folksonomy, like traditional indexing, is found to contain its own biases in worldview and subject representation."[107] They recommend integrating folksonomies into catalogs "to provide a partial improvement to the discoverability and subject representation of some 'non-dominant' resources … but with an awareness of the biases that it contains."[108] Sarah Hayman and Nick Lothian also see a value in using tagging for augmenting controlled vocabularies. They write that "[observation of] terms suggested, chosen, and used in folksonomies is a rich source of information for developing our formal systems so that we can indeed get the best of both worlds."[109] And Hong Zhang, Linda Smith, Michael Twidale, and Huang Gao, argue that "the weighting of subject terms [e.g., placing resulting hits

from subject headings higher in a retrieved list] is more important than ever in today's world of growing collections, more federated searching, and expansion of social tagging."[110]

Macgregor and McCulloch remark that "[i]t is curious to note that during the period in which collaborative tagging has emerged, a reaffirmation of controlled vocabularies has arisen in parallel. The requirement for improved information organisation and management within the corporate sector has facilitated the increased deployment and development of corporate taxonomies."[111] And, indeed, a perusal of the literature on tagging and folksonomies written since 2006 shows that much has been written about an alternative to free-for-all tagging – an alternative called "tag gardening," "structured folksonomy," "structured collaborative tagging," or "collaborative ontology engineering."[112] The idea presented in these studies is that with the increase in social sharing sites, traditional indexing is not feasible, but, at the same time, the more user tags there are, the more unruly they become, and then, in order for them to be useful, it becomes necessary to weed, seed, and fertilize (using the gardening analogy) or to impose facets or categories (using the structuring or engineering analogy).

Use of User Search Terms to Augment Controlled Vocabularies

Not quite the same as tagging/folksonomy is the idea that professional organizers can use the search terms of users (i.e., keywords) to expand and supplement controlled vocabularies. There is a large corpus of research dealing with "query expansion" – that is, the idea of reformulating a search query after observing retrieved results. Some of this research encourages use of a particular controlled vocabulary list to assist in finding synonyms to search, or finding terms that will broaden or narrow results or that will find related material. For example, Jane Greenberg reports an experiment examining whether thesaurus terms that are related to a search query in a

specified semantic way (e.g., synonyms, narrower terms, related terms, broader terms), could be identified as having a positive impact on retrieval effectiveness when added to a query through automatic query expansion, or, alternatively, when used for interactive query expansion.[113]

Although a majority of this corpus of research is beyond the scope of this paper and deserves its own literature review, a small portion of the group is concerned with making improvements to controlled vocabularies by incorporating and/or adapting users' search terminology (i.e., keywords). June Abbas, in writing about the creation of metadata for children's resources, notes that there is a significant body of research into adults' use of information systems, but there is much less research into children's understandings of such systems, or into use of their search terms as a source for controlled vocabulary.[114] Abbas posits that development of age-appropriate representations of objects is necessary for good retrieval.[115] She describes a study using the ARTEMIS Digital Library, a collection designed to provide high-quality age-appropriate resources for middle and high school science students. Transaction logs provided a source of search terms entered by students after they had composed the questions that they were trying to answer. One outcome of the study was the development of a list of 205 student-generated keywords; all of the terms in the list were unique and were not included in the controlled vocabulary used by the system.[116]

Prototype Tools

The fourth suggested solution to the keyword versus controlled vocabulary dilemma is to create searching tools that will find the appropriate search terms that both satisfy the information need and also match the language used in the information system. Karen Markey Drabenstott says: "Since end users will gravitate to subject searches, we need experimentation with interfaces that

help end users to accomplish these tasks and, at the same time, tell them why these tasks will benefit them."[117] Markey called specifically for work toward new interfaces, with researchers, practitioners, and system designers working together to create and test prototypes."[118] Creation of such tools is still in experimental stages.

Among the first tools provided to accomplish the purpose of helping end users with subject searching are various ontologies and integrated controlled vocabularies. For example, "[t]he Ontology Lookup Service (OLS) was created to integrate publicly available biomedical ontologies into a single database. All modified ontologies are updated daily. A list of currently loaded ontologies is available online."[119] Liu, Qin, Chen, and Park write about another successful integration of controlled vocabularies in a particular subject area: "While users of Internet search engines are generally not concerned about controlled vocabulary, the usefulness and effectiveness [of] controlled vocabulary in information retrieval has been proven in specialized search systems such as the Unified Medical Language System (UMLS)…. Most digital libraries built for educational purposes offer a search option for using controlled vocabulary."[120] A third unified ontology is the Open Biomedical Resources (OBR) described by Noy, et al.[121]

Vivien Petras introduces a "search term recommender," based on statistical associations between specialized language terms and controlled vocabulary terms.[122] Hubert and Mothe propose a search engine that will integrate both "browsing an ontology (via categories)" and "defining a query in free language (via keywords)."[123] Charles-Antoine Julien and Charles Cole describe the design and development of an interactive visual map of a collection's major subject headings and their relations. The resulting visualization prototype is a complement to keyword searching.[124] Julien, Catherine Guastavino, France Bouthillier, and John Leide developed a "virtual reality subject browsing and information retrieval prototype … [that] allows users to explore the LCSH

subject hierarchy and its assigned documents by travelling up and down the hierarchy of broad to narrow subjects. Integrated with keyword searching, users are able to visually inspect subject headings written on labels hovering hierarchy branches."[125]

Addition of TOCs and Summaries/Abstracts

A fifth solution proposed for the keyword versus controlled vocabulary dilemma is to add to bibliographic records tables of contents, summaries, or other metadata that can supply additional words for keyword searching. In a 1987 study Drabenstott and Calhoun analyzed catalog records from four large research libraries.[126] They found that the largest source of unique subject rich words (from 9 to 20 unique subject rich words per record) came from summary and contents notes. LCSH contributed from 3 to 7 unique subject rich words per record.

Subject rich words found in summaries and contents notes help recall, but they cause a problem for precision, because the terminology is not controlled. Nevertheless, users like summaries and contents notes, and have become accustomed to having them available through use of sites such as Amazon.com. Partly because of the additional metadata on such sites, the 2005 Bibliographic Services Task Force of the University of California Libraries Report recommends that the UC Libraries should: "Consider whether automated enriched metadata such as TOC, indexes can become surrogates for subject headings and classification for retrieval."[127] In a table of suggested responses to various user desires, the Task Force suggests that in order to provide better result sets, a library should "[i]ndex TOC, abstracts, [and] other enriched metadata for a wider variety of searchable metadata."[128] "Other enriched metadata" is defined elsewhere in the report as: "cover art, publisher promotional blurbs, content excerpts (print, audio or video), and bibliographies"[129] and "user-provided reviews."[130] Calhoun, in her 2006 report to LC, states that

"interviewees also suggested enrichment of the catalog with title page or jacket images, reviews, tables of contents and such…."[131] And later in the report she says, "As the UC report points out, automated enriched metadata such as TOCs can supply additional keywords for searching."[132]

Zhou, Yu, Smalheiser, Torvik, and Hong, in a 2007 paper about domain-specific knowledge, state: "[W]hile some experts may well be adept at choosing the right number and types of keywords, it is fair to say that for most others the literature search process is laden with considerable frustration.… One way to overcome these limitations may be to store what we term 'structured annotations' along with the full text of each publication. By tying keywords to specific contexts (unique to each scientific field) and by controlling the vocabulary for these annotations, many of these limitations may be avoided."[133]

OCLC's 2009 report also shows that users expect to find enriched metadata: "Both groups of respondents [i.e., end users and librarians] rely on and expect enhanced content, including summaries/abstracts and tables of contents.... The findings suggest that summaries are most important in searches for unknown items."[134] The report further states: "To aid in discovery, end users reported that they want *more subject information*, followed by the addition of evaluative information similar to what librarians predicted—*adding tables of contents and summaries/abstracts*."[135] The report then gives voice to concerns about cost: "To support these features, today's catalogs rely on labor-intensive practices for producing controlled subject headings. Given the growing concern that these traditional methods are not sustainable going forward, it may be necessary for libraries to find more economical means to achieve the benefits to end users that controlled subject vocabularies provide."[136]

However, research continues to suggest that controlled vocabularies are needed to provide unique search terms not available even in additional content. In the report of a 2009 study of overlap between author-assigned keywords and cataloger-assigned Library of Congress Subject Headings for a set of electronic theses and dissertations (ETDs) Rockelle Strader found:

