Text to accompany presentation slides: "Controlled Vocabulary in the Age of Google? Really?

Slide 2:

Just to make sure we are all on the same page with definitions….

Slide 3:

Also for the purposes of this presentation, controlled vocabulary is abbreviated "CV" in the slides (in order to save space).

Slide 4:

Here's an example of a controlled term from LCSH.

Slide 5:

Searching with Controlled Vocabulary was traditionally called Subject Searching and consisted of typing a controlled vocabulary term or phrase into a search box, usually starting with the left-most letter of the term or phrase and continuing to type the letters exactly as found in the Controlled Vocabulary. This required users to find the used term or phrase in the Controlled Vocabulary first and then move to the bibliographic file to search using the correct term or phrase.

Now, most catalogs allow a keyword search for any term in a Controlled Vocabulary phrase, and once a desired term or phrase has been found, a user can click on that term or phrase to retrieve a list of records to which that term or phrase has been applied.

Slide 6:

CV includes….

For the purposes of this presentation, I'm including categorization as a kind of controlled vocabulary. Even though the ones like Dewey and LC use numbers and letters, they also use words (terms); so in the sense that they bring together terms representing a concept, they are controlled vocabularies.

A drop down list does not fit the traditional definition of Controlled Vocabularies in that it does not bring together all terms or phrases representing a concept. But it does do a pretty good job of bringing together words that begin with the same letters.

Slide 7:

Let's suppose I want to find a track for walking near my home…

Search for "tra" ….

The terms above the line are my own personal "controlled vocabulary," because they are terms that have words beginning with "tra" that I have used successfully before.

Slide 8:

Add "ck" ….

Slide 9:

Add "s" ….

The problem is that even though it pulls together many forms of the word, it does not help with the concepts. OCLC's 2009 evidence-based study of what constitutes "quality" in catalog data states that keyword searching is king, but advanced search options and facets help end users refine searches, navigate, browse and manage large results sets.

In OCLC's research, end users wanted to be able to do a simple Google-like search and get results that exactly match what they expect to find. Even if the words they use in their searches have multiple meanings and depend on context, they still expect their searches to return appropriate materials on **exactly** what they want.

> OCLC, "Online Catalogs: What Users and Librarians Want: An OCLC Report," (Dublin, Ohio: OCLC Online Computer Library Center, 2009), p. 11, 14

Slide 10:

So … to summarize so far… We know that over the last couple of decades, it has been acknowledged that online public access catalogs (OPACs) are difficult for patrons to use, partly due to the complexity of subject searching in the catalog.

> 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, http://outgoing.typepad.com/outgoing/2005/09/why_our_catalog.html.
>
> Karen Markey, "The Online Library Catalog: Paradise Lost and Paradise Regained?," *D-Lib Magazine* 13, no. 1/2 (2007), accessed June 20, 2010, doi: 10.1045/january2007-markey.

Keyword searching has become the most often used, and, in fact, the preferred, method of conducting a search in any online system, including OPACs.

Slide 11:

So is there any value in having controlled vocabulary if everyone is doing keyword searches?

Almost 50 years ago 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," which means, of course, that just over one third of the titles did not have a match to a subject heading word.

> 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.

Carolyn Frost, comparing title words with LCSH in 1989, found that, "For 27% of her sample, there were no words from the title which matched any part of the subject heading."

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

Slide 12:

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. Keller found an overlap of 64%. Keller concluded that subject headings were still needed.

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

In 2003, Elaine Nowick and Margaret Mering compared keyword queries with three controlled vocabularies (*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."

> 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.

Slide 13:

Tina Gross and Arlene G. Taylor researched the effect of the presence of subject headings on keyword searching in 2005 and found that 36% of hits in keyword searches did not have the keywords anywhere in the records except in the subject headings.

> 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.

Our sample was limited to English language items. Also, this was before the beginning of the project to link tables of contents to bibliographic records in the catalog.

Slide 14:

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

Either:

- We should abandon controlled vocabulary in favor of keywords.

Or:

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

Slide 15:

In the last few years, there have been several calls to abandon traditional controlled vocabulary in favor of relying on keyword searching of metadata records, or indeed, keyword searching of full-text databases – just let searchers go to the full text and search that by keyword.

