Summer 9-13-2013

Wikipedia-type Disambiguation Functionality in LCSH: a Recommendation

Daniel CannCasciato
Central Washington University, dcc@cwu.edu

Wikipedia-type disambiguation functionality in LCSH: a recommendation

Head of Cataloging
CWU Brooks Library

Abstract: A proposal for a modest change in LCSH practice (and RDA when chapter 23 is written) to provide a Wikipedia-type disambiguation function for subject headings in library catalogs. Such a change is cost effective, scalable, familiar, and system-agnostic. Furthermore, it is necessary for fulfillment of our catalog objectives.

Keywords: Disambiguation; LCSH; Library of Congress Subject headings; RDA; Resource Description and Access;

Salsa

If a patron speaks or types the word "Salsa", without further qualification, then a confident definition of the term cannot be achieved. Salsa might be a dance, a musical style, or sauce of Spanish, Italian or Latin American origin. In other words, without some clarification, one is left with ambiguity. The process of clarification is generally called disambiguation. In the work of cataloging, especially subject analysis, disambiguation plays a prominent role in the establishment of terms in the *Library of Congress Subject Headings* authority file (LCSH).

Disambiguation by the use of qualifiers or other modifications is commonly used when a term can have multiple senses: the salsa example is one type. In LCSH there are three headings:

Salsa (Dance)
Salsa (Music)
Salsas (Cooking); this has a see reference from of *Salsa (Cooking)*

Additionally, the reason many LCSH terms have qualifiers (e.g.: *Ground reaction force (Biomechanics)*) is to clarify the context for the term and to disambiguate (clarify in context) its meaning. In the example given, a ground reaction force might be some type of Special Forces military group. That there is no existing heading for such a group or practice at this time does not eliminate the need for the qualifier; it is added just the same. (This practice is covered specifically in the *Subject Heading Manual* instruction sheet H357.)

However, neither of these examples achieves what a utilization of a clearer disambiguation practice would. The purpose of this paper is to recommend that when RDA tackles subject access, it institute a practice of providing Wikipedia-type disambiguation see-references of all ambiguous terms, rather than the somewhat inconsistent or non-existent practices currently in evidence in LCSH. Further, I propose that it is already possible for LCSH policies to be changed to implement such a practice and that such a change happen. Using *Salsa* again as the example, a Wikipedia-type disambiguation practice would make a reference presentation to our patrons of:

Salsa
1--> See Salsa (Dance)
2--> See Salsa (Music)
3 --> See Salsas (Cooking)

at the *top* of a results page to a search query. Enabling and supporting such a presentation of choices and guidance to our patrons assists them, and supports our traditional and our new cataloging objectives.

CATALOGING'S OBJECTIVES

Cataloging's objectives have remained steady more or less since Cutter first described them. *Resource description and access* (RDA) carries them through in some manner. (Cf. RDA 0.2.)

To paraphrase Cutter's Objects, the catalog's purpose is:

> ***To enable a person to find*** *a book of which the subject is known …* ***To show what*** *the library has on a given subject, and …* ***To assist in the choice*** *of a book as to its character (literary or topical)*

To paraphrase from RDA's purpose and scope:

> *RDA provides a set of guidelines and instructions on formulating data to support resource discovery. The data created using RDA to describe a resource are designed to assist users performing the following tasks:* find*—i.e., to find resources that correspond to the user's stated search criteria (0.0)… The data should enable the user to:* find *resources that correspond to the user's stated search criteria …* find *all resources on a given subject (0.4.2.1) … The data should meet functional requirements for the support of user tasks in a cost-efficient manner. (0.4.2.2).*

Very different language is used across the years, but the overall intent (perhaps *mandate* is a better word) is very clear: connect the patron with the information resources available based on the user's stated search criteria. When that stated criteria is vague, or demonstrates a lack of awareness of the vastness of resources potentially available, then it falls to us to assist the user, the patron, to be successful.

DISAMBIGUATION

This issue of ambiguity of terminology (including those due to homonymy and synonymy) has been well noted in information retrieval. Cutter addressed the matter in his rules in 1904 in a manner that might well be considered quaint at this time. Ide and Veronis provide a good overview of the issue (termed "word sense disambiguation" in computational linguistics, from 1998 (Ide, 1998). Beall and Kafadar (2008) have addressed the issue of ambiguity as well, utilizing a web environment rather than that of a library catalog; they examined the impact of the synonymy problem and the results (retrieved and omitted) presented to users.