> A notable result occurred when keywords and LCSH were matched against abstracts, which are included in the bibliographic records for OSU ETDs. Author-assigned keywords exactly matched words in the abstract 54.61 percent of the time, while cataloger-assigned LCSH exactly matched only 26.84 percent of abstract words. Keyword nonmatches occurred 10.59 percent of the time, and cataloger-assigned LCSH nonmatches occurred 31.08 percent of the time. Put another way, about one-tenth of the keywords and roughly one-third of the assigned LCSH are unique to the bibliographic records. This result corroborates Gross and Taylor's findings…. In terms of the discoverability of bibliographic records, the use of LCSH significantly complements keywords by providing further unique terms for searching and matching, even in the presence of enhancements such as abstracts.[137]

McCutcheon, in 2011, also discusses the issue of providing access to electronic theses and dissertations.[138] Because only sophisticated scholars seek out ETD repositories, metadata records need to be integrated with databases such as OCLC Worldcat. McCutcheon discusses the possibility of using the required metadata supplied by the authors of theses and dissertations, but, in comparing the author-supplied metadata for 92 ETDs with the actual works, she found that "in the abstract field alone, the student authors had spelling errors that impact findability in 12 ETDs (13%), and the total number of spelling errors in abstracts were 17."[139] She found that authors

also sometimes omitted or misspelled title words, and "[a]nother obstacle to access has to do with the representation of scientific symbols, diacritics, and some punctuation in author-supplied metadata."[140] She concludes that although "[k]eywords and controlled vocabulary each have their advantages and disadvantages …, keyword access alone cannot suffice for thorough and comprehensive retrieval by subject.… [F]or fullest access, and the best possible service to users who seek material on a subject, subject analysis and the assignment of subject headings is key to maximizing access by topic."[141]

In a 2012 publication Schwing, McCutcheon, and Maurer replicated Strader's research using electronic theses and dissertations in another catalog, with a smaller sample, but reporting in more detail. The authors found that both author-assigned keywords and cataloger-assigned LCSH provide unique terms that enhance access.[142]

Need for Controlled Vocabulary Even with Full Text Available

The idea of adding enhancements to bibliographic records invokes the same questions asked about full text databases, one of which is the question of why there should be any metadata at all, if every word of the text can be searched. Already mentioned above are the articles about enterprise search, which comprises full text searching in business databases. These and numerous other articles suggest that even in full text databases, controlled vocabulary can be used in conjunction with keyword searching to gain, essentially, the best of both worlds. Among the recent research articles found on this subject, only one suggested that there might be a way to do full text searching successfully without any controlled vocabulary. The article suggesting that controlled vocabulary may not be needed is one published in 2007 by Bradley Hemminger, Billy Saelim, Patrick Sullivan, and Todd Vision.[143] They write: "Significantly more articles were

discovered via full-text searching; however, the precision of full-text searching also is significantly lower than that of metadata searching.… By using the number of hits of the search term in the full-text to rank the importance of the article, performance of full-text searching was improved so that both recall and precision were as good as or better than that for metadata searching. This suggests that full-text searching alone may be sufficient, and that metadata searching as a surrogate is not necessary."[144]

The most common finding, however, is that searching of full text indexes is more successful when controlled vocabulary has been added. Arturo Montejo Raez and Ralf Steinberger, writing in 2004, present a typical assessment: "[T]he use of full text indexes has its limitations, especially in the multilingual context, and it is not a solution for further information access requirements…. We show that automatic indexing with controlled vocabulary keywords (descriptors) complements full-text indexing because it allows cross-lingual information access."[145] They also say, "We have shown that manual or automatic indexing of document collections with controlled vocabulary thesaurus descriptors is complementary to full-text indexing and that it provides both human users and machines with the means to analyse, navigate and access the contents of document collections in a way full-text indexing would not permit."[146]

One reason that full text presents difficulties for searching is explained by Zipf's Law. In simple terms, as the Law applies in this situation, George Zipf observes "that the number of meanings a word takes on in a given collection of documents is roughly equivalent to the square root of the number of times the word appears in that set of documents."[147] So if a keyword appears 9 times in a set of documents, it very likely appears with 3 different meanings. It is, of course, difficult to imagine coming up with a set of keywords for searching that will distinguish among the meanings, especially for a large collection. Hayman and Lothian, writing in 2007, note that

"[w]ithout even considering the issue of other languages, English itself has a huge number of words with multiple meanings. Vocabularies have been built for specific communities where the meanings chosen are appropriate for that context … but even within communities there can be ambiguities of meaning."[148] And if multiple languages are involved, there is the problem of words in different languages spelled the same as English words but having different meanings.

In the aforementioned 2007 article by Garrett on adding subject headings in ECCO, he writes: "This article extends arguments recently presented by Gross and Taylor (2005) in two directions: first, by considering the importance of subject headings for access to historical materials; and, second, by examining the value added by subject headings even when the full text of a work is available online."[149] Garrett asserts that important terms and concepts are found in subject headings in metadata that cannot be found in the full text itself:

> In response [to administrators wondering whether to fund subject analysis work], it can be readily shown that keyword searching in full-text databases is no substitute for searches run against OPACs or other bibliographic files with ample descriptors and subject headings. …. The demonstrable fact is that full-text searching of eighteenth-century texts often does not retrieve examples of terms that describe the work as a whole or even important topics or aspects of the work, especially as we might describe them today. Indeed, those researching the topic of urban sanitation in the eighteenth century might be surprised to learn that there is not a single valid occurrence of the word "sanitation" in the entire 26,000,000-page ECCO corpus.… With foreign-language works, of course, the disjunction approaches 100%.[150]

Additionally, as pointed out in a 2012 article by Buckland: "Even when the denotation is stable, the connotation or attitudes to the connotation may change. Always, some linguistic expressions are socially unacceptable. That might not matter much, except that what is deemed acceptable or unacceptable not only differs from one cultural group to another, but changes over time, and, especially during changes, may be the site of contest. The phrase "yellow peril" was widely used to denote what was seen as excessive immigration from East Asia, but it is now considered too offensive to use even though there is no convenient and acceptable replacement name and the phrase remains needed in historical discussion."[151]

In an article published in 2008 Sheila Bair and Sharon Carlson discuss a project to describe some Civil War diaries so as to make them accessible to an audience of historians, genealogists, and others. They report: "This paper [shows] how the addition of controlled vocabularies for personal, corporate, and geographic names, and pre-coordinated topic searches to transcribed and marked up primary texts increases their research value, provides searchability far beyond mere full-text keyword, and can facilitate scholar and student access to these materials."[152] After describing how the diaries were transcribed and tagged with names, terms, and definitions of obsolete terms, they write: "Inclusion of controlled vocabularies in the XML markup helps to disambiguate between names and commonly used words. For instance, the words cotton, hill, gray, wood, and cousin are also names of people and places in the diaries."[153] They further elaborate: "Librarians involved in this project have noted the increasing number of reference questions in the last decade about non-military aspects of Civil War history such as clothing, health, leisure, and religion. Because of the interest in these topics, a decision was made to incorporate subject analysis at the word level in the XML markup."[154] They conclude: "Primary sources, such as diaries and letters, are foundational to digital humanities research.… However,

merely scanning and providing full-text keyword searchability may not fully meet the needs of digital humanities scholars. Abbreviations, obsolete and regional word usage, idioms, misspellings and alternate spellings, and omissions in primary sources make keyword searching, especially across many items in online collections, unproductive."[155]

Beall, also writing about the needs of scholars in 2008, asserts: "Linguistic problems, the limitations of full-text search engines, and missing data combine to make full-text searching unreliable, incomplete, and insidiously imprecise, especially for serious information seeking, such as scholarly research."[156] And in their study of the synonym problem in full-text searching, Beall and Karen Kafadar found that, "The extent of the synonym problem in full-text searching depends on whether one searches the more common of the synonyms. Overall, the measure of what's missed is as high as 30% in a large (90%) fraction of common word-pairs. Information discovery systems need to take the synonym problem into account and develop solutions for it, both probabilistic and deterministic.… Additionally, the data demonstrate the value of vocabulary control and cross references in providing more precise search results."[157] Hans-Michael Müller, Arun Rangarajan, Tracy Teal, and Paul Sternberg, writing about the difficulty of searching thousands of neuroscience papers, observe that assigned categories can offer assistance.[158]