The Bibliographic Services Task Force of the University of California Libraries, after much discussion of their disagreements, made a recommendation to consider using controlled vocabularies only for name, uniform title, date, and place, and [to consider] abandoning the use of controlled vocabularies [LCSH, MESH, etc] for topical subjects in bibliographic records.

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

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. Her recommendation: "Abandon the attempt to do comprehensive subject analysis manually with LCSH in favor of subject keywords; urge LC to dismantle LCSH." [This is an astonishing recommendation! Let's just get rid of LCSH so no one will be tempted to use it ever again!!!]

> 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, accessed June 20, 2010, http://www.loc.gov/catdir/calhoun-report-final.pdf.

Both reports suggested: TOC and Indexes should be considered for being linked to bibliographic records.

We'll come back to this idea that tables of contents and indexes can make controlled vocabulary unnecessary.

Slide 16:

Many library administrators were pleased with these reports. Deanna Marcum, an administrator at the Library of Congress, in a discussion of how her audience should think about cataloging in the Age of Google, argued in 2005 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."

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

Suggestions to abandon LCSH were not 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), which issued a policy to keep LCSH, but to do more to simplify it, automate it, and to encourage Web application use of it.

> 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 June 20, 2010, http://www.loc.gov/catdir/cpso/pre_vs_post.html.

Then, LC appointed members from across the country to the Library of Congress Working Group on the Future of Bibliographic Control. In this group's report, they see keyword searching as an extremely useful addition to the arsenal of searching capabilities available to users, but not a satisfactory substitute for controlled vocabularies. One recommendation in this report is: "Optimize LCSH for Use and Reuse." But at the same time, the working group recommended recognizing the flaws in LCSH and working to overcome them.

> 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), p. 19, available: http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf.

Even so, the future of controlled vocabularies still seems precarious.

So what if every word of every text were to be available to be searched by keyword? Would that be acceptable?

Slide 17:

A question often asked about full text databases is: Why should there be any metadata at all, if every word of the text can be searched?

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. This article is one published in 2007 by Bradley Hemminger, Billy Saelim, Patrick Sullivan, and Todd Vision. They wrote: "Significantly more articles were discovered via full-text searching; however, the precision [i.e., % of relevant items] 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."

> 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.

Note that this refers to the number of times a term appears in a single document (i.e., an author may use the same word with the same meaning many times in one document)

One reason that full text presents difficulties for searching is explained by Zipf's Law (1949). In simple terms, as the Law applies in this situation, George Zipf observed "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." 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.

> George Kingsley Zipf, as described in Schymik, "Representational Indeterminacy and Enterprise Search," 2009, p. 4.

Sarah Hayman and Nick 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."

> 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.

How many meanings for the same word would we have in our field? If I were to ask for the meaning of "verbal," how many meanings would I get?

"Verbal" can mean:

1. Consisting of words
2. Words rather than meaning or substance
3. Spoken rather than written
4. Of, relating to, or formed from a verb
5. of or relating to facility in the use and comprehension of words
6. verbatim, word-for-word

Slide 18:

Jeffrey Garrett (2007), adding subject headings to Eighteenth Century Collections Online (ECCO), asserts that important terms and concepts are found in subject headings in metadata that cannot be found in the full text itself …. For example, 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….

> Jeffrey Garrett, "Subject Headings in Full-Text Environments: The ECCO Experiment," *College & Research Libraries*, 68, no. 1 (2007): p. 69, 75.

On the other hand, as pointed out in a 2012 article by Michael Buckland: "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." Other examples, gleaned from the work of Sanford Berman: "Gypsies are not from Egypt and prefer to be called Roma; the cross-reference 'Rogues and vagabonds see also Gypsies' exhibits prejudice. And one's own behavior is usually reflected as superior to that of others: Rebellions by slaves are named 'insurrections,' rebellions by Whites are more positively named 'revolutions.' "

> Michael K. Buckland, "Obsolescence in Subject Description," *Journal of Documentation* 68, no. 2 (2012): 154 - 161

In 2008 Sheila Bair and Sharon Carlson discussed a project to describe some digitized Civil War diaries so as to make them accessible to an audience of historians, genealogists, and others. After describing how the diaries were transcribed and tagged with names, terms, and definitions of obsolete terms, they wrote: "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." They conclude: "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."