The concept of disambiguation (especially name disambiguation) has been addressed by many writers. Elliott provides a good overview of both the topic and some projects that are addressing the issue. Bazzanella [et al.] looked at the issue of entities (persons, locations, events) and a means for disambiguation in a web environment (Bazzanella, 2011. Cota [et al.] looked at names and their appearance in bibliographic citations and they too worked on automated means of producing relevant results. Roman [et al.] also looked at the issue of names. Thomas examined the limitations of library catalogs in providing assistance to users and then examined the Web service provided that do a better job (e.g.; IMDb and Wikipedia). In asking how much information is needed to help a patron resolve the ambiguity, he wrote: "the descriptive phrase needs to be long enough to include the trigger words most likely to be recognized as relevant, but short enough to be displayed at the point of need." (Thomas, 2011; p. 227) On a separate topic, but showing the influence of Web services, Faiks [et al.] considered the idea of Google-izing the catalog (Faiks, 2007).

A complicating factor that has been noted is that of the size of the database being searched. A larger database creates more ambiguity simply by having more uses of terms in various contexts, for example the various meanings of the term "school," as Cutter pointed out (Cutter, 1904, p. 71) though even smaller database will present the same problem, though perhaps less frequently. Ward writes of his experience working with a very focused database (on powertrains) and the providing access to his patrons. Apparently, clarifying a user's request for an article he'd seen about a tank was in reality a request for an article about a howitzer. He, Ward, then noted in the database record for that article the it "Looks like a tank." (Ward, 2000, p. 69.). Even what might be considered a small niche can acquire its own taxonomy. In 1977, Roberts [et al.] did a sociological study that discussed the naming and categorization of used cars as practiced in a small neighborhood.

There's the issue of the increasing volume in a collection which can drive the need for precision of terms. As Buckland states it, ""Collections of millions do need detailed description in order to achieve sufficient fineness of sifting to select a handful rather than a flood of records." (Buckland, 2012, p.155). And Beall noted that search fatigue can hamper a patron's search success or even cause it fail.

SUBJECT CATALOGING PRACTICE

Regarding subject practice, RDA at the moment reads:

> *[Section 7 Chapter] 23*
> *GENERAL GUIDELINES ON RECORDING THE SUBJECT OF A WORK*
> *[To be developed after the initial release of RDA]*

This is nearly 30 years after William Studwell's article "Why not an "AACR" for Subject Headings?" Keep in mind some training materials already in existence for subject practice education is well over 600 pages in extent (*Basic subject cataloging using LCSH***,** 2007). And despite Gregor's and Mandel's plea from 1991 for simplification of descriptive and subject cataloging, simplification appears unlikely to be realized anytime soon. Studwell, I believe, still waits for an answer (Studwell, 1995).

In light of the voluminous rules already in existence, then I am making a very modest proposal to what will be a large undertaking, however, whenever, if-ever Chapter 23 is realized. Yet I believe this proposal would have a positive impact and should be considered. It could be implemented under the current procedures for LCSH but should definitely be included in any new manual that is created.

According to an email from Janis Young of May 6, 2013, to the SACO participant's discussion list, the Subject Heading Manual (SHM) uses *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, NISO standard.

An illustration from that standard illustrates the issue of ambiguation very well: (From section 5.3.1, 2005 , p. 13)

