A Visualization Software Tool for Library of Congress Subject Headings

Abstract
The aim of this study is to develop a software tool, VisuaLCSH, for effective searching, browsing, and maintenance of LCSH. This tool enables visualizing subject headings and hierarchical structures implied and embedded in LCSH. A conceptual framework for converting the hierarchical structure of headings in LCSH to an explicit tree structure is proposed, described, and implemented. The highlights of VisuaLCSH are summarized below: 1) revealing multiple aspects of a heading; 2) normalizing the hierarchical relationships in LCSH; 3) showing multi-level hierarchies in LCSH sub-trees; 4) improving the navigational function of LCSH in retrieval; and 5) enabling the implementation of generic search, i.e., the ‘exploding’ feature, in searching LCSH.

1. Introduction
Library of Congress Subject Headings (LCSH) serves as a de facto universal controlled vocabulary for libraries in many countries (O’Neill & Chan, 2003). Beyond its use in traditional libraries, LCSH also serves as a tool for interoperability among metadata schemes (Vizine-Goetz, et al., 2004), for creating cataloguing records of digital information (Frank & Paynter, 2004; Riley, 2007), and for integrating multiple information depositories (Koch & Day, 1997). Despite these expanded applications of LCSH, most users underutilize LCSH’s full power in part due to its complexity and the lack of an effective tool to guide users.

The conversion of hierarchical term references in the entire LCSH (called ‘hidden classification’ of LCSH) into an explicit tree structure has long been desired (Richmond, 1959). To date, there is no existing tool for displaying the complete ‘hidden classification’ of LCSH. Two popular systems for searching LCSH and LC authority records (source of LCSH) are currently available: a) the LC Authorities website (http://authorities.loc.gov) which is freely open to the public; and b) the OCLC Connexion service (http://www.oclc.org/connexion) which is accessible only to registered users. Both systems are primarily designed for accessing a single authority record at a time. To view the hierarchical structure surrounding a heading, users are required to take additional steps to visit each heading individually. In the recently released report on the future of bibliographic control, recommendations for the optimal use of LCSH include: 1) “transform LCSH into a tool that provides a more flexible means to create and modify subject authority data”; 2) “[e]nable its customized assembly into topical thesauri”; and 3) “evaluate the ability of LCSH to support faceted browsing and discovery” (Library of Congress, 2008, 35).

The following functions, valuable not only to general users but also to professionals responsible for development and maintenance of LCSH including the display of ‘hidden classification’, are only partially supported by current LCSH tools: 1) displaying the full hierarchical structure of LCSH; 2) showing multiple locations of a LC subject heading in LCSH hierarchies; and 3) displaying LCSH terms all of which are synonym/homograph.

The aim of this study is to develop a software tool for effective searching, browsing, and maintenance of LCSH by enabling the visualization of subject headings and hierarchical structures implied and embedded in LCSH. This new LCSH visualization tool (called VisuaLCSH) is proposed to fill the aforementioned gaps in current systems. It is targeted
for various user groups, from novices to professionals, and is applicable to a wide range of applications including the searching, browsing, updating and maintaining of LCSH, as well as allowing for the integration of LCSH to other knowledge organizations.

A primary function of the VisuaLCSH is to enable the visualization of the full hierarchical structure of LCSH in both macro and micro views, thereby enabling effective browsing and traversing of LCSH. VisuaLCSH also offers smart searching of LCSH by considering both the synonyms and the word-based or character-based matches of search terms. The major features supported by the LCSH visualization tool include: 1) the visual outlook of the headings, concepts, and structures of LCSH; 2) the traversal of LCSH via zooming-in and out; 3) the smart searching capability; 4) the ability to locate multiple positions of specific subject headings across the entire LCSH hierarchical structure; and 5) the ability to demonstrate the synonymy and polysemy control of LCSH. The aforementioned features will be demonstrated with VisuaLCSH in a later section.

