Capitalizing on Information Organization and Information Visualization for a New-Generation Catalogue

CHARLES-ANTOINE JULIEN, CATHERINE GUASTAVINO, AND FRANCE BOUTHILLIER

ABSTRACT
Subject searching is difficult with traditional text-based online public access library catalogues (OPACs), and the next-generation discovery layers are keyword searching and result filtering tools that offer little support for subject browsing. Next-generation OPACs ignore the rich network of relations offered by controlled subject vocabulary, which can facilitate subject browsing. A new generation of OPACs could leverage existing information-organization investments and offer online searchers a novel browsing and searching environment. This is a case study of the design and development of a virtual reality subject browsing and information retrieval tool. The functional prototype shows that the Library of Congress subject headings (LCSH) can be shaped into a useful and usable tree structure serving as a visual metaphor that contains a real world collection from the domain of science and engineering. Formative tests show that users can effectively browse the LCSH tree and carve it up based on their keyword search queries. This study uses a complex information-organization structure as a defining characteristic of an OPAC that goes beyond the standard keyword search model, toward the cutting edge of online search tools.

INTRODUCTION
The use of library catalogues is declining as traditional online catalogues no longer live up to the expectations of users (Calhoun, 2006; Sadeh, 2007). Libraries have to improve their online search tools to compete with leading innovations from the likes of Amazon and Google. Systematic organization of information is a defining characteristic of library collections that could attract users back to the library catalogue. Unfortunately, current online tools essentially hide the underlying information-organization
structure, making it difficult for users to browse the collection using the topics covered. The field of information visualization (IV) (Bederson & Shneiderman, 2003) offers potential solutions in the form of interactive visual representations of topic structures acting as metaphoric containers for the contents of the collection. This article presents a case study of the design and development of a prototype library catalogue where searchers can browse and search for information using an interactive three-dimensional (3D) visual representation of the Library of Congress subject headings (LCSH) hierarchy.

Since 2006, a new generation of discovery tools has been replacing the traditional online public access catalogue (OPAC) (see reviews by Mi & Weng, 2008; Yang & Wagner, 2010). Three innovative catalogues represent this latest generation of OPACs referred to as discovery layer: Endeca’s faceted searching (Endeca, n.d.), AquaBrowser (SerialsSolutions, n.d.), and WorldCat Local (Online Computer Library Center [OCLC], n.d.). Commercial integrated library systems vendors matched this latest generation of information search tools with their own discovery layers (e.g., ExLibris, n.d.; Innovative Interfaces, n.d.; SirsiDynix, n.d.). The dynamic field of open source also offers demonstrated solutions; for example, Evergreen (http://open-ils.org) and Koha (LibLime, n.d.). Once at the cutting edge of online search tools (Markey, 2007), is the OPAC condemned to playing a never-ending game of catch up with highly dynamic and well-funded commercial Web sites? Or should libraries attempt to better communicate the distinctive features of their collections through innovative OPACs?

**Subject Browsing**

Searching an OPAC via its controlled subject vocabulary (CV) can remedy problems associated with a natural language (NL) keyword searching strategy. CV searching increases the precision of results; this saves the time of the user by reducing the long lists of mostly irrelevant results often returned by overly general NL (Larson, McDonough, O’Leary, & Kuntz, 1996). An NL search also “fails to group related materials together” and, as a result, “much valuable material [to the user] may be missed” (Bates, 2003, p. 14). The value of CV is not lost on popular commercial Web sites; for example, Amazon organizes its product into a CV hierarchy of broad to narrow categories available in a TouchGraph visualization (TouchGraph, n.d.).

The design of traditional subject browsing OPAC interfaces is inadequate for subject browsing (Borgman, 1996; Papadakis, Stefanidakis, & Tzali, 2008, p. 19). There are experimental attempts to improve subject browsing by leveraging existing classification numbers such as Dewey Decimal (Allen, 1995) or Library of Congress Classification (LCC) (Chandler & LeBlanc, 2006), and the North Carolina State University Libraries (http://www.lib.ncsu.edu/catalogue/) allow online searchers to filter
their search using the first character of the LCC. A drawback of this approach is that a single classification number cannot truly describe all the topics covered by multidisciplinary works.