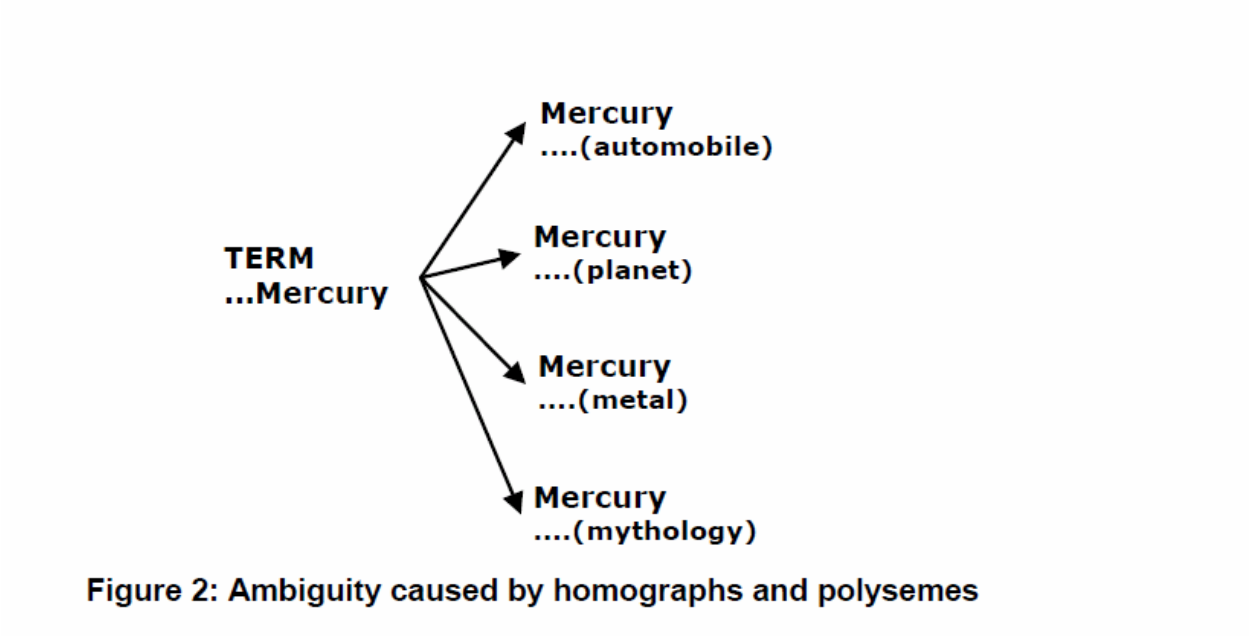


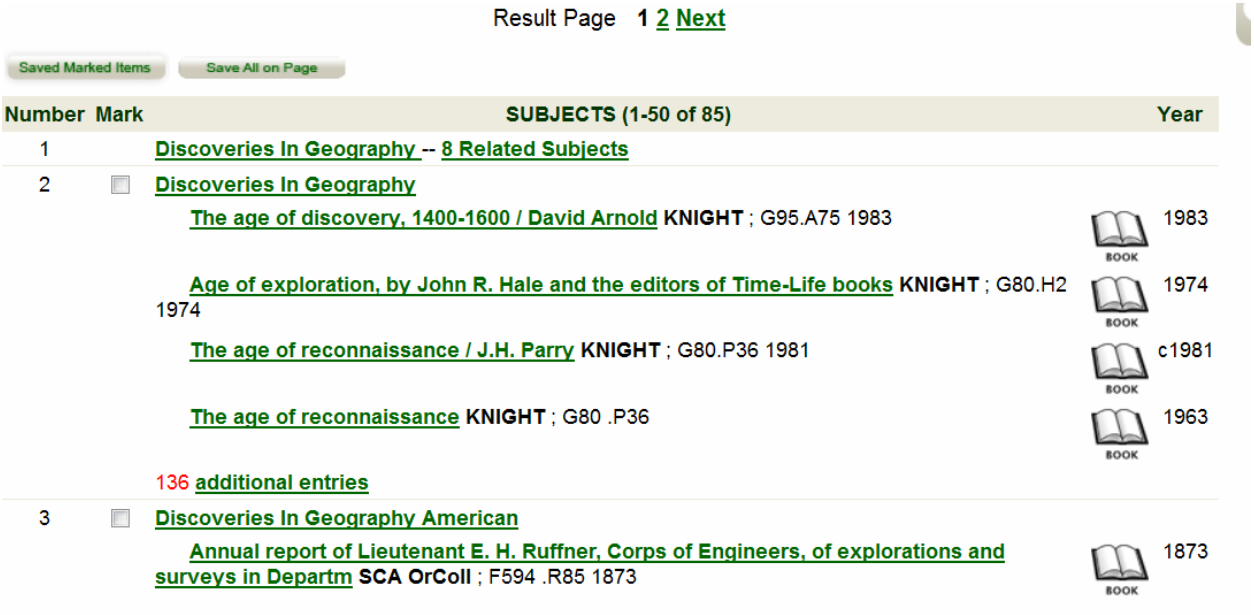
Figure 2: Ambiguity caused by homographs and polysemes

My modest proposal, then, is to institute something similar to the multi-arrow reference from one term of many possible meanings to those other terms. In short,