2. LCSH Tree
For this study, the hierarchical relationships of LC subject headings are presented in a tree structure, with the tree representing the entire LCSH being called LCSH Tree.

2.1 Conceptual Framework for LCSH Tree
A tree is a well-known data structure representing a hierarchical structure (Cormen et al., 2001). It consists of a set of nodes with a special node (called root node) placed at the highest level of the tree and links between the various nodes. There is only one root node per tree. A node placed one level immediately higher (lower) than the current node is called the parent (child) node. A node that does not have any child nodes is called a leaf node, and is the last node along the path starting from the root node. The root node of a tree is the tree’s only node that does not have a parent node.

Let LCSH Tree be a tree in which each term in LCSH is denoted by a tree node. Four types of terms can be found in a subject authority record: established terms, Related Terms (RTs), terms for Used For (UF), and Broader Terms (BTs). In LCSH, hierarchical relationships between terms are not always rigorously defined, particularly with regard to subject heading strings (i.e., combinations of main heading with one or more subdivisions). As a result, this study focuses on LC main headings only, due to difficulties involved in determining broader and narrower relationships among subject heading strings involving multiple facets.

A hierarchical structure can be built in such a way that a relationship between an established term and its BTs are denoted by child-parent nodes in the tree. RTs and terms for UF are excluded from constructing the hierarchical structure in the tree because they are not in a hierarchical relation with established terms. UF terms are included solely as access points in the searching of the VisuaLCSH system.

Once the algorithm for constructing a LCSH Tree is applied to the LC subject authority file, the hierarchical structure in the authority file is represented by a set of n trees. Each of the n trees is independent of the remaining n-1 trees, meaning that no hierarchical relation (link) exists across different trees. Each of the n trees is called a concept tree as it represents a concept. To connect all the n trees into one, an artificial node is created and used as a parent node for all the collective n trees. The resultant single tree becomes the LCSH Tree where the fictitious node becomes its root node. Hence, the LCSH Tree consists of n trees called LCSH sub-trees as they are all located under one root node. The
complex structure of the LCSH Tree (concatenation of n LCSH sub-trees) constitutes the hierarchical structure of LCSH.

2.2 Development of VisuaLCSH
In this study, a retrospective version of the LC subject authority file, consisting of 291,000 authority records and covering the years 1986 to 2005 available via the Library of Congress Cataloging Distribution Service, is utilized to build a LCSH Tree. Other resources, such as bibliographic records, are not considered because we want to build the Tree based upon subject headings authorized by LC only so that we have a set of subject headings validated by a single authority. The subject headings for topical terms (specifically headings in the MARC field 150 subfield code a) are used in building the LCSH Tree, excluding juvenile headings (beginning with sj).

An LCSH Tree is constructed based on the algorithm and dataset described above. The 2005 retrospective version of LC subject authority files produces 28,136 (equivalent to the n above) trees, i.e., 28,136 LCSH sub-trees. The LCSH Tree is composed of a concatenation of 28,136 trees and an artificial root node placed over the combined trees. The sizes of the LCSH sub-trees range from sub-trees consisting of one node to some with a few ten thousand nodes.

The LCSH visualization tool is developed to display the resulting LCSH Tree, equipped with text-based and structure-based searching features. The LC subject authority file is manipulated and imported into an Access database which becomes the experimental LC subject authority database. The visualization system for the LC subject authority file is implemented using Microsoft Visual Basic 6.0.

3. LCSH Visualization Tool: VisuaLCSH
Two key functions of the LCSH visualization tool are: 1) navigation of the LCSH tree; and 2) searching of LC subject headings. Some additional functions not fully implemented in pre-existing LCSH tools, which were discussed in the introductory section, will also be demonstrated in this section.

