Incorporating Discipline-Specific Classification into Faceted Browsing:

Research Proposal

Justine Withers

San Jose State University

Abstract

Why not use existing classification schemes to present faceted browsing in OPACs? The following study is proposed. We would measure the correspondence of “naive” sorting of topics between subject-specific classification and Dewey decimal classification (DDC). We would also compare retrieval results in two faceted browser interfaces: one based on DDC, the other on a subject-specific scheme. This study attempts to ascertain whether these subject-specific classifications aid searchers in a faceted browser. Do subject-specific classifications match user’s expectations for the organization of information? Do they provide better guidance to a specific topic than Dewey decimal classification and Library of Congress subject headings?

Incorporating Discipline-Specific Classification into Faceted Browsing

Faceted browsing can improve retrieval results by steering searchers in the right direction and removing ambiguity (Taylor, 2009). Its actual implementation in online public access catalogs and their complementary discovery layers has been clumsy, however (Withers, 2010). Part of the problem may lie in the background mapping between facets and bibliographic records. Another challenge might lie in the use of Library of Congress subject headings (LCSH) to display the knowledge of all disciplines. Harnessing existing taxonomies and classification schemes in a faceted browsing interface seems possible and helpful.

Purpose statement

In this study, we would measure the correspondence of “naïve,” participant-generated, sorting of topics between subject-specific taxonomies and Dewey decimal classification (DDC). We would also compare retrieval results in two faceted browser interfaces: one based on DDC, the other using subject-specific schema. This study attempts to ascertain whether these subject-specific classifications aid searchers in a faceted browser. Do subject-specific classifications match user’s expectations for the organization of information? Do they provide better guidance to a specific topic than Dewey decimal classification and Library of Congress subject headings? The actual mechanism for linking bibliographic records and multiple classification schemes is outside the scope of this paper. However, a simple cross-mapping technique is not outside the realm of possibility. Linked topics would allow catalogers to use existing LIS methods and make the resources available to others.

Definitions

classification: Spärck Jones' definition is useful here:

> Classification involves three distinct ideas: that we should divide the universe of objects; that we should do this in such a way that the subsets into which objects fall are held together by likeness among their members; and that the resulting description of the objects in terms of their class memberships should be simpler than their original description in terms of properties. (2005, p. 580)

In addition, we subscribe to Hjørland's caveat: "the selection of the properties of the objects must reflect the purpose of the classification" (2008, p. 334).

taxonomy: Here we rely on the idea of and functionality of a taxonomic tree structure, as described by Keshet (2011): "taxonomy generally organizes the knowledge of the world as a tree-like structure of broader–narrower, inclusive–included, superclass–subclass...relations between concepts" (p. 144). Our interface will reflect this structure.

facets: Jacob provides a useful description: "inductive, bottom-up schemes generated through a process of analysis and synthesis" (2004, p. 525). Jacob distinguishes faceted schemes as a controlled vocabulary rather than classification, while allowing their hierarchical structure, thus making it compatible with scientific taxonomies. In this study, we consider subject-specific classifications to be the primary facet upon which users will browse. The relevant characteristics here are that users can see all available subcategories (arrays) of a facet in logical order, to aid browsing (Mills, 2004).

subject-specific: Classification and terminology based on and relevant to a specific discipline, e.g. physics, biology, sociology, archaeology, and art (De Robbio, Maguolo, and Marini, 2001; Tudhope, 2003; Hjørland and Nicolaisen, 2004).

cross-walk: Tool for matching terminology in one organizational scheme with that in another (Taylor 2009, p. 116).

Tree of Life (ToL): Collaborative effort by biologists to create a modern taxonomy of the plant and animal worlds (Tree of Life Web Project, n.d.).

Physics and Astronomy Classification Scheme (PACS): A hierarchical classification system, similar to Dewey, in that there are ten main classes, each finely subdivided; maintained by the American Institute of Physics (2012).

discovery layer: User-oriented tools, usually harnessing Web 2.0 technologies and facets, that serve as a separate interface to an online catalog (Deng, 2010).