NO	**YES**
Salsa (Dance) *Salsa (Music)* *Salsas (Cooking)*	Salsa 1--> See Salsa (Dance) 2--> See Salsa (Music) 3 --> See Salsas (Cooking)

The reason a disambiguation practice is needed is due to what Yee and Layne identified as the filing elements used in online systems (Yee and Layne, 1998, p. 170). Many online cataloging systems ignore punctuation and display results as a normalized text string. In the current practice, a catalog search of the term salsa will require that a patron page through (or scroll through) numerous listings to move from *Salsa (Dance)* to *Salsas (Cooking)*. So even though each term is differentiated and unambiguous, much of the work of discovery of pertinent resources is left to the patron. As Beall has identified, search fatigue (overload) can hinder a

patron from being successful. A more illustrative example of this potential can be seen when doing a search for "Discoveries." An Innovative Interfaces catalog returned a display of:



Result Page 1 2 Next

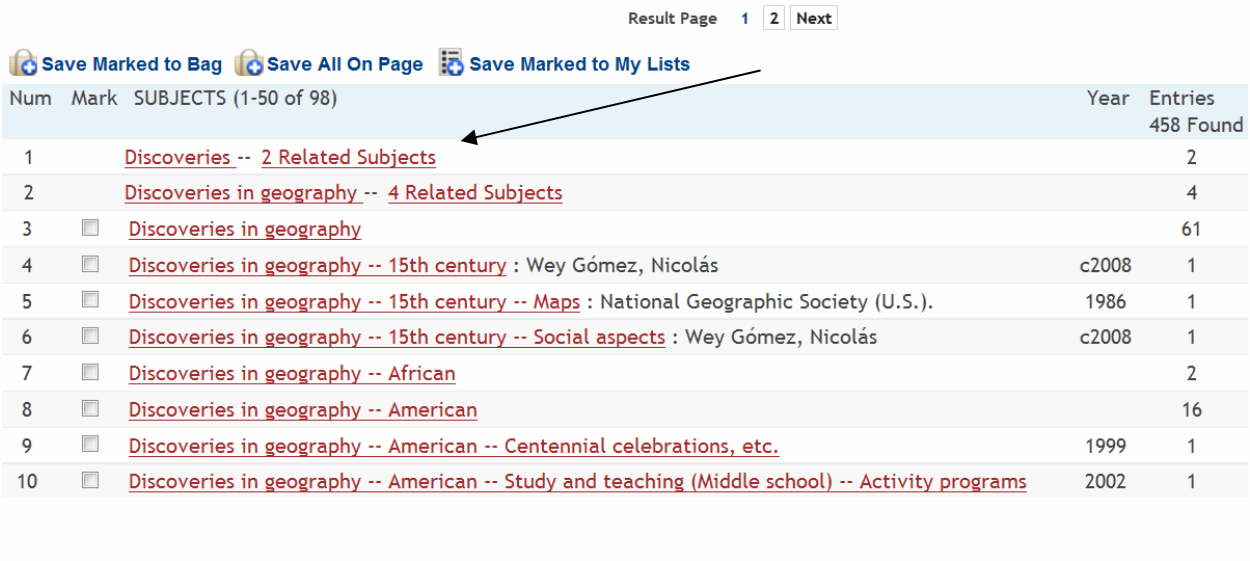
Saved Marked Items | Save All on Page

Number	Mark	SUBJECTS (1-50 of 85)	Year
1		Discoveries In Geography -- 8 Related Subjects	
2	☐	Discoveries In Geography	
		The age of discovery, 1400-1600 / David Arnold KNIGHT ; G95.A75 1983	1983
		Age of exploration, by John R. Hale and the editors of Time-Life books KNIGHT ; G80.H2 1974	1974
		The age of reconnaissance / J.H. Parry KNIGHT ; G80.P36 1981	c1981
		The age of reconnaissance KNIGHT ; G80 .P36	1963
		136 additional entries	
3	☐	Discoveries In Geography American	
		Annual report of Lieutenant E. H. Ruffner, Corps of Engineers, of explorations and surveys in Departm SCA OrColl ; F594 .R85 1873	1873