3.1 Browsing LCSH Tree
To browse the LCSH Tree, the LC subject authority Access database should be selected in the ‘Browsing LCSH’ section (see Figure 1). All the root nodes (top-level) of the 28,136 LCSH sub-trees will then automatically appear, in alphabetical order, under the ‘LCSH sub-trees’ section. Using scroll bars on the right and bottom, users are able to identify and locate top-level subject headings of the sub-trees. When a top-level subject heading is clicked, the corresponding LCSH sub-tree is displayed on the ‘Selected LCSH sub-tree’ section. Figure 1 shows a resultant screen when the ‘Child care’ top-level subject heading is selected. In the ‘Selected LCSH sub-tree’ section, the plus (+) symbol is shown in front of subject headings only when there exists at least one child node. By clicking the plus (+) symbol, the selected node will expand, leading to the display of the immediate child nodes of the selected node. On the contrary, when a minus (-) symbol is clicked, the sub-tree under the node with the minus (-) will be collapsed back into one node. By zooming-in and out using the symbols, every node in the selected LCSH sub-tree can be visited, and the full hierarchical structure of the selected LCSH sub-tree can be displayed.
3.2 Searching LCSH Tree
Figure 2 shows a VisualLCSH resultant screen when the term ‘child care’ is searched. There are two search options for matching: complete (or exact) word-matching and partial word-matching. With partial word-matching, matching between headings and search terms is made on a word-by-word (character-by-character) basis. For example, a search for ‘Child care workers’ will not match the term ‘Child care worker’ with the option of complete word-matching, but will retrieve it when using the partial word-matching option.

When a search term is typed in the ‘Searching LCSH’ section, the system identifies three LCSH sub-trees, each of which contains at least one node whose headings are com-
pletely word-matched with the search term. The three top-level subject headings (‘Child care’, ‘Child care workers’, and ‘Persons’) of the LCSH sub-trees are displayed in the ‘LCSH sub-trees’ section. Clicking on a top-level subject heading (‘Child care’ in Figure 2) triggers two subsequent events: 1) the hierarchical structure of the corresponding LCSH sub-tree appears in the ‘Selected LCSH sub-tree’ section; and 2) the complete paths from the top-level node to the nodes containing the search term are displayed in the ‘List of hierarchical paths to search term’ section. In the ‘Selected LCSH sub-tree’ section, the nodes matched with a search term are highlighted with a black background. In the ‘List of hierarchical paths to search term’ section, nodes of different levels are designated by the ‘>’ symbol with general headings on the left side of the symbol and specific headings on the right. The ‘I’ in parenthesis at the end of a path indicates that the attached node is an internal node (not a leaf node) of the sub-tree, whereas the ‘L’ designates a leaf node.

3.3 Multiple locations of a heading in LCSH hierarchies

Using the VisualLCSH search function, users are able to identify how many times a heading or search term appears in LCSH as well as where it is located within the hierarchy of the LCSH Tree. For example, results for the term ‘weblogs’ searched with the complete option show that ‘weblogs’ as a heading appears once in each of two LCSH sub-trees, of which the top nodes are ‘Auxiliary sciences of history’ and ‘Biographical sources.’ The paths to the nodes are: ‘Auxiliary sciences of history > Civilization > Social Sciences > Sociology > Communication > Information science > Information resources > Electronic information resources > Computer network resources > Web sites > Weblogs’ and ‘Biographical sources > Diaries > Weblogs.’

3.4 Multiple LCSH terms with synonym/homograph control

When searching normally, only main headings are used for matching. VisualLCSH, however, provides the option of searching synonyms via UF references. With the synonym option on, both the main heading and its UF references are taken into account in the search process. Let us repeat the same search conducted in Figure 2 with the synonym option on. This time, an additional heading (‘Nanny placement agencies’) is retrieved as a result of the search because a UF (‘Home child care workers placement agencies’) contains the search term (‘child care’) even though the heading itself does not contain it.