**Next-Generation OPACs**

Since 2006, this latest generation of online library catalogues allow users to filter their NL query results using a subject facet (Anderson & Hoffman, 2006). These subject facets are automatically extracted from existing CV, and usage statistics (Prestamo, 2007) suggest that facets are increasingly used by searchers. Subject facets are used to reduce the results of an existing query; they do not provide an overview of the collection and its predominant subjects. Subject faceting also ignores the network of relations between CV strings. This hinders topical browsing of the bibliographic collection.

**LCSH**

The LCSH is a CV maintained by the Library of Congress (LC) since 1898 for the creation of access points and their assignment to bibliographic records. Despite “perennial criticisms” (Taylor & Miller, 2006, p. 350) centered around cost, scalability, and consistency, LCSH is the standard list of topical CV for most large libraries in the United States and throughout the world (Chan & Hodges, 2007, p. 213). LCSH is a pertinent case study since it covers all knowledge domains; projects such as SKOS (http://www.w3.org/2004/02/skos/) suggest that the study of the LCSH structure may also benefit the semantic web.

A bibliographic collection organized with LCSH is comprised of two data sources: (1) bibliographic records that are indexed by topic with at least one topical LCSH string and (2) the LCSH authority records that contain the established topical LCSH strings and define the topical LCSH hierarchy. A topical LCSH string is composed of a topical main heading and possible subdivisions meant to further specify the main heading. The following definitions are used in this article:

- **LCSH string:** any topical LCSH.
- **Established (LCSH) string:** an LCSH string listed in the LC authority records. LCSH strings not listed in the LC authority records (i.e., nonestablished strings) are created by cataloguers without being added to the authority records.
- **Main heading:** the leftmost portion of an LCSH string indicating the main topic of the string, excluding any subdivisions. Note that every LCSH string has a main heading, and a string without subdivision(s) is equal to its main heading.
- **(LC) authority records:** the list of established LCSH strings. This includes all valid main headings and some subdivided main headings. Authority
records define the relationships between main headings (i.e., broader/ narrower or related terms).

- **LCSH (syndetic) structure**: the sum of relations between main headings as defined in the authority records.

Main headings can be related with other main headings in one of two ways: **broader** term (and the reciprocal narrower term) or a **related** term (Chan & Hodges, 2007). Broader terms are vertical relationships that define a hierarchy of broad (e.g., science, medicine) to narrow terms (e.g., mechanics, surgery). Related terms represent horizontal relations, which are not considered by the current study.

For example, let us assume “Aircraft accidents–Human factors–Quebec (Canada)” is an LCSH string assigned to a bibliographic record. “Aircraft accidents–Human factors” is the established LCSH string listed in the LC authority records. “Aircraft accidents” is the main heading whose authority record states that it has a broader term, “Transportation accidents”; a related term, “Airplane crash survival”; and several narrower terms, including “Runway incursions” and “Near misses (Aeronautics).”

**Information Visualization**

Awareness and usage of subject indexing by the general public may increase if access is facilitated by information-interaction techniques from the field of IV (Calhoun, 2006). IV techniques (Bederson & Shneiderman, 2003) can complement textual search interfaces with direct manipulation visual metaphors of information spaces. IV seems to be well suited for broad or introductory searches (Marchionini, 1995), and the primary aim of IV is to visually reveal patterns or identify characteristic features in large data sets (Bederson & Shneiderman, 2003, p. ix). A collection visualization might offer “many starting paths for exploration . . . while immediately familiarizing the user with the high-level information structure of the collection” (Yee, Swearingen, Li, & Hearst, 2003, p. 403).

Since the 1990s, visually enhanced information retrieval methods have been developed to support query formulation and result inspection (see review by Hearst, 1999). Many use spatial representations of information collection as an interaction metaphor capitalizing on ubiquitous preexisting knowledge derived from real world physical navigation and object manipulation (see review by Borner, Chen, & Boyack, 2003). One of the major appeals of 3D virtual reality IV is to capitalize “on the ‘natural’ navigation skills of users” (Plaisant, Grosjean, & Bederson, 2003, p. 362).

Three-dimensional cone trees (Robertson, Mackinlay, & Card, 1991) is one of few visualization techniques that have been systematically integrated into designs of information systems (Chen, 2004; Hearst & Karadi, 1997). A cone tree is built from the top of the hierarchy placed near the top of the screen “and is the apex of a cone with its children placed evenly
spaced along its base. The next layer of nodes is drawn below the first, with their children in cones” (Robertson et al., 1991, p. 190). In this manner a complete overview of the hierarchy is provided. This makes the technique well suited for overall structure comprehension but requires efficient navigation controls for specific node finding. Our study is unique in that it demonstrates a 3D cone trees–inspired visual interface to LCSH.