In their 2009 discussion of the high cost of full-text searching in businesses, Schymik, St. Louis, and Corral write: "This article explains why full-text search alone cannot yield the results sought by enterprise searchers..."[159] They observe that the "use of subject indexes has largely been replaced by the use of enterprise search appliances built on full-text web search engines. The indeterminacy of language leads to very large result sets being returned by such search engines. We have demonstrated that incorporating the search of subject metadata into the search process

dramatically reduces the size of the result set. In the case of enterprise search, we suggest that it might be better to automate, not obliterate, the traditional library search process."[160] In his related 2009 conference paper, Schymik observes that, "[a]s document collections get large, the complexities of language make it very difficult to define a set of query terms that will adequately describe the documents we search for yet sufficiently discriminate between relevant and irrelevant documents."[161] After describing Zipf's Law [as discussed above], Schymik says, "[G]iven the fact that the number of meanings a word takes on increases with the square root of the number of times the word appears in a given collection, it is … fairly obvious that, for reasonably large collections (those containing more than a few hundred documents) it is nearly impossible to choose a set of keywords that will discriminate relevant from irrelevant documents."[162]

Elaine Nowick, Daryl Travnicek, Kent Eskridge, and Stephen Stein, in a 2010 study, discuss use of controlled vocabulary and keywords identified by automated text analysis or word clustering techniques for documents in an online environment, and explore similarity among terms from users, from the documents themselves, and from controlled vocabularies. Their findings show that "the controlled vocabulary terms were better matched to both users' search terms and document terms than documents to users. Correlations between users and controlled vocabularies were 2-3 times higher [than] between users and documents.… This suggests that, through controlled vocabularies, libraries do provide a bridge between users and relevant documents.… These results would indicate that human catalogers are the ideal way to organize documents into a library. However, given the limitations of humans to undertake a complete catalog of the internet, there may be ways to refine cluster-based organizing algorithms for digital libraries."[163]

Corral, Schuff, Schymik, and St. Louis in 2010 "performed an experiment that measured the impact of adding subject metadata to keyword-based full-text searches."[164] They state that their experimental research supports the earlier findings of Voorbij and of Gross and Taylor, who found that subject metadata improves search results, and it "extends their findings beyond a search of the bibliographic record to an evaluation of the impact the addition of metadata search has on full-text search."[165]

The preponderance of the literature continues to show that controlled vocabularies are useful, and indeed are necessary in some cases, such as in searching full text. For keyword searching of bibliographic records, including those that have been given tags by users of the systems, most studies show that controlled vocabularies cannot be replaced by keyword searching for in-depth, scholarly work. Only three research studies were identified that address the issue of whether enhancements, such as tables of contents and summaries or abstracts, can replace controlled vocabulary. One is Strader's study of electronic theses and dissertations in 2009; another is McCutcheon's study in 2011; and the third is the 2012 study by Schwing, McCutcheon, and Maurer. All three found that LCSH significantly complements keywords. Because abstracts are, in a sense, "full text," this seems a logical finding in comparison to the studies of full-text searching that show that controlled vocabularies are also needed in full-text situations. The current study seeks to provide a sense of whether Strader's and Schwing, et al.'s findings are extendable to the more general set of records found in a university library catalog.

Research Questions

The research questions guiding this investigation expand upon the research question from the earlier study. In 2005, Gross and Taylor asked, "What proportion of records retrieved by a

keyword search has a keyword only in a subject heading field and thus would not be retrieved if there were no subject headings?"[166] This question applies to the current study as well. Beyond this question, however, the researchers also ask: (1) What proportion of records retrieved by a keyword search has a keyword only in a subject heading field in a catalog enriched with TOCs & summary notes?; and (2) What proportion of records retrieved by a keyword search has a keyword only in a subject heading field when the results are not limited to English? The purpose of this study is to revisit the research question from the first study in the context of the new questions posed.

Methodology

In order to replicate the first study so that results would be comparable, the authors employed the same methodology that was used in the 2005 Gross and Taylor study.[167] Conducting the searches in the "next generation catalog" (at the time the searches were performed, the University of Pittsburgh was using Aquabrowser) in addition to the OPAC was considered, but the authors concluded that while investigation of the role of subject headings in discovery layers would be essential future research, it would not be appropriate to address it in a study intended to respond to criticisms of the former study. As in the earlier study, captured searches from a transaction log were used to conduct a series of keyword searches to determine what proportion of the records retrieved by each user's search had a keyword only in a subject heading field and would not be retrieved if the subject headings were absent. The searches were conducted after the University of Pittsburgh library system began to use Blackwell's Table of Contents Enrichment service to add table of contents and summary notes to English language monographs that had been

published since 1992.[†] Each search was conducted twice, once with search results limited to English language materials (as was done in the 2005 study) and again with no language limit placed on the searches. Except where indicated, data in this report correspond to searches performed with no language limit.

The search terms used in the current research were the same as those in the 2005 Gross and Taylor study. The terms were taken from a March 2000 transaction log of 3,397 keyword searches from the catalog of the library at Winthrop University, Rock Hill, South Carolina. The searches ranged from single terms to multi-word phrases. De-duplicating the search terms repeated in the transaction log reduced the number of possible terms to 2,270. A sample size of 227 searches was selected based on a common statistical formula for determining sample size.[168]

Keyword searches on each set of terms were conducted in PittCat Classic, the traditional interface to the University of Pittsburgh's online public access catalog, which contains more than six million[169] titles from all of the university's libraries. To minimize the impact of duplicate holdings while including a broad range of materials, the searches were limited to the holdings of the Pittsburgh campus libraries (the University Library System, Law, and Health Sciences libraries). Stopwords, including "a," "an," "and," "by," and nine others, were omitted from the searches.

[†] *1992 was the earliest date for which TOC enrichment data was available from Blackwell at the time, and it appears to continue to be the date before which TOC enrichment is not yet available. The former Blackwell service is now provided by Yankee Book Peddler (http://www.ybp.com/MARCenrichmentservice.html), which offers "coverage dating back to 1992." The authors could not identify any existing service that offers TOC enrichment for earlier publications.*

A small number of searches in the sample yielded zero hits with the keywords anywhere, and were excluded from the analysis. Also excluded were searches that retrieved more than 10,000 hits, the maximum that PittCat will display. Since the total number of hits for these searches was unknown, the proportion of hits lost in the absence of subject headings could not be determined.

For each search in the sample, the following data were collected:

1. number of hits with all keyword(s) anywhere
2. number of hits with all keyword(s), and at least one in subject, but not all in title
3. number of the first fifty hits from the second search with at least one keyword in subject only (or, when the second search had fifty or fewer hits, the total number of hits with at least one keyword only in a subject,)

The steps used to collect this data are best explained with a concrete example. In the rest of this section, a search from the sample, *horror films* (with no language limit), is used to demonstrate each step in the data collection process.

The first step was to determine the number of hits with all of the keyword(s) anywhere. The search *horror films* retrieved 1017 hits with the keywords anywhere. Like most of the sets retrieved, this was too large to examine each hit manually, and so a second search was performed to reduce the number of records that would have to be viewed.

The second step was to perform a search for the number of hits containing all of the keywords, with at least one keyword in the subject fields, but not all of them in the title fields (see figure 1).

(Insert Figure 1)

This second search eliminated many of the hits that would still have been retrieved if the subject headings had not been present because all of the keywords were present in a title field. In figure 2, for example, one can see that both keywords are in the title, as well as in the subject headings.

(Insert Figure 2)

By performing the second search, records like the one in figure 2 were excluded from the set to be examined manually. *Horror films* had 823 hits with all keywords somewhere in the record and at least one in a subject heading, but not all keywords in a title field.

Because keywords can appear in many parts of a bibliographic record, including author, series, notes, and publication/distribution information, it was still necessary to view individual records to determine if any keywords were present only in the subject headings.

The third step was to view the first fifty hits from the second search (or all of the hits, when there were fifty or fewer).

In the 2005 Gross and Taylor study, "the first fifty were used rather than sampling because PittCat displays results of keyword searches in reverse chronological order and thus the most recent, and presumably the most useful, hits appear first."[170] The use of random sampling to select fifty hits to be viewed manually was tested by the researchers for possible inclusion in this study, but no statistically significant difference was found between using the first fifty hits and using fifty random hits.[171]

Of the 823 records from the second search for *horror films*, the first 50 were viewed to determine that 37 of them had at least one keyword in a subject field only. For example, the record in figure

3 contains the keywords only in the subject heading *Horror films*—United States—History and criticism.