Significance

Without getting too deeply mired in an epistemological debate, we can safely say that this study comes from the pragmatic view of Hjørland and Pedersen (2005; Hjørland, 2008). Responding to Spärck Jones' assertion that "there is no one correct or natural way of classifying a universe of objects" (2005, p. 577), Hjørland and Pedersen (2005) argue that the LIS field should draw on classification schemes used in the topic at hand when cataloging. (To be clear, Hjørland (2012, p. 314) distinguishes himself as a "traditional" classifier rather than someone aligned with faceted or user-oriented approaches.) Mills (2004), Svenonius (2004), and

Thellefson (2004) offer more theoretical support for hierarchies over thesauri in reflecting domain knowledge and perspectives.

Hjørland and Pedersen note that "different communities use the same sign [i.e. words with different meanings assigned] without any trouble as long as their literatures are not merged into one database (2005, p. 590). We take this idea in a slightly different direction: namely, instead of re-cataloging resources, adding metadata to existing bibliographic records, or creating multiple interfaces, might we create crosswalks between topics so that they are correctly placed in the relevant taxonomy?

This study does not go as far as the Idea Collider (Smiraglia and van den Heuven, 2011), attempting to map a multiverse of knowledge, although we are not philosophically opposed to such a venture.

Previous explorations of cross-disciplinary classification have focused on using LCSH in a wider environment (Chan, 2000) or reported mixed feelings about LCSH in a qualitative study (Calhoun, 2006, p. 133). User studies have investigated interface metaphors (Dørum and Garland, 2011), the use of context in searching (Xie, 2000; Byström and Hansen, 2005; Kelly, 2006), perception of relevance (Savolainen and Kari, 2006), considering user's mental models (or preconceived ideas about a topic) (Novotny, 2004; Keshavarz, 2008), and keyword searches in federated databases (Williams, Bonnell, and Stoffel, 2009). Roszkowski (2011) identifies typical characteristics of subject-specific taxonomies. Additional studies have explored the feasibility of faceted browsing (Gabridge, et al., 2005; Sadeh, 2008; Emanuel, 2009) and discipline-specific headings in database searches (Deng, 2010). Of particular interest is a specialized faceted browser for retrieving information using the Art & Architecture Thesaurus® (AAT) (Tudhope, 2003). However, they do not touch on interdisciplinary searches, as would conceivably be conducted on a general use OPAC.

To the best of our knowledge, no studies have measured information retrieval from an integrated faceted browser using subject-specific classification.

Subject-specific taxonomy displays concepts using terminology users are familiar with. It is likely that discipline-specific headings will group topics more closely than DDC. We know it reflects the domain knowledge of experts. This study explores the idea that it triggers at least vague recognition among laypeople. Additionally, it is hoped that non-experts can learn as they browse and feel increased confidence when relying on expert knowledge. We hope to harness language familiar to users without resorting to folksonomy (Keshet, 2011).

This study presumes the usefulness of faceted browsing and information visibility, as proposed in Sadeh (2008) and Mansourian, et al. (2008). Faceted browsing eases information retrieval by providing context to terms, guiding a searcher to related topics, and giving preliminary hints for search queries. One way of looking at it is that the OPAC offers information and a faceted browser offers knowledge, as differentiated in Zins (2006). Faceted browsing combined with subject-specific classification might solve Hjørland and Pedersen's concern that "users are normally not able to specify criteria in a literature of which they are not very knowledgeable" (2005, p. 591). Users avoid entering a specific search query that might be wildly inaccurate.

Other researchers have explored cross-mapping of subject-specific classification and DDC headings and this study gratefully takes advantage of their work (De Robbio, Maguolo, and Marini, 2001). Out of the scope of this study but relevant to future research are the complications of cross-mapping multiple classifications, changes in subject knowledge (see Hjørland and Pedersen, 2005, for an example of DDC adjustments to reflect changes to the treatment of "race" in sociology), the question of who would manage cross-mappings and metadata implementation, and the ideal design of the faceted browser.