**Current LCSH Browsing Systems**

Current LCSH browsing systems (Library of Congress Authorities, http://authorities.loc.gov/; OCLC Connexion, http://connexion.oclc.org/) do not provide explicit features to navigate the LCSH structure, forcing users to visit each LCSH individually. Traditional alphabetical displays of LCSH strings have long been recognized as inadequate (Drabenstott & Weller, 1996) since they hide the structure of interrelated concepts (Larson, 1992, p. 130; Richmond, 1959). Inadequate LCSH browsing features may partly explain why its “hierarchical nature is largely ignored” (Frank & Paynter, 2004, p. 214) by the searcher. According to an LC report (Library of Congress, 2008), LC should support initiatives into building a true LCSH hierarchy, and efforts should be made to increase LCSH usage in modern library discovery layers. The few LCSH browsing interfaces (Papadakis, Stefanidakis, & Tzali, 2009; Yi & Chan, 2008) are designed as interfaces to the empty LCSH structure from LC, which does not contain a bibliographic collection. Our study differs from prior endeavors since it builds the LCSH hierarchy containing a specific collection so that searchers are able to browse a collection by exploring its specific portion of the LCSH hierarchy. The following sections describe the design and development of a functional prototype that

- reduces the LCSH structure to reflect the contents of a real world collection;
- automatically shapes this structure into a valid tree;
- simplifies the constructed LCSH tree to facilitate visual exploration;
- integrates visual browsing of the LCSH tree with keyword searching of the collection;
- has been iteratively tested and refined with four test users.

**Development**

**Data**

Yi and Chan (2010) analyze the LCSH structure and report a skewed distribution of relationships; in other words, the LCSH structure offers a few large groups of highly interconnected main headings, but most main headings are connected to few. The most connected groups are associated with *Science*, consequently, this domain is chosen for this study since it offers the most complex structure from which to build an LCSH tree.
The specific data sets are composed of 204,430 topical authority records provided by the McGill University Libraries in January 2008 and 130,940 bibliographic records housed in the Schulich Science and Engineering Library.

**Constructing the LCSH Tree**

Recognizable hierarchies, such as computer file explorers, are called trees because they follow a set of specific structural rules. A tree consists of a set of nodes, including a root node that is a common ancestor to all nodes (Di Battista, 1999). Nodes can have multiple child nodes, and those without children are called leaf nodes. Each node has a single parent node except the root node, which has no parent. A tree must be *acyclic*, meaning that a node cannot be both an ancestor and a descendent of another. The LCSH structure cannot be called a tree since it has known structural issues:

- It is a set of independent branches without a common root node (Yi & Chan, 2009) since there is a large number of orphan established strings that have no broader term (see review by Shubert, 1992, pp. 59–60).
- Main headings can have multiple parents.
- It may contain cycles.

These issues have been addressed in order to transform the collection-specific LCSH structure into a valid tree (see details in Julien, Tirilly, Leide, & Guastavino, 2012a). Figure 1 illustrates the data manipulation process used by this study to produce an LCSH tree that covers a real world collection. It shows that the LCSH structure is first reduced to contain only the established LCSH strings assigned to the specific collection being browsed (left portion of fig. 1). This requires matching each LCSH string found in the bibliographic records with an established string; this is hampered by the systematic cataloguing practice of adding subdivisions. The remaining LCSH structure contains the collection (middle portion of fig. 1); this structure is not a tree since some established strings have no relations with others. The LCSH tree is produced by modifying the LCSH structure (right portion of fig. 1) to adhere to the rules defining tree structures (i.e., a single root node, no cycles, a single parent per node). The resulting structure is a valid tree of established LCSH strings that contains the specific collection.

The resulting LCSH tree is too large to be visually represented; however, it can be simplified by pruning most established strings that are assigned to very few bibliographic records (see details in Julien, Tirilly, Leide, & Guastavino, 2012b). Sacrificing some information in order to gain simplicity is sometimes justified (Small, 2000) since “what is sought in designs for the display of information is the clear portrayal of complexity” (Tufte, 1983). The result is an LCSH tree that clearly reveals the few
established LCSH strings assigned to the vast majority of the bibliographic records.

