Controlled Vocabulary in the Age of Google? Really?

Presentation for the School of Information Sciences, University of Tennessee – Knoxville
November 7, 2012

Arlene G. Taylor

Keywords

- A **keyword** is a term that is chosen, either from actual text or from a searcher's head that is considered to be a "key" to finding desired information
- **Keyword searching** is the use of one or more keywords as the intellectual content of a search command

2

Controlled Vocabulary (CV)

- … is a **list of terms** in which terms or phrases representing a concept are brought together.
- Often, a **preferred term** or phrase is designated for use in metadata records in a retrieval tool.
- **Terms not to be used** (e.g., synonyms, near-synonyms) often have references from them to the chosen term or phrase.
- **Relationships** (e.g., broader terms, narrower terms, related terms) among used terms may be identified.
- **Scope notes** may explain terms

3

Entry from LCSH

Sexism ***(May Subd Geog)***
Here are entered works on sexism as an attitude as well as works on attitude and overt discriminatory behavior. Works dealing solely with discriminatory behavior directed toward both of the sexes are entered under **Sex discrimination**.

UF	Gender bias
	Sex bias
BT	Attitude (Psychology)
	Prejudices
	Sex (Psychology)
	Social perception
RT	Sex role
NT	Heterosexism
	Sex discrimination

4

Subject Searching

- Searching with CV is traditionally called Subject Searching
- Left-anchored searching is very difficult for most users – requires finding correct term or phrase in the CV first
- Most catalogs allow a user to find a desired subject heading in a record and to click on that to retrieve records that have been assigned that subject heading.

5

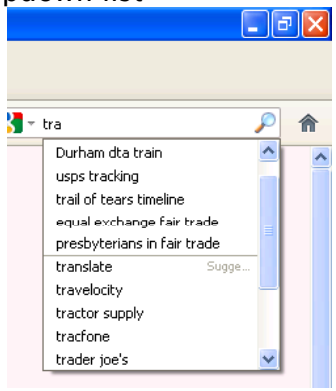
Controlled Vocabulary

- Includes
  - Subject heading lists, e.g., Library of Congress Subject headings (LCSH)
  - Thesauri, e.g., Art & Architecture Thesaurus (AAT)
  - Ontologies, e.g., WordNet
  - Classification/Categorization schemes, e.g., Dewey Decimal Classification (DDC), National Library of Medicine (NLM) classification scheme, taxonomies
  - Drop down list within a searching text box, e.g., Google

6

"Controlled vocabulary"

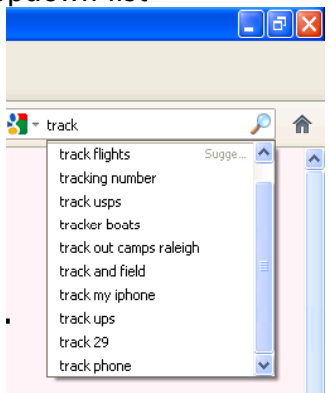
- Example Google dropdown list



7

"Controlled vocabulary"

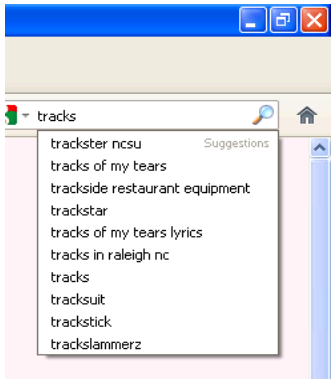
- Example Google dropdown list



8

"Controlled vocabulary"

- Example Google dropdown list



9

Controlled Vocabulary vs. Keyword Searching

- Online public access catalogs (OPACs) – difficult for patrons to use, partly due to the complexity of subject searching
- Keyword searching has become the preferred method of conducting a search in any online system

10

Is CV Useful in a Keyword search?