> Sheila A. Bair and Sharon Carlson, "Where Keywords Fail: Using Metadata to Facilitate Digital Humanities Scholarship," University Libraries Faculty & Staff Publications. Paper 12, 2008, p. 2-3, p. 6, p. 15.

Slide 19:

A very large issue in full-text searching is synonyms. According to Jeffrey Beall and Karen Kafadar, "Overall, the measure of what's missed is as high as 30% in … 90% … of common word-pairs. Information discovery systems need to take the synonym problem into account and develop solutions for it,…. Additionally, the data demonstrate the value of vocabulary control and cross references in providing more precise search results."

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

Elaine Nowick, Daryl Travnicek, Kent Eskridge, and Stephen Stein, in a 2010 study, compared controlled vocabulary with keywords that had been identified by automated text analysis or word clustering techniques for documents in an online environment. They explored similarity among terms from users, from the documents themselves, and from controlled vocabularies. Their findings show that controlled vocabulary terms were better matched to users' search terms and also better matched to document terms than were 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 provide a bridge between users and relevant documents."

> 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), p. 5-6.

Slide 20:

So…, what if we were to abandon CVs?

Several writers have remarked on the fact that when controlled vocabulary is removed in favor of keyword only, the cost of subject analysis is moved to users.

George Macgregor and Emma McCulloch, discussing a 2005 blog post by Ian Davis, 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."

> Ian Davis, ''Why tagging is Expensive,'' (2005), available: http://blogs.capita-libraries.co.uk/panlibus/2005/09/07/why_tagging_is_/.
>
> George Macgregor and Emma McCulloch, "Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool," *Library Review* 55, no. 5 (2006): 296.

And Markey observed in 2007: "Because many people are searching online systems for something they do not know, their behavior is neither targeted nor direct." "Searching for Something One Does Not Know Is Frenetic, Aimless, and Random."

> Karen Markey, "The Online Library Catalog: Paradise Lost and Paradise Regained?," *D-Lib Magazine* 13, no. 1/2 (2007), accessed June 20, 2010, doi: 10.1045/january2007-markey.

Slide 21:

Some types of information resources *require at least* manually assigned keywords, if not controlled vocabulary. It should be obvious that non-textual resources are unsearchable by keyword unless *someone* adds words as metadata.

People start with Google. Markey says:

"The World-Wide Web has become the people's *encyclopedia* of choice. Google and other web search engines give people a good start, and, in fact, with Wikipedia links in hand, it gives them a *running* start, for building on their bare-bones, basic knowledge of a topic." But, "Asked about the reliability, accuracy, and objectivity of the information they retrieve on the web, people express concern, but there is little evidence that they act on their concern."

> Karen Markey, "The Online Library Catalog: Paradise Lost and Paradise Regained?," *D-Lib Magazine* 13, no. 1/2 (2007), accessed June 20, 2010, doi: 10.1045/january2007-markey.

I, personally, think Google is great! I use it all the time for quick, basic information.

"but" says Hsieh-Yee, "for more in-depth or extensive searches, the limitations of keyword searching, … result in many irrelevant items … to wade through."

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

Slide 22:

Let's look at these limitations of keyword searching that result in so many irrelevant items.

I've borrowed the list on this and the next 3 slides from Jeffrey Beall who has described, in a 2008 journal article, the ways in which keyword-based full-text searching can fail.

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

Synonyms ….

Variant spellings …

Word forms …

Different languages or dialects …

Obsolete terms …

Etc. …

Slide 23:

Homonyms …

Uncontrolled personal names – The only way I have given my name as an author for the last 30 years is Arlene G. Taylor, but various indexes use these various forms.

False cognates - the word "location" in French doesn't mean "location" in English; it means a rental or a lease.

Inability to employ facets - Pure, full-text searching fails at these tasks, because the search engine doesn't know the format (DVD's) or the subject (agriculture) or the publication date (2005) of the documents it searches.

Clustering – River banks does not eliminate resources about financial banks because a resource about financial banks can contain the word ‘river’ (e.g., ‘river of data’ or 209 River Street)

Slide 24:

Inability to sort …

Spamming - Most large search engines ignore subject metadata (often referred to here as "keywords") added into a document's meta tags for fear that it is spam…. Thus this potential added value, that is, the value of rich subject metadata, is often lost in the jungle of the World Wide Web.

Aboutness - "A classical problem for document retrieval systems is the failure of keywords to identify the conceptual content of documents.”