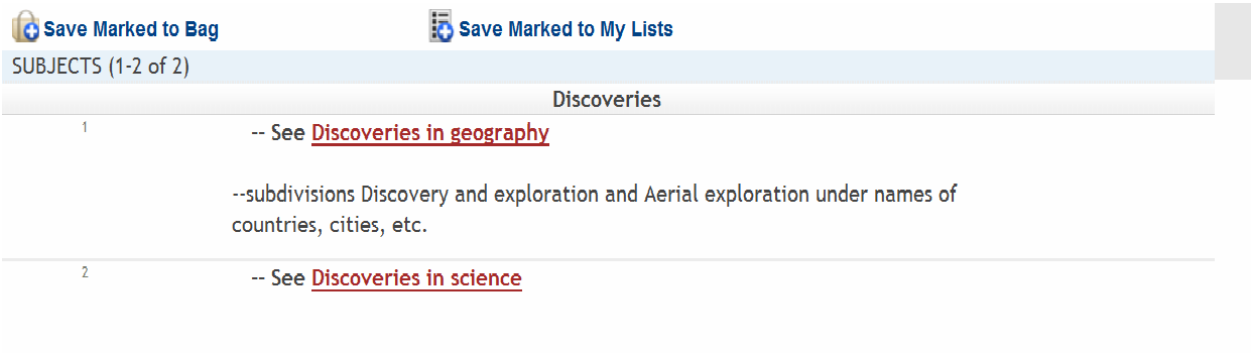
Which, unfortunately gave no hint that further down the list were materials with the heading *Discoveries in science*, on row 59 some materials are listed.

57	☐	Discoveries In Geography Spanish History	
		Atlas de los exploradores españoles / [Luis Conde-Salazar Infiesta, editor literario y redactor ; Man KNIGHT ; G288 .A85 2009	c2009
58	☐	Discoveries In Geography Spanish History Sources Translations Into Italian	
		Nuovo mondo. Gli spagnoli, 1493-1609 / a cura di Aldo Albònico e Giuseppe Bellini KNIGHT ; E123 .N86 1992	c1992
59	☐	Discoveries In Science	
		Abduction, reason, and science : processes of discovery and explanation / Lorenzo Magnani SCIENCE ; Q175.32.A24 M34 2001	c2001
		Accelerating scientific discovery through computation and visualization [microform] / James S. Sims . DOC-US/MF ; C 13.58:6709	2001
		Accelerating scientific discovery through computation and visualization [electronic resource] / James C 13.58:6709	2001
		Accelerating scientific discovery through computation and visualization II [microform] / James S. Sim DOC-US/MF ; C 13.58:6877	2002

However, for disambiguation purposes and guidance, I'm suggesting we need to provide a see-reference in the authority record that then creates a line up at the top of the results that gives the patron a hint of the possibilities:

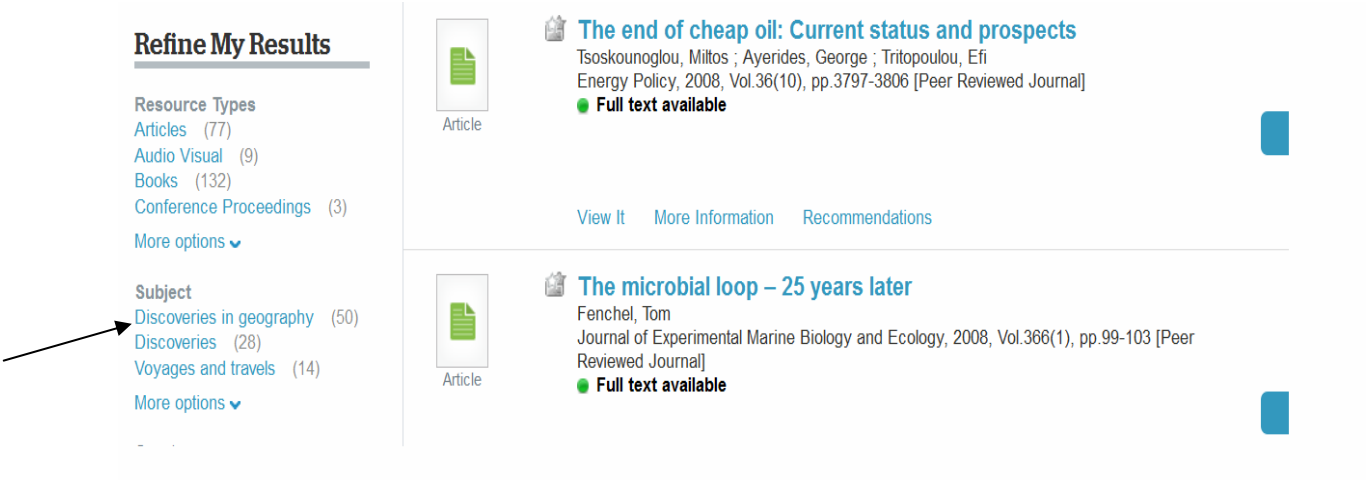


When clicked, the ILS would display to the following information to the patron:

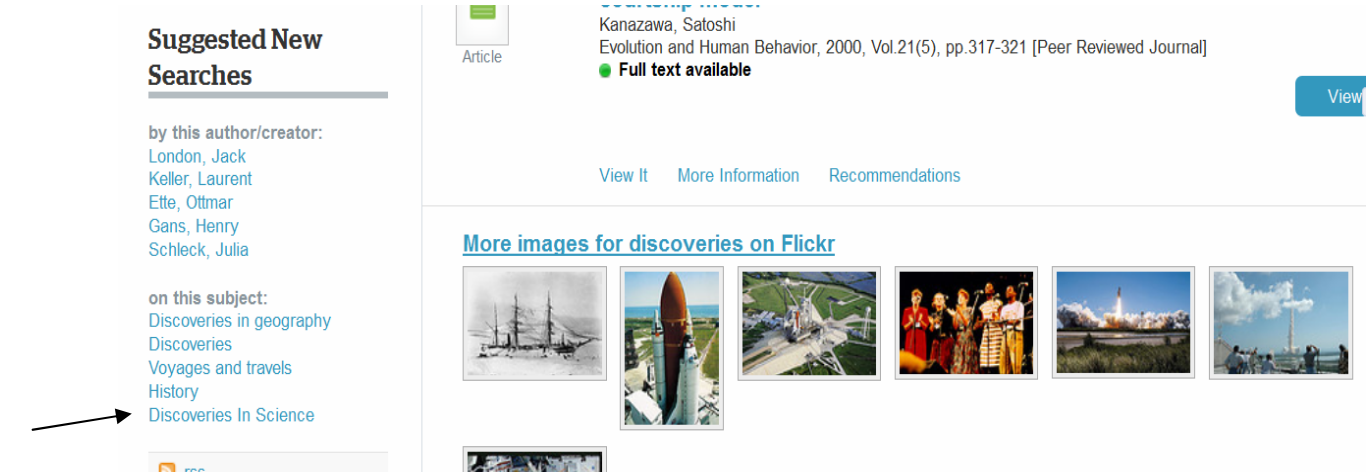


This simple change alerts the patron of the possibilities and allows her or him to jump directly to pertinent results. The result of the ambiguous search, rather than providing only potentially wearisome or *hidden* results, also provides the patron with assistance. The see-references help to disambiguate - - if not the search - - then the breadth of the displayed results.

In some other catalogs, the results of a search are more difficult to refine. In a Primo/Alma catalog the results of an *advanced* search brought about an option for refinement only for "Discoveries in geography."