(Insert Figure 3)

These 37 hits are 74 percent of the first fifty hits. Applying this proportion to the 823 hits from the second search, it was projected that the total number of hits with at least one keyword present only in a subject field in a search for *horror films* would be 609.02.

The final step was to determine the percentage of hits that would be lost out of the total number of hits, based on the number of hits with a keyword only in the subject headings identified in the second step. For *horror films*, there were 1017 hits with the keywords anywhere, and a projected 609.02 hits with at least one keyword in a subject field. Therefore, for the search *horror films*, an estimated 59.9 percent of the hits would not have been retrieved without the subject headings. Data from all searches is available in St. Cloud State University's institutional repository.[172]

Limitations

The most significant limitation of this study is that results with no language limit (not limited to English) cannot be compared to results in the pre-enhancement catalog, since data for searches with no language limit was not collected in the 1995 study. A comparison of search results before and after systematic TOC and summary enhancement can only be made for searches limited to English.

A second limitation is that the enhancement data added to the University of Pittsburgh's catalog was available only for English language monographs published since 1992. This study did not attempt to limit search results to exclude publications from before 1992, or to limit the analysis

to bibliographic records that had received enhancement. Instead, it compares the hits that would be lost without subject headings in the real search results provided by a large academic library's catalog before and after implementation of available TOC and summary enhancement, measuring the impact of actually existing enhancement services.

However, because the third step in the methodology employed used the proportion of records that would be lost from the first fifty hits (those with the most recent publication dates, since reverse chronological order is the default sort in PittCat Classic) to project the proportion of all hits that would be lost for each search, the proportion associated with records for very recent publications may be overrepresented in the results.

Findings

When search results included materials in all languages, the mean percentage of hits that would be lost in the absence of subject headings in a catalog with summary and contents data enrichment was 27 percent, and the median was 17.6 percent. The overall percentage of hits that would be lost when the results of all searches were aggregated was 27.7 percent (45,086.14 out of 162,574 hits).

For about 20.4 percent of the search sample (39 out of 191), the percentage of hits with a keyword only in a subject field was 50 percent or greater. This means that for about 1 out of every 5 successful keyword searches, half or more of the hits now retrieved would not be retrieved if there were no subject headings.

Searches with three keywords (36 out of 191, or 18.8% of the sample) would lose an average of 36.6 percent of retrieved hits if the subject fields were not present. Searches with four or more

keywords (16 out of 191, or 8.4% if the sample) would lose an average of 40 percent of retrieved hits (see figure 4). The average proportion of hits that would be lost appears to increase as the number of keywords increases, but regression analysis did not suggest any significant difference depending on the number of keywords.[173]

(Insert Figure 4)

There were many searches, using what appeared to be common terms for popular topics, for which the number of the hits that would not be found in the absence of subject headings was higher than two thirds, such as *film criticism*, *businesswomen*, and *hispanic americans* (see figure 5).

(Insert Figure 5)

Limited to English

The searches were also performed with the results limited to English, as was done in the 2005 study. With that limit, the mean percentage of hits that would be lost in the absence of subject headings was 24.8 percent (compared to 27% when not limited to English). The overall percentage of hits that would be lost when the results of all searches were aggregated was 27.9 percent (43,964.52 out of 157,618 hits).

The average percentage of hits that would be lost in searches for materials in all languages was 2.2 percent higher than the percentage lost in searches limited to English.

With and Without Table of Contents/Summary Data Enrichment

The 2005 study found that in a catalog before systematic TOCs and summary enhancement, the average percentage of hits that would be lost in searches limited to English in the absence of subject headings was 35.9 percent. The current study found that in a catalog after systematic enhancement, the average percentage of hits lost in searches limited to English was 24.8 percent, 11.1 percent less than without enhancement.

Future Research

The importance of controlled vocabulary in library catalogs and other databases consisting of metadata is established by a significant body of research, including the present study. Research that looks at the effect of controlled subject vocabulary in discovery layers and web-scale discovery tools has begun to appear, and in the near term, these rapidly changing environments are the domain in which the impact of subject headings needs to be investigated most urgently. In the long term, the ultimate test of the importance of controlled vocabulary will be its effect in full text environments. While most studies that have looked at the role of subject metadata in full text searching indicate that controlled vocabulary is needed in full text environments, research in this area needs to continue and expand as the extent and accessibility of full text resources increases.

Most studies on the value of controlled vocabulary in keyword searching, whether looking at searches performed on surrogate metadata or on full text, have focused on the presence of keywords without any consideration of relevance. The present study asks what proportion of hits would be lost if no subject headings were present in catalog records, but does not attempt to determine what proportion of hits – of those lost in the absence of subject headings, or of those that would be retrieved without subject headings - would be deemed relevant by the users

performing the searches. Arguably, it could be surmised that a larger proportion of the lost one-fourth of hits would be relevant to the users than would be the case in the retrieved three-fourths because the lost one-fourth all contain at least one keyword in a subject heading, while the retrieved three-fourths may or may not. Research examining relevance in addition to the presence of keywords in records is needed.

Conclusion

The 2005 study of the effect of controlled vocabulary on the results of keyword searching found that an average of 35.9 percent of hits in keyword searches would be lost if subject headings were to be removed from or no longer included in catalog records. The current study found that with the addition of tables of contents and summaries or abstracts, an average of 27 percent of hits would be lost if the subject headings were not present in the records. While the proportion of hits that would be lost in the absence of subject headings is reduced with the addition of contents and summary data, it still represents a significant proportion of total hits (more than one fourth). This study also found that when limited to English, the loss is 24.8 percent, demonstrating that subject headings in English are, indeed, helpful in locating materials in other languages.

As demonstrated in reviewing the literature, there are many additional advantages to including controlled vocabulary in metadata records, such as grouping synonyms and variant spellings and word forms, providing references from and to obsolete terms, distinguishing among variant meanings of the same term, and providing hierarchical references, not to mention the usefulness of providing searchable text for non-textual resources.

Emerging and future uses of controlled vocabulary are also significant. The use of subject headings to support faceted searching and relevance ranking is only in its early stages. The

potential applications of LCSH and other vocabularies as linked data have only begun to be explored. Indeed, as the cataloging world turns toward linked data, the notion that tables of contents and subject keywords obviate the need for controlled subject vocabulary seems anachronistic. Implementing a linked data framework for bibliographic metadata means that access points based on text strings will need to be replaced with Uniform Resource Identifiers (URIs). As the mantra heard in discussions about the Bibliographic Framework Transition Initiative goes, we need to use "things, not strings."[174] Linked data requires the use of URIs to uniquely identify things likes names, resources, and subjects on the web, and URIs for subjects cannot be based on uncontrolled keywords.

Assertions that controlled subject vocabulary is no longer needed contradict the vast majority of research results, and appear to disregard primary emerging methods of providing subject access. This study adds to mounting evidence that controlled vocabulary continues to be an essential tool for assisting users to find the resources that they seek.

Endnotes

[1] Christine L. Borgman, "Why are Online Catalogs Hard to Use?," *Journal Of The American Society For Information Science* 37, no. 6 (1986): 387-400; Borgman, "Why are Online Catalogs Still Hard to Use?," *Journal Of The American Society For Information Science* 47, no. 7 (1996): 493-503; Thom Hickey, "Why Our Catalogs Don't Work," *Outgoing*: *Library Metadata Techniques and Trends*, September 15, 2005, available: http://outgoing.typepad.com/outgoing/2005/09/why_our_catalog.html.

[2] Karen Markey, "The Online Library Catalog: Paradise Lost and Paradise Regained?," *D-Lib Magazine* 13, no. 1/2 (2007), available: http://www.dlib.org/dlib/january07/markey/01markey.html.

[3] Library of Congress Cataloging Policy and Support Office, "Library of Congress Subject Headings Pre- vs. Post-Coordination and Related Issues. March 15, 2007," Library of Congress, accessed, available: http://www.loc.gov/catdir/cpso/pre_vs_post.html.

[4] The Library of Congress Working Group on the Future of Bibliographic Control, "On the Record, Report of the Library of Congress Working Group on the Future of Bibliographic Control," (2008), 19, available: http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf.