Kai A. Olsen, Kenneth M. Sochats, & James G. Williams, "Full Text Searching and Information Overload," *International Information & Library Review* 30 (June, 1998): 105-122.

Figurative language – about "happiness"? Or maybe about people who are divorced from reality?

Word lists …

Abstract topics …

Slide 25:

Search term not in database …

An addition to Beall's list by Thomas 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."

> 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, p. 4. Available: http://www.guild2910.org/searching.htm.

Almost every discipline, for example, has information about children and toys. But the disciplines will not be in any order or collocated in any way.

Slide 26:

Recent articles in 14 subject areas talk about the need for controlled vocabulary. I have citations for these which you can ask about afterward. I'd like to tell you about the articles from Business on the next few slides.

These subject areas are listed here in order of date of article:

WATER QUALITY

> 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.

PHYSICS

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

MEDICAL THESES

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

WOMEN'S STUDIES

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

BIOINFORMATICS

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.

GENOMICS

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, p. 655-662.

TISSUE ENGINEERING

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.

MEDICINE

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, p. 1-10.

NEUROSCIENCE

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.

BIOMEDICINE

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, p. W170-W173.

VETERINARY MEDICINE

> Kristine M. Alpi, Elizabeth Stringer, Ryan S. DeVoe, 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.

ASTRONOMY

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

CLINICAL NURSING

> 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.

Slide 27:

Karen Corral, Gregory Schymik, Robert St. Louis, David Schuff, Ozgur Turetken, in various combinations, have written four articles about enterprise search systems in business corporations.

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…."

> 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, p. 2.

[†] *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."*

Slide 28:

They cite Google for "data indicating that knowledge workers are wasting almost half of their time as a direct result of failed searches."

"[Workers] 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…."

Some workers give up and re-do the work.

> 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, p. 1.

Slide 29:

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 have presented 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."

> 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), p. 146-148, 156

The authors 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 this improved search.

Slide 30:

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."

> 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.

Slide 31:

Let's look at each of these additional solutions separately.

Slide 32:

Use Both!! I've tried to teach this to students ever since online catalogs emerged. A number of recent articles also make this point.

Jack Leong argued in 2010 that the somewhat separate areas of metadata schemas and bibliographic control are converging. 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."

> 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.

Slide 33:

Another suggested solution to the keyword versus controlled vocabulary dilemma is to make use of tagging systems. George Macgregor and Emma 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."

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

Folksonomies are touted because of the perception that no formal thesaurus can keep up with user needs and therefore new terminology is added faster. Sarah Hayman and Nick Lothian propose that if we observe the taggers' terms, [this] will give us a rich source of information for developing our formal systems so that we can indeed get the best of both worlds."

> 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.

It turns out that tags and folksonomies have the same issues already identified with keyword searching. Macgregor and McCulloch observe that "[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."

In 2011 Jo Bates and Jennifer Rowley examined LibraryThing's folksonomy from a British perspective and found it dominated by United States taggers, which has an impact especially 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."

> 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): 431.

Slide 34:

Furthermore, according to several authors, tagging has the additional issue of tags that are personal (e.g., 'to read'), are silly, or are purposely misleading.

Peter Rolla wrote in 2009 that a "comparison of LibraryThing's user tags and LCSH suggests that while user tags can enhance subject access to library collections, they cannot replace the valuable functions of a controlled vocabulary like LCSH…. 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."

> 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.

Finally 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."

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

Slide 35:

The third suggested solution is to develop prototype tools.

Karen Markey (2010) – "Before mass digitization projects make significant headway, the library community must act on building the future online catalog, joining forces with researchers, practitioners, and system designers in related and allied fields to: (1) gather relevant information, (2) test prototype post-digitization-era catalogs, (3) evaluate results and make decisions, (4) assign tasks to willing parties [a task being accomplished by the DPLA (Digital Public Library of America)], and (5) execute them."

> Karen Markey, "The Online Library Catalog: Paradise Lost and Paradise Regained?," *D-Lib Magazine* 13, no. 1/2 (2007), accessed June 20, 2010, doi: 10.1045/january2007-markey.

Among the first tools provided to accomplish the purpose of helping end users with subject searching have been various unified 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.

> 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.

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)."

> 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, p. 1-10.