**Virtual Reality Subject Browsing and Information Retrieval**
The developed application is a 3D visual LCSH tree acting as a searchable information container for the specific collection. Through animated viewpoint movements (also called egocentric view), searchers use keyboard arrows and point-and-click to explore the broad-to-narrow relations between established LCSH strings. Three levels of detail are available:

- An LCSH tree overview (see fig. 2) shows only the established LCSH strings that contain the majority of the bibliographic records. Branches can be dynamically expanded or retracted by clicking on a floating label.
- When the user selects a specific established string, an animated transition transports the viewpoint toward a two-dimensional plane called a subject map (see fig. 3) where the string’s narrower terms are represented as circles sized in proportion with the number of bibliographic records they contain. Established strings leading to additional narrower terms are identified by “More Below.”
- A textual list of bibliographic record details is available on demand (see fig. 2, top left).

Clicking on a floating label prunes the branches that are not direct descendants of the chosen established LCSH string and reveals additional tree depth if applicable. This also updates the textual list to show the bibliographic records assigned to the established string. Clicking on an LCSH string in the textual list (fig. 2, top left) triggers an animated movement toward the position of its established string in the tree. Floating labels are consistently drawn at the same position in space, allowing searchers to progressively build a mental model of the collection. Previously visited
established strings are indicated with a dark green border.

Keyword searching dynamically filters and prunes the tree to reveal the portions that contain matching bibliographic records (see fig. 4), and the textual list (fig. 2, top left) is filled with a classic list of ranked results. The label colors are changed to reflect the number of matching bibliographic records contained in the established string (see fig. 4): red labels contain the highest number of matching bibliographic records, and orange labels contain some matching bibliographic records. Search results often cluster around specific areas of the tree; this can facilitate navigation by creating conspicuous and stable information landmarks.

**Formative Testing**
The prototype was iteratively tested with four different participants at various stages of the development process using a think-aloud protocol. The first participant tested basic navigation and the following three test users performed browsing and information-retrieval tasks. To maximize the benefit of each formative evaluation, fixes and modifications were integrated into the prototype before being tested by the next participant. Some of the main redesigns resulting from these formative tests are briefly described next.

**Single Search Index for Structure and Collection**
Early tests clearly showed users had difficulty understanding the conceptual difference between their keywords matching an established LCSH
Figure 3. Close up view of a subject map showing the immediate narrower terms belonging to an established string.

Figure 4. Integration of keyword searching which prunes the visible portions of the LCSH subject hierarchy.
string synonym versus matching bibliographic metadata elements (i.e., title, series). This relates to the classic OPAC distinction between searching authority records versus searching bibliographic records; this distinction was lost on users. To address this, the indexing unit includes semantically salient bibliographic fields (i.e., title, series, LCSH strings); additionally, for each LCSH string, established string synonyms (i.e., see from relations) are also added to the indexing unit. The result is an indexing unit that includes fields from the bibliographic record (i.e., title, series, LCSH strings) and the established string synonyms from the authority records.

The combination of bibliographic and authority data into a single indexing unit simplifies the integration and usage of keyword searching; however, a search cannot return strictly an established LCSH string even if it exactly matches the user query. For example, a keyword search for the specific main heading “Computer software” does not directly return this heading in the result list; instead, it returns matching bibliographic records that are likely to have been assigned to the main heading.

**Easy Navigation in Three Dimensions**

The first version of the system required users to navigate the 3D environment using an airplane metaphor. An airplane reaches a specific target in space if the pilot chooses the correct approach trajectory; this proved much too difficult for untrained users. Tests showed that users intuitively understood that keyboard arrow keys would allow forward/back and left/right movements on a two-dimensional *floor beneath ones’ feet*. Two additional arbitrary keyboard keys (i.e., T for Top and B for Bottom) provide the ability to move up and down along a vertical axis pointing *above ones’ head* and *below one’s feet*; this allows users to navigate to any other *floor* in space. The result is a quasi-helicopter metaphor where speed is automatically controlled by the system based on the proximity of the viewpoint with the LCSH tree (i.e., fast when far away, slow when close to an individual subject map).