[5] Tina Gross and Arlene G. Taylor, "What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results," *College & Research Libraries* 66, no. 3 (2005): 213.

[6] Ibid., 223.

[7] Karen Calhoun, "The Changing Nature of the Catalog and its Integration with Other Discovery Tools: Final Report, Prepared for the Library of Congress. March 17 2006," Library of Congress, available: http://www.loc.gov/catdir/calhoun-report-final.pdf.

[8] Ibid., 46.

[9] Jennifer Rowley, "The Controlled versus Natural Indexing Languages Debate Revisited: A Perspective on Information Retrieval Practice and Research," *Journal of Information Science* 20, no. 2 (1994): 108–19.

[10] Donald H. Kraft, "A Comparison of Keyword-in-context (KWIC) Indexing of Titles with a Subject Heading Classification System," *American Documentation* 15 (Jan. 1964): 48.

[11] Carolyn O. Frost, "Title Words as Entry Vocabulary to LCSH," *Cataloging & Classification Quarterly* 10, no. 1-2 (1989): 176.

[12] Barbara Keller, "Subject Content Through Title: A Masters Theses Matching Study at Indiana State University," *Cataloging & Classification Quarterly* 15, no. 3 (1992): 78.

[13] Henk J. Voorbij, "Title Keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences," *Journal of Documentation* 54, no. 4 (Sept. 1998): 466-76.

[14] Elaine A. Nowick and Margaret Mering, "Comparisons between Internet Users' Free-Text Queries and Controlled Vocabularies: A Case Study in Water Quality," *Technical Services Quarterly* 21, no. 2 (2003): 15.

[15] Gross and Taylor, "What Have We Got to Lose?," 223.

[16] Caimei Lu, Jung-ran Park, and Xiaohua Hu, "User Tags Versus Expert-assigned Subject Terms: A Comparison of Library Thing Tags and Library of Congress Subject Headings," *Journal of Information Science* 36, no. 6 (2010): 775-776.

[17] OCLC, "Online Catalogs: What Users and Librarians Want: An OCLC Report," (Dublin, Ohio: OCLC Online Computer Library Center, 2009), 11.

[18] Ibid., 14.

[19] Kayo Denda, "Beyond Subject Headings: A Structured Information Retrieval Tool for Interdisciplinary Fields, *Library Resources & Technical Services* 49, no. 4 (2005): 266.

[20] Lois Mai Chan, "Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources: Issues and Challenges," (In Library of Congress, *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium*, 2000), 5. Available: http://www.loc.gov/catdir/bibcontrol/chan_paper.html.

[21] Ibid., 2.

[22] Rebecca Donlan and Rachel Cooke, "Running with the Devil: Accessing Library-licensed Full Text Holdings through Google Scholar," *Internet Reference Services Quarterly*, 10, nos. 3-4 (2005): 155-156.

[23] Ibid., 156.

[24] Jeffrey Garrett, "Subject Headings in Full-Text Environments: The ECCO Experiment," *College & Research Libraries*, 68, no. 1 (2007): 72.

[25] Ibid., 74.

[26] Nancy J. Fallgren, "Users and Uses of Bibliographic Data: Background Paper for the Working Group on the Future of Bibliographic Control," (Feb. 25, 2007), available: http://www.loc.gov/bibliographic-future/meetings/docs/UsersandUsesBackgroundPaper.pdf.

[27] Sue Ann Gardner, "Changing Landscape of Contemporary Cataloging," *Cataloging & Classification Quarterly*, 45, no. 4 (2008): 88.

[28] Oksana L. Zavalina, "Collection-Level Subject Access in Aggregations of Digital Collections: Metadata Application and Use" (PhD dissertation – University of Illinois at Urbana-Champaign, 2010), 113.

[29] Ibid.

[30] Bibliographic Services Task Force of the University of California Libraries, "Rethinking How We Provide Bibliographic Services for the University of California: Final Report," 2005, 23, available: http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf.

[31] Ibid.

[32] Ibid., 24 [brackets in the original].

[33] Deanna B. Marcum, "The Future of Cataloging: Address to the Ebsco Leadership Seminar Boston, Massachusetts January 16, 2005, 10, available: http://www.loc.gov/library/reports/CatalogingSpeech.pdf.

[34] Calhoun, "The Changing Nature of the Catalog," 33.

[35] Ibid., 46.

[36] Ibid., 18.

[37] Library of Congress Working Group on the Future of Bibliographic Control, "On the Record," 19..

[38] Ibid., 34-35.

[39] Ingrid Hsieh-Yee, "Search Tactics of Web Users in Searching for Texts, Graphics, Known Items and Subjects," *The Reference Librarian* 28 (1998): 79.

[40] Daniel N. Joudrey, "Book Reviews: The Changing Nature of the Catalog and its Integration with Other Discovery Tools; Rethinking How We Provide Bibliographic Services for the University of California: Final Report," *Library Resources & Technical Services* 50, no. 4 (2006): 296.

[41] Michael A. Weber, Stephanie A. Steely, and Marilou Z. Hinchcliff, "A consortial authority control project by the Keystone Library Network," *Cataloging & Classification Quarterly* 43, no. 1 (2006): 78.

[42] Thomas Mann, "What is Going on at the Library of Congress?" Prepared for AFSCME 2910, The Library of Congress Professional Guild, 2006, 18-19; available: http://www.guild2910.org/AFSCMEWhatIsGoingOn.pdf; Thomas Mann, " 'On the Record' but Off the Track: A Review of the Report of The Library of Congress Working Group on The Future of Bibliographic Control, With a Further Examination of Library of Congress Cataloging Tendencies," Prepared for AFSCME 2910, The Library of Congress Professional Guild, 2008, 11; available: http://www.guild2910.org/WorkingGrpResponse2008.pdf.

[43] X. Liu, K. Maly, M. Zubair, Q. Hong, and C. Xu, "An OAI Compliant Federated Digital Library," Citeseer, 2008, 1; available: http://citeseerx.ist.psu.edu.

[44] Daniel E. O'Leary, "A multilingual knowledge management system: A case study of FAO and WAICENT," *Decision Support Systems* 45 (2008): 642.

[45] Ying-hsang Liu, The Impact of MeSH Terms on Information Seeking Effectiveness, Thesis (PhD) – Rutgers University, 2009, 108-9.

[46] Marcia J. Bates, "Task Force Recommendation 2.3: Research and Design Review: Improving User Access to Library Catalog and Portal Information: Final Report (Version 3)," Library of Congress, from *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium*, 2003, 49. Available: http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf.

[47] Clay Shirky, "Ontology is overrated: Categories, Links, and Tags," 2005, 5. Available: http://shirky.com/writings/ontology_overrated.html.

[48] Ibid., 12-13.

[49] Ibid., 13.

[50] Calhoun, "The Changing Nature of the Catalog," 33.

[51] Karen Antell and Jie Huang, "Subject Searching Success: Transaction Logs, Patron Perceptions, and Implications for Library Instruction," *Reference & User Services Quarterly* 48, no. 1 (2008): 69-70.

[52] Chan, "Exploiting LCSH, LCC, and DDC," 4.

[53] Antell and Huang, "Subject Searching Success," 68.

[54] Athena Salaba, "End-user Understanding of Indexing Language Information," *Cataloging & Classification Quarterly* 47, no. 1 (2009): 30.

[55] Garrett, "Subject Headings in Full-text Environments," 70.

[56] Jeffrey Beall, "The Weaknesses of Full-Text Searching," *Journal of Academic Librarianship* 34, no. 5 (2008): 439-443.

[57] Thomas Mann, "Will Google's Keyword Searching Eliminate the Need for LC Cataloging and Classification?," Prepared for AFSCME 2910, The Library of Congress Professional Guild, 2005, 4. Available: http://www.guild2910.org/searching.htm.

[58] Chan, "Exploiting LCSH, LCC, and DDC," 4.

[59] Davis, Ian, ''Why tagging is Expensive,'' (2005), available: http://blogs.capita-libraries.co.uk/panlibus/2005/09/07/why_tagging_is_/.

[60] George Macgregor and Emma McCulloch, "Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool," *Library Review* 55, no. 5 (2006): 296.

[61] Mann, " 'On the Record' but Off the Track."

[62] William Badke, "The Treachery of Keywords," *Online* 35, no. 3 (2011): 52-54.