By using VisualLCSH, users can also benefit from the terminological control of homographs in LCSH. Type-in ‘mercury’ as a search term with default search options (complete word-matching and no synonym searching) to see how the homographs of the term are controlled. The search reveals the following results, each of which presents a different meaning of the word ‘mercury’:

- Minerals > Ores > Metals > Liquid metals > Mercury
- Transportation, Automotive > Automobiles > Ford automobile > Mercury automobile
- Gods, Roman > Mercury (Roman deity)

Note that other homographic examples, such as ‘Mercury (Planet)’, are not searchable with the current version of this system as the heading is defined under MARC field 551 (heading — geographic name) which is not included in the authority Access database.
3.5 LCSH’s irregular hierarchical structures

A well-recognized problem with LCSH is its irregular hierarchical structure: “the relationships [of equivalence, association, and hierarchy] are inconsistent and may not exist at all on older terms.” (Library of Congress, 2008, 34) This new visualization tool does not point them out automatically. It greatly helps, however, in identifying where such flaws occur, especially when there are suspicious locations or terms. For example, while inspecting the lineage of mammals, the conflicting hierarchical structures embedded in LCSH are found as follows:

- Life (Biology) > … Vertebrates > Mammals > Primates > Cercopithecidae
- Life (Biology) > … Vertebrates > Mammals > Primates > Monkeys
- Life (Biology) > … Vertebrates > Mammals > Primates > Monkeys > Cercopithecidae

According to the first and third chains of subjects, the subject Cercopithecidae is immediately connected to both the subject Primates and to the subject Monkeys on different levels of the hierarchy. If the first and second chains are true, then the third must be false. If the second and third are true, however, the first must be false. Thus, two existing contradictory lineages are identified.

Another example of irregularity can be found in a search on ‘literature.’ It can be expected to see a single LCSH sub-tree beginning with the ‘literature’ heading on the top. However, the system returns approximately 100 different LC sub-trees, each of which appears to represent the literature of a specific region or genre. For example, headings at the top-level in the sub-trees include: ‘Greek literature’; ‘Haitian literature’; ‘Romani literature’; and ‘West African literature (French)’. There are no hierarchical links between any of the subject-trees, thereby indicating that they are separate concepts in terms of hierarchy, which is not truly the case.

4. Conclusion

This project proposes a conceptual framework of LCSH and develops a new tool for visualizing the structure of LCSH. Major functions of the new tool are described and demonstrated above, including features that have not been embedded in existing tools. The highlights of VisualLCSH are summarized below: 1) it reveals multiple aspects of a heading; 2) it is a tool for normalizing the hierarchical relationships in LCSH; 3) it shows multi-level hierarchies of terms in LCSH sub-trees; 4) it improves the navigational function of LCSH in retrieval; and 5) it enables the implementation of generic searching, i.e., the ‘exploding’ feature, in LCSH.

The conceptual framework shown in the LCSH Tree and VisuaLCSH may contribute to the utilization, development, maintenance, and application of the current LCSH in the following ways: 1) Enhancing subject searching in LCSH: In most cataloguing systems, one can search a heading or a word, but not the relationships between headings (Weinberg, 1998, 371); VisuaLCSH provides an enhanced means of searching by presenting the complete hierarchy of LCSH to users; 2) Building a hierarchy: The developed algorithm and software may be applied to other controlled vocabulary lists in which a hierarchy is embedded; the hierarchical display facilitates the generation of mini-thesauri based on a subset, such as LCSH sub-trees of the full structure; 3) Maintaining and controlling quality: This tool can be utilized to review and verify the hierarchical structure of LCSH; 4) Automating organization/classification based on LCSH: LCSH-based automated classification can play a central role in building digital libraries using traditional library classification schemes (Koch & Day, 1997; Yi, 2005); the LCSH concept trees can be used...
as a basis for organizing electronic resources; 5) Mapping to folksonomies: Facilitating mapping of LCSH to folksonomies or user-assigned tags; 6) Enabling interoperability with other controlled vocabularies: A number of previous studies attempted to link LCSH to other controlled vocabularies (Vizine-Goetz et al, 2004); the notion of concept trees may be useful in future interoperability research.

Acknowledgement
The authors would like to thank unknown reviewers for their valuable comments and suggestions.

References