A third unified ontology is described by Noy, et al., who write: "The Open Biomedical Resources (OBR) component automatically indexes the metadata for important biomedical data sets available online …, and links the underlying data sets to the terms in the ontologies in BioPortal and UMLS. OBR allows biomedical investigators to use the terms in the BioPortal ontologies to enhance their ability to search for relevant online data in a manner that is not possible with conventional keyword search strategies."

> 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, p. W170-W173.

Example ontology – WordNet

Slide 36:

WordNet definition of "ontology"

Slide 37:

Vivien Petras introduced in 2006 a "search term recommender," based on statistical associations between specialized language terms and controlled vocabulary terms.

> Petras, Vivien, "Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages," (2006), p. 1

Gilles Hubert and Josiane Mothe proposed in 2009 a search engine that will integrate both "browsing an ontology (via categories)" and "defining a query in free language (via keywords)."

> 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.

Julien and Cole (2009) described 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.

> 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.

Julien, Catherine Guastavino, France Bouthillier, and John Leide, in 2010, 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."

> 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.

Slide 38:

The fourth suggestion (and the one that some library administrators have believed would allow them to do away with subject cataloging) is to add tables of contents and summaries or abstracts linked to bibliographic records, using publisher data already in existence.

Users seem to like having summaries and contents notes available, as is evident from their 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 recommended that the UC Libraries should: "Consider whether automated enriched metadata such as TOC [and/or] indexes can become surrogates for subject headings and classification for retrieval."

> Bibliographic Services Task Force of the University of California Libraries, "Rethinking How We Provide Bibliographic Services," 2005, p. 23-24.

OCLC's 2009 report also shows that users expect to find enriched metadata: "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*." [italics original] But the report then gives voice to concerns about cost of providing subject headings, saying that it may be necessary for libraries

to find more economical means to achieve the benefits to end users that controlled subject vocabularies provide.

> OCLC, "Online Catalogs: What Users and Librarians Want: An OCLC Report," (Dublin, Ohio: OCLC Computer Library Center, 2009), p. 17, p. 48, p. 52.

Subject rich words found in summary notes and tables of contents help **recall** (many retrievals, both relevant and not relevant), but they cause a problem for **precision** (fewer retrievals, but most of them relevant), because the terminology is not controlled.

Slide 39:

Research continues to suggest that controlled vocabularies are needed to provide unique search terms that are 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 that a notable result occurred when keywords and LCSH were matched against abstracts. Almost one-third of the assigned LCSH were unique to the bibliographic records, even in the presence of the abstracts."

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

McCutcheon also looked at the possibility of using the 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 many cases the student authors omitted title words, misspelled words, and misrepresented symbols and diacritics. So she concluded that keyword access alone cannot suffice for retrieval by subject.

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

In a 2012 publication, Schwing, McCutcheon, and Maurer replicated Strader's research using electronic theses and dissertations in another catalog. The authors found that in 24% of cases a term in the LCSH was either a variant of the term in the abstract (and therefore would not be matched to a keyword search using the abstract's term), or did not exist in the abstract at all.

> 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.

Slide 40:

As already mentioned (slide 13), Gross & Taylor – 2005 – 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." The results were persuasive, but two key criticisms were raised. Some argued that the 2005 study should not have been limited to English language materials. Others suggested that the addition of links to tables of contents and summary notes in catalog records would affect the results of the study and minimize the need for controlled vocabulary.

Our current 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 second difference was that the study looked at search results that included materials in all languages, not just English language materials. The average percentage of hits that would be lost in the absence of subject headings in a catalog with summary and contents data enrichment was 28 percent. So the drop was from 36% to 27% (8%).

For about 20% of the searches, though, the percentage of hits with a keyword only in a subject field was 50% 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.

The average proportion of hits that would be lost appears to increase as the number of keywords increases. Searches with three keywords would lose an average of 37% of retrieved hits if the subject fields were not present. Searches with four or more keywords would lose an average of 40% of retrieved hits.

Slide 41:

How can we incorporate the use of controlled vocabularies into algorithms to create search results? SKOS

Slide 42:

So, do we need controlled vocabulary in the age of Google? Really? For most searches, such as those to win a bet about a date or other defined piece of information, No. But it appears that serious researchers and scholars, along with businesses that have enterprise systems, can still benefit from some form of controlled vocabulary.