[63] Yee, "Will the Response of the Library Profession to the Internet be Self-immolation?," 5.

[64] Bibliographic Services Task Force of the University of California Libraries, "Rethinking How We Provide Bibliographic Services," 24.

[65] Donna Slawsky, "Building a Keyword Library for Description of Visual Assets: Thesaurus Basics," *Journal of Digital Asset Management* 3, no. 3 (2007): 134.

[66] Library of Congress Working Group on the Future of Bibliographic Control, "On the Record," 20.

[67] Cosmin Munteanu, Useful Transcriptions of Webcast Lectures, Thesis (PhD) – University of Toronto, 87.

[68] Nowick and Mering, "Comparisons between Internet Users' Free-Text Queries and Controlled Vocabularies."

[69] Arturo Montejo Ráez and Ralf Steinberger, "Why Keywording Matters," *High Energy Physics Libraries Webzine*, Issue 10 (December 2004): 1-16.

[70] Maria Ansari, "Matching between Assigned Descriptors and Title Keywords in Medical Theses," *Library Review* 54, no. 7 (2005): 410-414.

[71] Kayo Denda, "Beyond Subject Headings: A Structured Information Retrieval Tool for Interdisciplinary Fields," *Library Resources & Technical Services* 49, no. 4 (2005): 266-275.

[72] Richard G. Côté, Philip Jones, Rolf Apweiler, and Henning Hermjakob, "The Ontology Lookup Service, a Lightweight Cross-platform Tool for Controlled Vocabulary Queries," *BMC Bioinformatics* 7, no. 97 (2006): 1-7.

[73] Wei Zhou, Clement Yu, Neil Smalheiser, Vetle Torvik, and Jie Hong, "Knowledge-intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature," SIGIR 2007 Proceedings, July 23–27, 2007, Ámsterdam, The Netherlands, 2007, 655-662.

[74] Abhishek Jain, Prakash Velayutham, Michael Wagner, and David L. Butler, "Accessing the Tissue Engineering Literature: A New Paradigm, *Tissue Engineering Part A*. 14, no. 3 (2008): 459-460.

[75] Xiaozhong Liu, Jian Qin, Miao Chen, Ji-Hong Park, "Automatic Semantic Mapping between Query Terms and Controlled Vocabulary through Using WordNet and Wikipedia," ASIS&T 2008 Annual Meeting, Columbus, Ohio, October 24-29, 2008, 1-10.

[76] Hans-Michael Müller, Arun Rangarajan, Tracy K. Teal, Paul W. Sternberg, "Textpresso for Neuroscience: Searching the Full Text of Thousands of Neuroscience Research Papers," *Neuroinform* 6 (2008): 195-204.

[77] Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen, "BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse," *Nucleic Acids Research* 37, Web Server issue, published online 29 May 2009, W170-W173.

[78] Kristine M. Alpi, Elizabeth Stringer, Ryan S. DeVoe, and Michael Stoskopf, "Clinical and Research Searching on the Wild Side: Exploring the Veterinary Literature," *Journal of the Medical Library Association* 97, no. 3 (2009): 169-170.

[79] Alasdair J.G.Gray, Norman Gray, Christopher W. Hall, and Iadh Ounis, "Finding the Right Term: Retrieving and Exploring Semantic Concepts in Astronomical Vocabularies," *Information Processing and Management* 46 (2010): 470-478.

[80] Susan B. Stillwell, Ellen Fineout-Overholt, Bernadette Mazurek Melnyk, and Kathleen M. Williamson, "Searching for the Evidence: Strategies to Help you Conduct a Successful Search," *AJN: American Journal of Nursing* 110, no. 5 (2010): 41-47.

[81] Gregory Schymik, Robert St. Louis, and Karen Corral, "Order of Magnitude Reductions in the Size of Enterprise Search Result Sets Through the Use of Subject Indexes," Americas Conference on Information Systems (AMCIS) *Proceedings*, paper 195, 2009, 2.

[82] Ibid.

[83] Ibid., 3.

[84] Ibid., 4.

[85] Ibid., 7.

[86] Gregory Schymik, "Representational Indeterminacy and Enterprise Search: The importance of subject indexes," Proceedings of the Fifteenth Americas Conference on Information Systems, San Francisco, California August 6th-9th 2009, 1.

[87] Karen Corral, David Schuff, Robert D. St. Louis, and Ozgur Turetken, "A Model for Estimating the Savings from Dimensional vs Keyword Search," In *Advanced Principles for Improving Database Design, Systems Modeling, and Software Development*, edited by Keng Siau and John Erickson, (Hershey, NY: Information science reference, 2009), 146.

[88] Ibid., 147-8.

[89] Ibid., 156.

[90] Karen Corral, David Schuff, Gregory Schymik, and Robert St. Louis, "Strategies for Document Management" *International Journal of Business Intelligence Research* 1 (no. 1): 78-9.

[91] Chan, Exploiting LCSH, LCC, and DDC, 4.

[92] Nowick and Mering, "Comparisons between Internet Users' Free-Text Queries and Controlled Vocabularies," 30.

[93] Elizabeth S. Jenuwine and Judith A. Floyd, "Comparison of Medical Subject Headings and Text-word Searches in MEDLINE to Retrieve Studies on Sleep in Healthy Individuals," *Journal of the Medical Library Association* 92, no. 3 (2004): 349.

[94] Mohammad Reza Davarpanah and Mohammad Iranshahi, "A Comparison of Assigned Descriptors and Title Keywords of Dissertations in the Iranian Dissertation Database," *Library Review* 54, no. 6 (2005): 383.

[95] Weber, Steely, and Hinchcliff, "Consortial Authority Control Project," 2006, 78-79.

[96] Pamela S. Morgan, The Availability of MeSH in Vendor-Supplied Cataloguing Records, as Seen Through the Catalogue of a Canadian Academic Health Library," *Partnership: the Canadian Journal of Library and Information Practice and Research* 2, no. 2 (2007): 1.

[97] Judith B. Herman, "A Word about a Library Tool," Letters to the Editor, *Los Angeles Times* (Nov. 29, 2009).

[98] Gilles Hubert, and Josiane Mothe, "An Adaptable Search Engine for Multimodal Information Retrieval," *Journal of the American Society for Information Science and Technology*, 60, no. 8 (2009): 1625.

[99] Sevim McCutcheon, "Keyword vs Controlled Vocabulary Searching: The One with the Most Tools Wins," *The Indexer* 27, no. 2 (2009): 62.

[100] Jack Hang-tat Leong, "The Convergence of Metadata and Bibliographic Control? Trends and Patterns in Addressing the Current Issues and Challenges of Providing Subject Access," *Knowledge Organization* 37, no. 1 (2010): 29-41.

[101] Ibid., 40.

[102] Sarah Hayman and Nick Lothian, "Taxonomy Directed Folksonomies: Integrating User Tagging and Controlled Vocabularies for Australian Education Networks," World Library and Information Congress: 73rd IFLA General Conference and Council, 19-23 August 2007, Durban, South Africa, 27 p.; Katrin Weller, "Folksonomies and Ontologies. Two New Players

in Indexing and Knowledge Representation," in H. Jezzard (Ed.), *Applying Web 2.0. Innovation, Impact and Implementation. Online Information 2007 Conference Proceedings*, London (2007), 108-115; Peter J. Rolla, "User Tags versus Subject Headings: Can User-Supplied Data Improve Subject Access to Library Collections?," *Library Resources & Technical Services* 53, no. 3 (2009): 174-184; Tom Steele, The New Cooperative Cataloging, *Library Hi Tech* 27, no. 1 (2009): 68-77; Sue Yeon Syn and Michael B. Spring, "Tags as Keywords - Comparison of the Relative Quality of Tags and Keywords," *Proceedings of the American Society for Information Science and Technology* 46, no. 1 (2009): 1-19; Kwan Yi and Lois Mai Chan, "Linking Folksonomy to Library of Congress Subject Headings: an Exploratory Study," *Journal of Documentation* 65, no. 6 (2009): 872-900; Margaret E. I. Kipp, "Searching with Tags: Do Tags Help Users Find Things?," *Knowledge Organization : KO* 37, no. 4 (2010): 239-255; Caimei Lu, Jung-ran Park, and Xiaohua Hu, "User Tags versus Expert-assigned Subject Terms: A Comparison of LibraryThing Tags and Library of Congress Subject Headings," *Journal of Information Science* 36, no. 6 (2010): 763–779; Brian Matthews, Catherine Jones, Bartlomiej Puzon, Jim Moon, Douglas Tudhope, Koraljka Golub, Marianne Lykke Nielsen, "An evaluation of enhancing social tagging with a knowledge organization system," *Aslib Proceedings* 62, no. 4/5 (2010): 447 – 465; Jo Bates and Jennifer Rowley, "Social Reproduction and Exclusion in Subject Indexing: A Comparison of Public Library OPACs and LibraryThing Folksonomy," *Journal of Documentation* 67, no. 3 (2011); Hong Zhang, Linda C. Smith, Michael Twidale, and Fang Huang Gao, "Seeing the Wood for the Trees: Enhancing Metadata Subject Elements with Weights," *Information Technology and Libraries* 30, no.2 (2011): 75-80.