Methodology

This study will collect both quantitative and qualitative data.

Design

This study will compare the taxonomy of the Dewey Decimal Classification (DDC) with those of the Physics and Astronomy Classification Scheme (PACS) and the Tree of Life (ToL). It will consist of two phases: 1) a card-sorting exercise to measure user expectations with existing classifications and 2) a think-aloud simulation of information seeking. We use scientific classifications to leverage existing, established taxonomies.

We start with two null hypotheses:

Hypothesis TM: User-defined taxonomies will show insignificant commonality among themselves or between themselves and DDC, PACS, or ToL.

Hypothesis PS: Retrieval time and perceived success will be the same in a faceted browser based on DDC and one based on PACS and ToL.

Phase one

In phase one, participants will sort cards with various scientific topics in lay terms (cf. Pisanski and Zumer, 2010, for methodology). They will create a tree structure that best represents how they organize the topics at hand. Phase one will provide background information on how users organize concepts without the influence of a provided taxonomy. Understanding their mental model (cf. Ahmed, 2009) will help us understand how they interact with a given classification scheme.

Phase two

In phase two, we will create two faceted browser interfaces: one based on DDC and LCSH (as presented in WebDewey) and one based on PACS and ToL. We will use the cross-mapping scheme proposed by De Robbio, Maguolo, and Marini (2001) for PACS and create a similar crosswalk for ToL. Both assume a user would begin with the concept of "science."

We will give each participant three tasks that require retrieving scientific information (Kelly, 2006; Kuhlthau, 1991). As they conduct their search, we will record their thoughts using think-aloud protocol analysis (Novotny, 2004). We will time their search and interview them to gauge their perceived level of success.

Ellis (1996) shows that testing on hypothetical indexing devices such as our mock browser can yield fruitful results.

Sample Pool

We have prioritized tiers of participants, depending on time and funds available.

Comparing a group of career scientists (n=30) with a group of community college students (n=15 communications majors, 15 health science majors; age range=30-40) will produce the most direct comparison. Scientists will presumably use their subject domain knowledge to guide their search. Community college students would presumably be exposed to scientific subjects at a layperson's level.

If resources permit, conducting the same protocol on a pool of tenth-graders (n=30) and graduate students (n=30) would allow us to compare domain knowledge and presumed searching techniques. Tenth-graders have taken at least basic science classes, are old enough to self-report their behavior, and are used to browsing web pages. Graduate students would have more specific knowledge of a topic and more experience with search queries to find content.

Analysis

For the card-sorting exercise, we will measure the similarity between the user's scheme and that of DDC, PACS, and ToL. For each topic, we will count the number of nodes away from the root of each structure and also how that number differs among the three schemes. This will create a measure of accuracy for the user's perception. For each user scheme, we will measure its breadth and depth and compare these to DDC, PACS, and ToL. This will create a measure of topical breadth.

For the retrieval exercise, we will collate participants' comments, especially those regarding perceived ease of use (using Al-Maskari and Sanderson's criteria, 2010), mental models (Ahmed, 2009), and use of domain knowledge.

Budget and Schedule

This study attempts to harness existing materials and maximize data-gathering.

Budget

We anticipate the following expenses:

Expense	Projected cost		Rationale
Recruiting	$	100	paper for informational flyers
Space	$	0	use existing lab space
Supplies	$	100	3x5 index cards tape printer paper and toner digital camera (already owned)
Remuneration	$	1500	Sixty (60) $25 gift cards (assuming initial group of 60; double if second group is deemed feasible)
Data gathering and analysis	$	3960	3 x 88 hrs x $15/hour 3 researchers Four (4) research sessions of two (2) hours each (additional $360 if second group in study) Forty (40) hours collating data Forty (40) hours analyzing data
Writing	$	3600	3 x 80 hrs x $15/hour 3 researchers Two (2) weeks of writing time
Total	$	9260	(plus $1560 for second group)