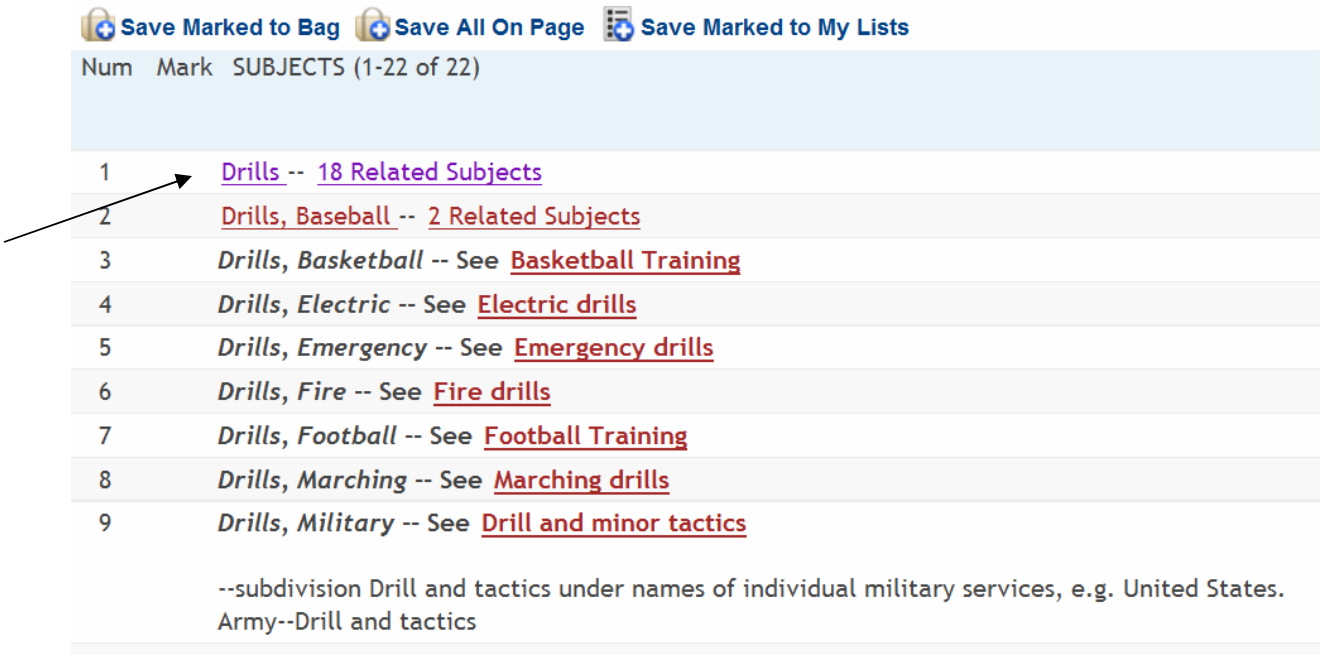


It wasn't until further down the left navigation pane that the second subject was made apparent, *Discoveries in science*.



The results of a browse search in the Primo/Alma catalog looked much as the results in the Innovative Interfaces catalog.

Another illustrative example using more homonyms, using the Innovative Interfaces catalog is for the heading "*Drills*": In LCSH the term does not stand alone, yet for purposes of disambiguation-type display, there could be a see reference on numerous authority records, creating a guide to the literature held (and the multiplicity of topics that could be covered by the term, such as:



Here's a partial image of the 18 related subjects

SUBJECTS (1-18 of 18)

	Drills
1	-- See Baseball for children Training
2	-- See Baseball Training
3	-- See Basketball Training
4	-- See Drill and minor tactics --subdivision Drill and tactics under names of individual military services, e.g. United States. Army--Drill and tactics
5	-- See Drills (Planting machinery)
6	-- See Electric drills
7	-- See Emergency drills
8	-- See Fire drills
9	-- See Football Training

As Thomas mentioned about providing assistance to resolve ambiguity, "the descriptive phrase needs to be long enough to include the trigger words most likely to be recognized as relevant, but short enough to be displayed at the point of need." This approach fits those requirements with very little effort or cost.

As have Faiks and Thomas, I have drawn upon a Web services as a model. For a comparison, here's a partial image of Wikipedia's disambiguation page for the term "Drill":

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Toolbox
Print/export
Languages
Deutsch
Français
日本語

Drill (disambiguation)

From Wikipedia, the free encyclopedia

A **drill** is a tool or machine for cutting holes in a material.

Drill may also refer to:

Military

- The movements performed on a military parade
- Former name of the United States Army Reserve's Battle Assembly
- Exhibition drill, a form of military drill
- Drill commands
- When applied as an adjective, a practice version of something, e.g., drill round

In **music**

- Drill (music genre), a subgenre of gangsta rap
- Drill (band), an alternative rock band
- *Drill* (album), a 1996 album by the band of the same name
- The Drill (band), an electro house band
- Drill (UK band), an industrial rock band from England
- *Drill* (EP), a 1992 EP by Radiohead

There is a clear similarity in function and use.

IMPLEMENTATION

As to the implementation of such a practice, it has many features that make it readily possible and beneficial.

It is inexpensive: see references are easy to add into existing subject authority records and do not require subsequent alterations to bibliographic records.