[103] Macgregor and McCulloch, "Collaborative Tagging," 292.

[104] Ibid., 294.

[105] Ibid., 295.

[106] Rolla, "User Tags versus Subject Headings," 182.

[107] Bates and Rowley, "Social Reproduction and Exclusion in Subject Indexing," 431.

[108] Ibid., 446.

[109] Hayman and Lothian, "Taxonomy Directed Folksonomies," 2.

[110] Zhang, Smith, Twidale, and Gao, "Seeing the Wood for the Trees," 79.

[111] Macgregor and McCulloch, "Collaborative Tagging," 298.

[112] Isabella Peters and Katrin Weller, (2008). "Tag Gardening for Folksonomy Enrichment and Maintenance." *Webology*, 5(3), Article 58. Available: http://www.webology.ir/2008/v5n3/a58.html; Katrin Weller and Isabella Peters, "Seeding, Weeding, Fertilizing – Different Tag Gardening Activities for Folksonomy Maintenance and Enrichment," in *Proceedings of I-SEMANTICS '08* Graz, Austria, September 3-5, 2008:110-117; Koraljka Golub, Catherine Jones, Brian Matthews, Bartlomiej Puzon, Jim Moon, Douglas Tudhope, Marianne Lykke Nielsen, "EnTag: Enhancing Social Tagging for Discovery," Joint Conference on Digital Library, *JCDL '09*, June 15-19, 2009, Austin, TX, p. 163-172; Louise F. Spiteri, "Incorporating Facets into Social Tagging Applications: An Analysis of Current Trends," *Cataloging & Classification Quarterly* 48, no. 1 (2010): 94-109; Wolfgang G. Stock, "Concepts and Semantic Relations in Information Science," *Journal of the American Society for Information Science and Technology* 61, no. 10 (2010): 1951–1969; Wolfgang G. Stock, Isabella Peters, Katrin Weller, "Social Semantic Corporate Digital Libraries: Joining Knowledge Representation and Knowledge Management," in Anne Woodsworth (ed.) *Advances in Librarianship, Volume* 32, Emerald Group Publishing Limited, (2010), pp.137-

158; H. H. Kim, "Toward Video Semantic Search Based on a Structured Folksonomy," *Journal of the American Society for Information Science and Technology*, 62, no. 3 (2011): 478–492; Hemalata Iyer, Lucy Bungo, "An Examination of Semantic Relationships between Professionally Assigned Metadata and User-generated Tags for Popular Literature in Complementary and Alternative Medicine" *Information Research* 16, no. 3 (September 2011): 26; Maciej Gawinecki, Giacomo Cabri, Marcin Paprzycki, and Maria Ganzha, "Evaluation of Structured Collaborative Tagging for Web Service Matchmaking" in *Semantic Web Services* (2012): 173-189; Stefanie Haustein, Isabella Peters, "Using Social Bookmarks and Tags as Alternative Indicators of Journal Content Description," *First Monday* 17, no. 11 (5 November 2012): available: http://firstmonday.org/ojs/index.php/fm/article/viewArticle/4110; Yi-Ling Lin, Lora Aroyo, "Interactive Curating of User Tags for Audiovisual Archives," in Genny Tortora, Stefano Levialdi, and Maurizio Tucci (eds.), *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)*, New York, NY, ACM, (2012), pp. 685-688; Lohmann, Steffen, "Conceptualization and Visualization of Tagging and Folksonomies," dissertation, Universidad Carlos III de Madrid. Departamento de Informática (2013).

[113] Jane Geenberg, Automatic Query Expansion via Lexical-Semantic Relationships," *Journal of the American Society for Information Science and Technology* 52, no. 5 (2001): 402-415; and "Optimal Query Expansion (QE) Processing Methods with Semantically Encoded Structured Thesauri Terminology," *Journal of the American Society for Information Science and Technology* 52, no. 6 (2001): 487-498.

[114] June Abbas, "Creating Metadata for Children's Resources: Issues, Research, and Current Developments," *Library Trends* 54, no. 2 (Fall 2005): 303-317.

[115] June Abbas, "Out of the Mouths of Middle School Children: I. Developing User-Defined Controlled Vocabularies for Subject Access in a Digital Library," *Journal of the American Society for Information Science and Technology* 56, no. 14 (2005): 1512-1524.

[116] Ibid, p. 1520.

[117] Karen Markey Drabenstott, "Information Retrieval Systems for End Users: Primetime Players that Just Don't Make the Grade," *Journal of Education for Library and Information Science* 45, no. 2 (2004): 175-176.

[118] Markey, "The Online Library Catalog," 8.

[119] Côté, Jones, Apweiler, and Hermjakob, "The Ontology Lookup Service, a Lightweight Cross-platform Tool for Controlled Vocabulary Queries," 1.

[120] Liu, Qin, Chen, and Park, "Automatic Semantic Mapping between Query Terms and Controlled Vocabulary through Using WordNet and Wikipedia," 2.

[121] Noy, Shah, Whetzel, Dai, Dorf, Griffith, Jonquet, Rubin, Storey, Chute, and Musen, "BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse," W171.

[122] Vivien Petras, "Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages," (2006), 1

[123] Hubert and Mothe, "An Adaptable Search Engine," 1625.

[124] Charles-Antoine Julien and Charles Cole, "Capitalizing on Controlled Subject Vocabulary by Providing a Map of Main Subject Headings: An Exploratory Design Study," *Canadian Journal of information and Library Science* 33, no. 1/2 (2009): 67-83.

[125] Charles-Antoine Julien, Catherine Guastavino, France Bouthillier, and John E. Leide, "Subject Explorer 3D: a Virtual Reality Collection Browsing and Searching Tool," Information Science: Synergy through Diversity, Concordia University, Montreal, Quebec, June 2 - 4 2010, Conference Proceedings, 8 p.

[126] Karen Markey Drabenstott and Karen Calhoun. "Unique Words Contributed by MARC Records with Summary and/or Contents Notes." In ASIS '87, Proceedings of the 50th ASIS Annual Meeting, edited by Ching-chih Chen, (Medford, NJ: Learned Information, 1987), 153–162.

[127] Bibliographic Services Task Force of the University of California Libraries, "Rethinking How We Provide Bibliographic Services," 24.

[128] Ibid., 31.

[129] Ibid., 25.

[130] Ibid., 18.

[131] Calhoun, "The Changing Nature of the Catalog," 39.

[132] Ibid., 46.

[133] Zhou, Yu, Smalheiser, Torvik, and Hong, "Knowledge-intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature," 655-662.

[134] OCLC, "Online Catalogs: What Users and Librarians Want: An OCLC Report," 17.

[135] Ibid., 48 [italics in the original].

[136] Ibid., 52.

[137] C. Rockelle Strader, "Author-Assigned Keywords versus Library of Congress Subject Headings," *Library Resources & Technical Services* 53, no. 4 (2009): 249.

[138] Sevim McCutcheon, "Basic, Fuller, Fullest: Treatment Options for Electronic Theses and Dissertations," *Library Collections, Acquisitions & Technical Services* 35 (2011): 64-68.

[139] Ibid., 65.

[140] Ibid., 66.

[141] Ibid., 66-67.

[142] Theda Schwing, Sevim McCutcheon, and Margaret Beecher Maurer, "Uniqueness Matters: Author-Supplied Keywords and LCSH in the Library Catalog," *Cataloging & Classification Quarterly*, 50, no. 8 (2012): 903-928.