We envision the following schedule:

Activity	Duration	Rationale
Recruiting	2 (two) months	Identify potential recruits Gather participants and receive permission Give three (3) weeks notice for scheduling Concurrently, conduct practice sessions

Research sessions	3 (three) weeks	Conduct 2 (two) sessions a week One week padding to accommodate scheduling conflicts
Data collation	2 (two) weeks	
Data analysis	3 (three) weeks	
Writing	3 (three) weeks	

Future research

If this study is successful, we see applications to several other areas of research. Future studies might focus on expanding cross-mapping to other subjects, especially the “soft” sciences. User studies could identify best practices for offering access to multi-disciplinary faceted browser interfaces. A query log study (similar to Strohmaier and Kroll, 2012) could track how users narrow and broaden their searches. With a strong cross-mapping schematic, a user study could measure precision and recall for an entire catalog. If cross-mapping proves fruitful, users might be able to jump from one taxonomy to another, facilitating the interdisciplinary research Szostak advocates (2008), without necessitating universal classification. Again the debate between the two is beyond the scope of this study. On a related topic, knowing how much established taxonomies assist users would contribute to the examination of taxonomy-folksonomy hybrids (Keshet, 2011). A further usability study could incorporate more theoretical schema, such as the Integrative Web Classification (International Society for Knowledge Organization, 2010) or Bliss Classification (Mills, 2004).

References

Ahmed, S.M.Z., McKnight, C., and Oppenheim, C. (2009). A review of research on human-computer interfaces for online information retrieval systems. *The Electronic Library, 27*(1), 96–116.

Al-Maskari, A. and Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology, 61*(5), 859–868.

American Institute of Physics. (2012). “What is PACS?” Retrieved from http://www.aip.org/pacs/pacs2010/about.html

Byström, K. and Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology, 56*(10), 1050–1051.

Calhoun, K. (2006). The changing nature of the catalog and its integration with other discovery tools Retrieved from http://www.loc.gov/catdir/calhoun-report-final.pdf

Chan, L.M. (2000) Exploiting LCSH, LCC, and DDC to retrieve networked resources: issues and challenges. Retrieved from http://www.loc.gov/catdir/bibcontrol/chan_paper.html

De Robbio, A., Maguolo, D., and Marini, A. (November, 2001). Scientific and general subject classifications in the digital world. *High Energy Physics Libraries Webzine, 5.* Retrieved from http://library.web.cern.ch/library/Webzine/5/papers/4

Deng, S. (2010). Beyond the OPAC: creating different interfaces for specialized collections in an ILS system. *OCLC Systems & Services: International Digital Library Perspectives, 26*(4), 253–262.

Dørum, K. and Garland, K. (2011). Efficient electronic navigation: a metaphorical question? *Interacting with Computers, 23,* 129–136.

Ellis, D. (1996). The filemma of measurement in information retrieval research. *Journal of the American Society for Information Science, 47*(1), 23–36.

Emanuel, J. (2009). Next generation catalogs: What do they do and why should we care? *Reference & User Services Quarterly, 49*(2), 117–120.

Gabridge, T.A., Hennig, N., Lubas, R., and Wenzel, S.G. (2005). When a librarian's not there to ask: creating an information resource advisory tool. Proceedings of the ACRL 12th National Conference, April 7–10, Minneapolis, MN.

Hjørland, B. and Pedersen, K.N. (2005). A substantive theory of classification for information retrieval. *Journal of Documentation, 61*(5), 582–597.

Hjørland, B. (2008). Core classification theory: a reply to Szostak. *Journal of Documentation, 64*(3), 333–342.

Hjørland, B. (2012). Is classification necessary after Google? *Journal of Documentation, 68*(3), 299–317.