- First research almost 50 years ago – 1964 – Kraft investigated KWIC indexing of titles
  - 64% of title entries contained as keywords one or more of the subject heading words
  - [note that this means that just over one third of titles did not contain one of the subject words]
- 1989 – Frost compared title words with LCSH
  - 27% of title entries did not match subject heading words
  - [note that this means that a little under one third of titles did not contain one of the subject words]

11

Is CV Useful in a Keyword search?

- 1992 – Keller compared titles of master's theses with the first word of LCSH on their bibliographic records
  - Found 64% overlap
  - [which means that 36% (a little over one third) did not match]
- 2003 – Nowick and Mering compared keyword queries with LCSH and 2 thesauri
  - Between 30% and 40% of the queries were exact matches to a term in one of the controlled vocabularies
  - [about one third again!]

12

Is CV Useful in a Keyword search?

- 2005 – Gross and Taylor researched effect of the presence of subject headings on keyword searching
  - 36% of hits in keyword searches did not have the keywords anywhere in the records except in the subject headings.
  - [just over one third again!!]
  - [sample included English only and study was completed before the catalog had tables of contents linked to records]

13

CV vs. Keyword

- Literature on CV vs. keyword seems to fall into two groups
- Either:
  - We should abandon controlled vocabulary in favor of keywords
- Or:
  - Successful keyword searching relies on controlled vocabulary

14

Abandon Controlled Vocabulary?

- 2005 – Bibliographic Services Task Force of the University of California Libraries
  - Agreed that CV is still needed for name, uniform title, date, and place
  - Did not agree that CV is still needed for topical subjects
- 2006 – Calhoun – report to LC on changing nature of the catalog – users do not like LCSH – she recommended abandoning and dismantling LCSH
- Both suggested linking tables of contents (TOC) and indexes, instead of creating subject headings and classification in metadata records

15

Abandon Controlled Vocabulary?

- Many library administrators were pleased with these reports – it would save lots of money not to pay for subject analysis – D. Marcum
- 2007 – LC Cataloging Policy and Support Office – keep LCSH, but do more to simplify it, automate it, and to encourage Web application use of it
- 2008 – LC Working Group on the Future of Bibliographic Control– optimize LCSH (i.e., keep the best while trying to overcome flaws)

16

Why Any Metadata at all, if Full-Text Searching is Available?

- If every word of a text can be searched, why do we need metadata?
- Among numerous research articles on full-text searching, only one suggests that full-text searching alone may be sufficient
  - Hemminger et al. – 2007 – used the number of hits of the search term in the text to rank the importance of **a document**
- Zipf's Law (1949) – the number of meanings a word has in a given **collection of documents** is roughly the square root of the number of times the word appears in that set of documents

17

Full-text Searching Limitations

- Importance of CV in full-text historical collections
  - Garrett – 2007 – ECCO – certain concepts don't use same words now as they did historically
  - Buckland – 2012 – cultural changes of acceptable terminology over time
- Importance of CV in digitized collections of primary materials (e.g., diaries)
  - Bair & Carlson – 2008 – Civil War diaries – words like "cotton", "hill", and "wood" are also names of people and places

18

Full-text Searching Limitations

- Synonym problem – high for people involved in scholarly research
  - Beall & Kafadar – 2008 – common word synonym pairs – 30% of relevant items missed
- Nowick et al. – 2010 – study comparing CV with automated text analysis and with word clustering techniques – controlled vocabulary matches user search terms better than do either of the automated sources of keywords, and also matches documents better than either. **CV is a bridge between users and relevant documents**

19

What if We Abandon Controlled Vocabularies?

- Cost moved to users
  - 2006 – Macgregor, McCulloch, Davis – any economies achieved by not doing subject analysis or classification are simply moved onto the price users pay in time to find what they need
  - 2007 – Markey observed that people often search online systems for something they do not know and such searching is often chaotic, aimless, and random

20

Can Keyword Searching Alone Be Successful?