> It should be noted that RDA mentions that "*The data should meet functional requirements for the* ***support of user tasks in a cost-efficient*** *manner. (0.4.2.2).*" Cost-efficient is not defined. It could refer to the cost of production of the data, or the cost of information resources going unused (Cf. Kent, 1979) or the cost of user's time being wasted and perhaps bearing no fruit through search fatigue and hidden results. Clearly,

the statement though does require the support of user tasks and this modest change would do so at both the find stage if not also that of identify.

It is scalable: such references will work regardless of the number of instances,

It is a familiar practice: initialisms and acronyms on name authority records are unqualified see-references to their authorized headings. Consider references from the acronym DOE, which has 10 see-references to corporate bodies from Fiji to Zanzibar. The initialism DDC has 15 see-references.

It is interoperable: Such see-references will function in all systems that make use of authority records; they might be handled and displayed differently, as we've seen with the two systems, Innovative Interfaces and ExLibris Primo/Alma, but they can function. Thus, it is system-agnostic.

CONCLUSION

I have proposed a Wikipedia-type disambiguation functionality in LCSH and that it carry through to RDA and chapter 23 when it is created. Such a modest change is cost effective, scalable, familiar, and system-agnostic. Furthermore, it is necessary for fulfillment of our catalog objectives.

Bibliography

Basic subject cataloging using LCSH. Instructor's Manual (Washington, D.C. : ALCTS/SAC-PCC/SCT Joint Initiative on Subject Training Materials, 2007)
Version 1h, October, 2007, Minor revisions, July 2009.

Bazzanella, Barbara, Heiko Stoermer, and Paolo Bouquet. "Entity Type Disambiguation In User Queries." Journal of Information & Knowledge Management 10.3 (2011): 209-224.

Beall, Jeffrey. "Search Fatigue." American Libraries 38.3 (2007): 46-50

Beall, Jeffrey, and Karen Kafadar. "Measuring The Extent Of The Synonym Problem In Full-Text Searching." Evidence Based Library & Information Practice 3.4 (2008): 18-33.

Buckland, Michael K. "Obsolescence in Subject Description." *Journal of Documentation* 68.2 (2012): 154-161.

Cota, Ricardo G. et al. "An Unsupervised Heuristic-Based Hierarchical Method For Name Disambiguation in Bibliographic Citations." *Journal of The American Society For Information Science & Technology* 61.9 (2010): 1853-1870

Cutter, Charles Ammi. *Rules for a Dictionary Catalogue*. 4th ed. (Washington : Government Printing Office, 1904), p.70-71

Elliott, Sarah. "Survey Of Author Name Disambiguation: 2004 To 2010." *Library Philosophy & Practice* (2010): 1-10

Faiks, Angi, Amy Radermacher and Amy Sheehan. (2007) "What about the book? Google-ising the catalog with tables of contents." *Library Philosophy and Practice,* (June) *LPP special issue on Libraries and Google*.

Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (Bethesda, Md: NISO Press, 2005)

Gregor, Dorothy and Carol Mandel. (1991) "Cataloging must change!" *Library Journal* Vol. 116, Issue 6 (April 1, 1991): p42-47.

Ide, Nancy and Jean Veronis. "Word sense disambiguation: the states of the art." *Computationsal linguistics (24:1)* p. 1-41.

Kent, Allen, et al. (1979) "Use of library materials : the University of Pittsburg study." New York : Marcel Dekker, 1979.

Resource description and access. (Chicago : American Library Association, 2010-)

Roberts, J.M. "Used car domain: an ethnographic application of clustering and multidimensional scaling." In *Classifying social data* (San Francisco : Jossey-Bass, 1982): p. 13-38.

Roman, Jorge H. et al. "Entity Disambiguation Using Semantic Networks." *Journal of The American Society For Information Science & Technology* 63.10 (2012): 2087-2099.

Studwell, William. "Why not an "AACR" for Subject Headings" (CCQ 6, no. 1: 3-9 (1985).

Studwell, William. "Ten years after the question: has there been an answer?" (CCQ 20, no. 3: 95-98 (1995).

Thomas, Bob. "Name Disambiguation-Learning From More User-Friendly Models." *Cataloging & Classification Quarterly* 49.3 (2011): 223-232.

Ward, Martin. "Phenomenological Warrant: the Case for Working From rhe User's Viewpoint." *Managing Information* 7.9 (2000): 68-71.

Wikipedia web site, September 12, 2013.

Yee, Martha M. and Sara Shatford Lane. Improving online public access catalogs. (Chicago : American Library Association, 1998)