International Society for Knowledge Organization. (November 2010). "How ILC works" Retrieved from http://iskoi.org/ilc/ilc/how.htm

Jacob, E. (2004). Classification and categorization: a difference that makes a difference. *Library Trends, 52*(3), 515–540.

Kelly, D. (2006). Measuring online information seeking context, Part 2: findings and discussion. *Journal for the American Society for Information Science and Technology, 57*(14), 1862–1874.

Keshavarz, H. (2008). Human information behaviour and design, development and evaluation of information retrieval systems. *Program: electronic library and information systems, 42*(4), 391–401.

Keshet, Y. (2011). Classification systems in the light of sociology of knowledge. *Journal of Documentation, 67*(1), 144–158.

Kuhlthau, C. (1991). Inside the search process: information seeking from the user's perspective. *Journal for the American Society for Information Science and Technology, 42*(5), 361–371.

Mansourian, Y., Ford, N., Webber, S., and Madden, A. (2008). An integrative model of "information visibility" and "information seeking" on the web. *Program: Electronic Library and Information Systems, 42*(4), 402–417.

Mills, J. (2004). Faceted classification and logical division in information retrieval. *Library Trends, 52*(3), 541–570.

Novotny, E. (2004). I don't think, I click: A protocol analysis study of use of a library online catalog in the Internet age. *College & Research Libraries, 65*(6), 525–537.

Pisanski, J. and Zumer, M. (2010) Mental models of the bibliographic universe. Part 1: Mental models of description. *Journal of Documentation, 66*(5), 643–667. doi: 10.1108/00220411011066772

Pisanski, J. and Zumer, M. (2010) Mental models of the bibliographic universe. Part 2: Comparison task and conclusions. *Journal of Documentation, 66*(5), 668–680. doi: 10.1108/00220411011066781

Roszkowski, M. (2011). Using taxonomies for knowledge exploration in subject gateways. Presented at INFORUM 2011: 17th Conference on Professional Information Resources, Prague, May 24–26.

Sadeh, T. (2008). User experience in the library: A case study. *New Library World, 109*(1/2), 7–24.

Savolainen, R. and Kari, J. (2006). User-defined relevance criteria in web searching. *Journal of Documentation, 62*(6), 685–707.

Smiraglia, R.P. and van den Heuvel, C. (2011). Idea collider: from a theory of knowledge organization to a theory of knowledge interaction. *Bulletin of the American Society for Information Science and Technology, 37*(4), 43–47.

Spärck Jones, K. (2005) Some thoughts on classification for retrieval. *Journal of Documentation, 61*(5), 571–581.

Strohmaier, M. and Kroll, M. (2012). Acquiring knowledge about human goals from search query logs. *Information Processing and Management, 48,* 63–82.

Svenonius, E. (2004). The epistemological foundations of knowledge representations. *Library Trends, 52*(3), 571–587.

Szostak, R. (2008). Classification, interdisciplinarity, and the study of science. *Journal of Documentation, 64*(3), 319–332.

Taylor, A.G. and Joudrey, D.N. (2009). The Organization of Information. Westport, Conn.: Libraries Unlimited.

Thellefson, T. (2004). Knowledge profiling: the basis for knowledge organization. *Library Trends, 52*(3), 507–714.

Tree of Life Web Project. (n.d.) “Tree of Life Web Project. ” Retrieved from http://tolweb.org/tree/

Tudhope, D. (2003). “The FACET project.” Retrieved from https://www.comp.glam.ac.uk/~FACET

Williams, S.C., Bonnell, A., and Stoffell, B. (2009). Student feedback on federated search use, satisfaction, and web presence. *Reference & User Services Quarterly, 49*(2), 131–139.

Withers, J. (2010) Uncovering the AquaBrowser Discovery Layer at the Saint Paul Public Library. [unpublished]

Xie, H. (2000). Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. *Journal of the American Society for Information Science, 51*(9), 841–857.

Zins, C. (2006). Redefining information science: from “information science” to “knowledge science.” *Journal of Documentation, 62*(4), 447–461.