- CV needed for non-textual resources (e.g., pictures, films, class-lecture recordings, etc. and variant-language resources)
- CV needed for scholarly research
  - People start with Google, where Wikipedia links to both online and print sources
  - Is this information complete, reliable, accurate, and objective?
  - For in-depth or extensive searches, limitations of keyword searching (e.g., no control over synonyms, spelling, language, etc.) result in many irrelevant items to wade through

21

Reasons Against Relying on Keyword Searching

- Failures of keyword searching
  - **synonyms** (rate, price, cost, charge, tariff, tab)
  - **variant spellings** (color, colour; donut, doughnut)
  - **word forms** (essay, essays; web site, website)
    - Abbreviations, acronyms, initialisms (Tyrannosaurus Rex, T-Rex)
  - **different languages or dialects** (football in American English is different from football in British English)
  - **obsolete terms** ('French distemper' is archaic way of referring to syphilis)

22

Reasons Against Relying on Keyword Searching

- Failures of keyword searching
  - **homonyms** ('boxers': dogs? sports persons? underwear?)
  - **uncontrolled personal names** (AG Taylor, A.G. Taylor, Arlene G. Taylor, Arlene Taylor)
  - **false cognates** ('location' in French, not same as 'location' in English; 'die' not same in German and English)
  - **inability to employ facets** (can't find "all DVDs on agriculture published after 2005")
  - **clustering** ('river banks' does not eliminate resources about financial banks)

23

Reasons Against Relying on Keyword Searching

- Failures of keyword searching
  - **inability to sort** (can't give oldest first or most recent first)
  - **spamming** (addition of text to the meta tags for a document to cause that document to be retrieved when it is actually irrelevant)
  - **aboutness issues** (individual words [e.g., 'debate'] in a resource do not necessarily map to conceptual content [e.g., resource about debates])
  - **figurative language** ("she's floating on a cloud")
  - **word lists** (keywords may match to many scanned dictionaries, glossaries, indexes)
  - **abstract topics** (good health, free will)

24

Reasons Against Relying on Keyword Searching

- Failures of keyword searching
  - **search term not in database** (descriptions of a person's political career may not use the words 'political career')
  - **search term unknown** (searcher does not know scientific or medical term for a concept)
  - **paired topics are difficult to search** ('Art and mental illness')
  - **Can't distinguish right words in conceptual contexts apart from the appearance of the same words in the wrong contexts** ('children' and 'toys' may appear in resources from almost every discipline)

25

CV in Particular Fields of Study

Recent articles in fourteen subject areas indicate that CV should be used when searching databases in these disciplines:

Astronomy	Bioinformatics	Biomedicine
Business	Clinical Nursing	Genomics
Medical theses	Medicine	Neuroscience
Physics	Tissue engineering	Veterinary Medicine
Water quality	Women's studies	

26

Business and "enterprise search"

- Many corporations have "enterprise search systems"
  - "Enterprise Search" – term used to describe software for searching full-text information within an enterprise (business or corporation)
    - File systems, intranets, document management systems, e-mail, databases
  - Contrasted with web search, which searches documents on the open web
  - Enterprise systems may have both structured and unstructured data

27

Business Management

- The authors cite Google as the source of the following data:
  - Knowledge workers waste almost half their time as a direct result of failed searches
  - Another 25% of time is spent conducting "successful" searches for information
  - About one quarter of time then is spent on truly value added activity
- Middle managers say that often, the information found in "successful" searches is wrong
- Some workers give up before finding a known document and re-do the work that was already done in the not-found document

28

Business Management – cont.

- Model to justify the up-front cost of determining and entering the CV data
  - Entering subject metadata eliminates ambiguity of words through pre-defined categories
  - Reduces the number of irrelevant documents returned in the result set
  - Results imply that it is cost effective for almost any business organization to implement
  - Break-even point: for a firm with 1000 employees and 100,000 documents in the database, an average of 25 searches per employee would justify the cost of encoding the metadata

29

Business Management and CV

Corral, et al. (2010) – experiment that measured the impact of adding subject metadata to keyword-based full-text searches:

"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."