**Animated Viewpoint Rotation**

Early formative tests revealed that users often repeated sequences of movement and gaze-direction shifts in order to inspect objects from different viewpoints. This was tedious and difficult. A participant suggested there should be a way to automatically rotate the viewpoint around the tree. Rotation requires an axis and its location would not be explicitly specified by the user. Rotation around the center of mass of the currently visible portion of the tree is impractical since unbalanced tree parts have an axis of rotation positioned far from expectations. We determined that the best option was to rotate around the last established LCSH string label selected by the user, and its position is usually close to the tree center of mass.
Wall-Mounted Landmarks
Conspicuous textual labels acting as wall-mounted landmarks were added to provide relative directional cues. This was suggested by a participant as he was rotating around the structure and wondered if he had traveled back to his starting viewpoint. Like cardinal directions, wall-mounted landmarks provide a fixed referential system that anchors the tree within a larger static world.

Mouse Wheel Zooming
We observed that the zooming features mapped to the keyboard were used sparingly. A participant suggested replicating the mouse wheel zooming features of online maps offered by search engines. This was implemented as a way to move forward and back along the current gaze direction.

These formative test sessions resulted in a significant evolution of the prototype. The first test user could barely use the system without uncovering navigation issues and bugs. After successive modifications and testing, the fourth test user confirmed the prototype was stable and usable. This last participant was able to effectively perform information retrieval tasks ranging from hierarchy exploration (e.g., find a specific established LCSH string) to complex retrieval where the vocabulary of the question is not sufficient for keyword searching (e.g., find books that might help the reader fix a broken computer).

Conclusion
This article described the design, development, and formative evaluation of a virtual reality browsing and searching tool for a real world collection organized with the LCSH. The design and development process revealed that LCSH broad-to-narrow terms can be transformed into a usable visual tree; however, the LCSH tree containing the information collection was significantly simplified to facilitate its visual presentation and exploration. The simplified LCSH tree clearly shows the few LCSH branches that contain the vast majority of the collection; in other words, relatively empty LCSHs are pruned for the sake of simplified visual navigation. Searchers can progressively browse the hierarchy of broad-to-narrow LCSHs and carve the tree using keyword searching. In addition, a traditional textual list of ranked results allows for a quick travel to individual established LCSH strings. Four rounds of successive formative tests have refined the prototype into a usable proof-of-concept.

This study capitalizes on existing controlled subject vocabulary to deliver a functional prototype OPAC that integrates visual browsing of the information-organization structure and keyword searching. Investigating new ways of interacting with topically organized library collections is a step towards a new generation of online search tools that capitalize on existing information-organization investments. Future directions include formal
usability testing with a larger number of participants on a wide range of information-retrieval tasks and comparison with text-only information-retrieval systems in terms of performance and affective reactions.

ACKNOWLEDGMENTS
The authors wish to thank McGill Libraries for providing the experimental data used in this work. Part of this work was funded by the Fonds Québécois de Recherche sur la Société et la Culture.

REFERENCES


Charles-Antoine Julien is an assistant professor at the McGill University School of Information Studies. He earned a doctorate in information studies specializing in information visualization of information organization structures. He is an industrial engineer with professional experience as a business information systems programmer, analyst, and trainer. His thesis demonstrated novel techniques allowing searchers to dynamically simplify subject structures. He has published on information visualization and simplification of information-organization structures, next-generation library catalogues, and human–computer interaction. Dr. Julien was also a postdoctoral fellow at the School of Information Studies of the University of Wisconsin in Milwaukee.

Catherine Guastavino is an associate professor at McGill University in the School of Information Studies and a member of CIRMMT, the Centre for Interdisciplinary Research in Music Media and Technology. She has published more than eighty articles on the fields of human–computer interaction, music information perception, archiving and retrieval, and multisensory information processing. She directs the Multimodal Interaction Laboratory and teaches in the areas of multimedia systems, database design, and research principles.

France Bouthillier is director and an associate professor at the School of Information Studies, McGill University. She has published on the evaluation of competitive intelligence software applications, competitive intelligence, knowledge management, business information services, and the management of library/information services. She has conducted research on public library services, the information needs of small businesses, the design of information services for the business community, and the use of information technology by competitive and strategic intelligence practitioners. As a consultant, she has advised corporations and government agencies on the design of intelligence software and on the structure and impact of intelligence services to improve their decision-making process.