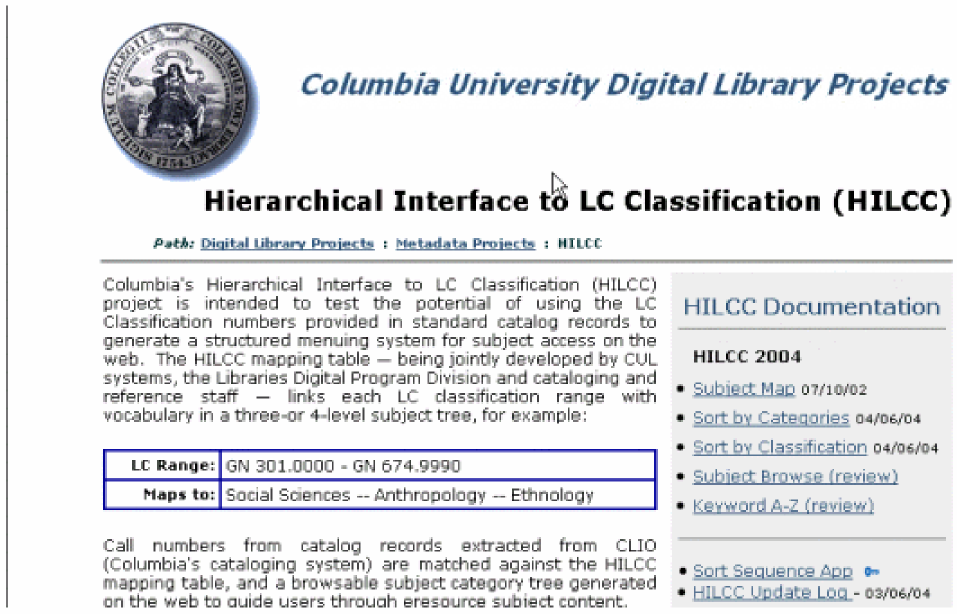
HOME

Taking it to the HILCC:
Automated classification and subject analysis under study in CTS

Providing enhanced subject analysis to library resources via automated means is a long-standing research problem in the cataloging community. "Self-declaring resources", automated classification from subject headings, and other means of using machines to analyze information content and provide access via subjects or index terms have been subjects of study, both theoretical and practical. Some real-life applications have been realized. For example, OCLC has offered mapping from Library of Congress Subject Headings to Dewey Decimal Classification numbers via WorldCat for several years. But for the most part, the task of distilling the "aboutness" of a given resource and assigning it to a particular category or classification scheme has largely been resistant to complete automation.

Two librarians in CTS are currently working with a new tool, the Hierarchical Interface to LC Classification, or HILCC, to explore the possibilities of using automation to expand and potentially customize subject access to library resources. HILCC was developed at Columbia University Libraries and is currently being used there to generate a structured menuing system for subject access to electronic resources. Library of Congress call number ranges are mapped to a table of related subject terms. The table allows a browsable subject category "tree" to be generated to assist users in navigating through e-resource subject content on the Web. Columbia is currently using the HILCC software in its browsable E-Journal lists and an A-Z subject list for electronic resources of all kinds.



Columbia University Digital Library Projects

Hierarchical Interface to LC Classification (HILCC)

Path: Digital Library Projects : Metadata Projects : HILCC

Columbia's Hierarchical Interface to LC Classification (HILCC) project is intended to test the potential of using the LC Classification numbers provided in standard catalog records to generate a structured menuing system for subject access on the web. The HILCC mapping table — being jointly developed by CUL systems, the Libraries Digital Program Division and cataloging and reference staff — links each LC classification range with vocabulary in a three-or 4-level subject tree, for example:

LC Range:	GN 301.0000 - GN 674.9990
Maps to:	Social Sciences -- Anthropology -- Ethnology

Call numbers from catalog records extracted from CLIO (Columbia's cataloging system) are matched against the HILCC mapping table, and a browsable subject category tree generated on the web to guide users through eresource subject content.

HILCC Documentation

HILCC 2004

- Subject Map 07/10/02
- Sort by Categories 04/06/04
- Sort by Classification 04/06/04
- Subject Browse (review)
- Keyword A-Z (review)

- Sort Sequence App
- HILCC Update Log - 03/06/04

The HILCC Web site at Columbia

In early February, Karen Calhoun, Associate University Librarian for Technical Services, Adam Chandler, CTS Information Technology Librarian, and Jim LeBlanc, head of CTS Post-Cataloging Services, traveled to New York to visit colleagues at Columbia and talk about the HILCC system. Columbia library staff and administrators graciously agreed to share HILCC with CUL for research purposes, and CTS is now working with HILCC to explore its potential in a different way. Currently, Adam and Jim are comparing the Columbia HILCC classification to the call numbers contained in the Uris collection. Their goal: to see if it is feasible to use the HILCC topics as a navigation aid to an undergraduate print collection. Uris Library currently holds about 150,000 titles. One of the research questions under study is whether HILCC "scales-up" as a browsing method for a collection of this size. Initial observations suggest that, at least in its current form, it does not. One reason is a mismatch between the HILCC subject distribution and that of the Uris materials. For example, two Uris call number ranges contain over 13,000 titles, while dozens contain none or only a few; the working assumption is the distribution should be more even.

The next phase of the plan involves comparing the results of the Uris frequency distribution to see if it may be possible to adjust HILCC to better suit the Uris collection. Adam and Jim are working to match the Uris collection HILCC histogram to the complete LC Classification tables for certain HILCC subjects. It seems clear that HILCC would need to be modified at the high end, to break subjects like "Languages & Literatures -- English -- English Literature" into smaller pieces. To keep this work moving, a working assumption about what the upper browse threshold is likely to be (300 titles? 500?) needs to be established. One possible outcome of the HILCC research would be a user-friendly, Web-based interface that will assist undergraduates in finding print resources appropriate to their needs and interests. Assuming that model were to work, it might be possible to further divide the catalog into other subject-focused browse lists, so that specialists in a particular subject area could look for resources tailored to their needs. Adam and Jim plan to publish the findings of their work in a forthcoming article.

For more about HILCC, visit the HILCC site at:

http://www.columbia.edu/cu/libraries/inside/projects/metadata/hilcc/

CUL's investigations into HILCC can be viewed here:

http://www.library.cornell.edu/cts/browseandextend/