30

Additional Solutions Offered for the CV vs. Keyword Dilemma

- Both CV and keyword searching
- Tagging Systems
- Prototype Tools
- Addition of TOC and Summaries/Abstracts
- SKOS (Simple Knowledge Organization System)

31

Use Both CV and Keyword Searching

- Numerous writers have urged that CV and keyword searching be used together – thus making use of the strong points of each
- 2010 – Leong – areas of metadata schemas and bibliographic control are converging
- "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."

32

Tagging and Folksonomy Systems…

- Have become ubiquitous on the web
- Has the advantage of being able to provide terminology for new technology and new events of yesterday
- Have same issues as found with keyword searching
  - No control over synonyms, homonyms, abbreviations, etc.
  - Bates and Rowley from British perspective find LibraryThing to be dominated by U.S. taggers

33

Tagging and Folksonomy Systems…

- Have the additional issue of tags that are personal, silly, or purposely misleading
- Should supplement CV, not replace it
- Tagging and a reaffirmation of CV have arisen in parallel

34

Prototype Tools

- Markey – 2010 – Create searching tools to find appropriate search terms that will both satisfy an information need and match the language used in the information system
- Unified ontologies and integrated controlled vocabularies; examples:
  - Ontology Lookup Service (OLS)
  - Unified Medical Language System (UMLS)
  - Open Biomedical Resources (OBR)
- Example ontology
  - Wordnet – "lexical database for English"

35

Wordnet

WordNet Search – 3.1– WordNet home page – Glossary – Help

Word to search for: ontology

Noun

- S: (n) **ontology** ((computer science) a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations)
  - *domain category*
    - S: (n) computer science, computing (the branch of engineering science that studies (with the aid of computers) computable processes and structures)
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) arrangement, organization, organisation, system (an organized structure for arranging or classifying) *"he changed the arrangement of the topics"; "the facts were familiar but it was in the organization of them that he was original"; "he tried to understand their system of classification"*
- S: (n) **ontology** (the metaphysical study of the nature of being and existence)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*

36

More Prototype Tools

- Petras – 2006 – Search term recommender
- Hubert & Mothe – 2009 – Integrate both browsing an ontology and defining a query in free language
- Julien & Cole – 2009 – design and development of an interactive visual map of a collection's major subject headings and their relations as a complement to keyword searching
- Julien et al. – 2010 – 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

37

Addition of TOC and Summaries/ Abstracts

- Supplies additional words for keyword searching
- Users like summaries and contents notes – evident from use of sites such as Amazon.com
  - U.Calif Report – 2005 – consider whether TOC or indexes can substitute for CV
  - OCLC – 2009 – Users want more subject information, including tables of contents and summaries/abstracts
- Great for **recall** (many retrievals, both relevant and not relevant), but a problem for **precision** (fewer retrievals, but most of them relevant)

38

Addition of TOC and Summaries/ Abstracts – Research

- Research continues to suggest a need for CV to provide additional unique search terms not available in TOC and summaries
  - Strader – 2009 – 31% of the time, LCSH are unique when compared with abstracts of dissertations
  - McCutcheon – 2011 – in author-supplied metadata, authors omit title words, misspell words, and misrepresent symbols and diacritics
  - Schwing, et al. – 2012 – 24% of the time, LCSH are either variants of words in the abstract or do not appear at all in abstracts of theses and dissertations

39

New Study by Gross, Joudrey, Taylor

- With TOC/summary data enrichment
- All languages
- In a catalog with TOC linked, 28% of hits would be lost if there were no subject headings – a drop of 8% from the first study
- Still, more than one quarter of hits would be lost
- And in many individual searches (20% of searches), the loss is 50% or greater
- Searches with 3 keywords lose 37%, and searches with 4 or more keywords lose 40%

40

SKOS (Simple Knowledge Organization System)

- http://www.w3.org/2004/02/skos
- developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web

41

Controlled Vocabulary in the Age of Google? Really?

- A search to win a bet? No CV needed!
- A search for a quick definition? No CV needed!
- Business enterprise system? Yes!
- Serious research? Yes!

42

Thank you!

Questions?

ataylor@sis.pitt.edu

43