[143] Bradley M. Hemminger, Billy Saelim, Patrick F. Sullivan, and Todd J. Vision, "Comparison of Full-Text Searching to Metadata Searching for Genes in Two Biomedical Literature Cohorts," *Journal of the American Society for Information Science and Technology* 58, no.14 (2007): 2341–2352.

[144] Ibid., 2341.

[145] Ráez and Steinberger, "Why Keywording Matters," 1.

[146] Ibid., 12.

[147] George Kingsley Zipf, as described in Schymik, "Representational Indeterminacy and Enterprise Search," 4.

[148] Hayman and Lothian, "Taxonomy Directed Folksonomies," 15-16.

[149] Garrett, "Subject Headings in Full-Text Environments," 69.

[150] Ibid., 75.

[151] Michael K. Buckland, "Obsolescence in Subject Description," Journal of Documentation 68, no. 2 (2012):159.

[152] Sheila A. Bair and Sharon Carlson, "Where Keywords Fail: Using Metadata to Facilitate Digital Humanities Scholarship," University Libraries Faculty & Staff Publications. Paper 12, 2008, 2-3.

[153] Ibid., 6.

[154] Ibid., 12-13.

[155] Ibid., 15.

[156] Beall, "The Weaknesses of Full-Text Searching," 444.

[157] Jeffrey Beall and Karen Kafadar, "Measuring the Extent of the Synonym Problem," *Evidence Based Library and Information Practice* 3, no. 4(2008): 28-29.

[158] Muller, Rangarajan, Teal, and Sternberg, "Textpresso for Neuroscience," 195.

[159] Schymik, St. Louis, and Corral, "Order of Magnitude Reductions," 2.

[160] Ibid., 8.

[161] Schymik, "Representational Indeterminacy and Enterprise Search," 3.

[162] Ibid., 4.

[163] Elaine A. Nowick, Daryl Travnicek, Kent Eskridge, and Stephen Stein, "A Comparison of Term Clusters for Tokenized Words Collected from Controlled Vocabularies, User Keyword Searches, and Online Documents," *Library Philosophy and Practice* (Nov. 2010), 5-6.

[164] Corral, Schuff, Schymik, and St. Louis, "Strategies for Document Management," 78.

[165] Ibid., 80.

[166] Gross and Taylor, "What Have We Got to Lose?," 215.

[167] Ibid., 215-219.

[168] David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 2nd ed. (New York: Freeman, 1993), 438. The formula used was N=(z*σ/m)2, with the values z*=1.96 for 95% confidence; σ=.3 for the standard deviation estimated from preliminary searching, and m=.04 for a 4% margin of error. The resulting equation was ((1.96)(.3)/.04)2 = 216.09. Since the total number of unique searches (2270) divided by the desired sample size (216.09) came out to 10.5, we decided for simplicity's sake to select every tenth unique search for the sample, although this made the sample size slightly larger than it needed to be to achieve 95% confidence and a 4% margin of error.

In the 2005 Gross and Taylor study, there was an error in de-duplicating search terms that were repeated in the transaction log. Both "prayers schools" and "schools prayer" were included in the sample of 227 searches. In order to maintain consistency and generate results that could be compared to those of the 2005 study, it was necessary to use the same sample, without alteration, for the present study.

[169] The tremendous growth in the number of titles found in the online catalog since the 2005 Gross and Taylor study is likely due to the batchloading of bibliographic records for

electronic resources in external databases. This was not being done on a significant scale when the data for the 2005 study was collected, but has become a regular practice in recent years.

[170] Gross and Taylor, "What Have We Got to Lose?," 218.

[171] When the results were not limited to English, there were 91 searches with 50 or more hits with keyword anywhere and 1 or more in subjects but not all in title (column D). A sample was tested using 50 random hits selected using Research Randomizer (http://www.randomizer.org). The proportion of hits that would be lost without the presence of the subject headings went up for some searches and down for others. The average percentage of hits lost increased by 1.37%. A paired t-test found no evidence of difference.

[172] Data spreadsheet: http://repository.stcloudstate.edu/lrs_facpubs/39.

[173] A regression analysis on the percentages of loss and number of keywords showed no evidence of a linear relationship, even after performing a a logarithmic transformation of the data (log10(percent lost + .01)) to correct for heteroscedasticity of the percentages of loss.

[174] Amit Singhal, "Google Official Blog: Introducing the Knowledge Graph: things, not strings" Posted May 16, 2012, http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html; Eric Miller, "MARC into Linked Data: An update on the Bibliographic Framework Initiative" (presented as a NISO/DCMI Webinar, January 23, 2013), http://dublincore.org/resources/training/NISO_Webinar_20130123/nisodcmi-webinar-bibframe-20130123.pdf ("things, not strings" is presented on slides 15-16)

Figure 1. Second search performed to reduce hits needing to be viewed manually.

Figure 2. Record with keywords in subject headings and also in title.

Comedy-horror films : a chronological history, 1914-2008 / Bruce G....

Title: Comedy-horror films : a chronological history, 1914-2008 / Bruce G. Hallenbeck.

Author: Hallenbeck, Bruce G., 1952-

Publisher: Jefferson, N.C. : McFarland, c2009.

ISBN: 9780786433322 (softcover : alk. paper)
0786433329 (softcover : alk. paper)

Physical Description: vii, 247 p. : ill. ; 26 cm.

LC Subject Heading(s): Horror films --History and criticism.
Comedy films --History and criticism.
Motion pictures --History --20th century.

Notes: Includes bibliographical references (p. 235-236) and index.
Includes filmographies: p. 209-234.

Summary: "This guide takes a look at the comedy-horror movie genre, from the earliest stabs at melding horror and hilarity during the nascent days of silent film, to its full-fledged development. Selected short films such as Tim Burton's Frankenweenie are also covered. Photos and promotional posters, interviews with actors and a filmography are included"--Provided by publisher.

Contents: The silents: unheard punchlines and subtitled screams
The thirties: old dark houses and gorilla suits
The forties: killer zombies and comedy teams
The fifties: elderly monsters and black humor
The sixties: gothic castles and cleavage galore
The seventies: naked vampires and young Frankensteins
The eighties: American werewolves and toxic avengers
The nineties: screams and cemetery men
Comedy-horror in the new millennium.

Figure 3. Record with keywords only in subject headings.

Going to pieces : the rise and fall of the slasher film, 1978-1986 / Adam...

Title: Going to pieces : the rise and fall of the slasher film, 1978-1986 / Adam Rockoff.
Author: Rockoff, Adam, 1974-
Publisher: Jefferson, N.C. : McFarland & Co., c2002.
ISBN: 0786412275 (ill. case bdg. : alk. paper)
Physical Description: ix, 214 p. : ill. ; 26 cm.
LC Subject Heading(s): Horror films --United States --History and criticism.
Notes: Includes bibliographical references (p. 203-205) and index.
Contents: 1. What Is a Slasher Film?
2. Pre-History of the Slasher Film
3. Halloween: The Night He Came Home
4. Deadly Prank Calls, Driller Killers and an Angry Young Woman
5. Friday the 13th, Prom Night and a Head in the Fish Tank
6. Trains of Terror, Funhouses, Horrible Holidays and a Maniac
7. Campus Killers, Slashing for Laughs and One Human Brain
8. Prowlers, Spaghetti Slashers and the Joys of Summer Camp
9. Nightmare on Elm Street, Sequels Galore and the Decline of the Slasher Film
10. Resurgence
App. Alternative Titles.

Figure 4. Results by number of keywords in search.

	all searches	***1 keyword***	***2 keywords***	***3 keywords***	***4 or more keywords***
# of searches	191	40	99	36	16
median # of hits	218	876	243	85	23.5
average % lost	27%	16.4%	25.8%	36.6%	40%
median % lost	17.6%	7.6%	16.4%	28.9%	25.6%

Figure 5. Individual searches with more than two thirds of hits lost without subject headings.

keyword(s)	***number of hits***	***number of hits with a keyword in subject headings only***	***% of hits retrieved that would be missed without subject headings***
juvenile folk tales	81	81	100%
indian pottery	553	502.74	90.9%
film criticism	3114	2800	89.9%
interprofessional relations	118	103.68	87.9%
businesswomen	326	276	84.7%
baptists united states	1014	842.8	83.1%
teaching foreign language	2182	1789.2	82%
hispanic americans	1322	1042.36	78.8%
mass media politics	1306	863.52	66%
violence motion pictures	321	207.46	64.6%