Exploring the Potential of a Virtual Undergraduate Library Collection Based on the Hierarchical Interface to LC Classification (HILCC)

Authors:

Adam Chandler
Information Technology Librarian
Central Technical Services
Cornell University Library
Ithaca, NY 14853
alc28@cornell.edu

Jim LeBlanc
Head, Database Management Services
Central Technical Services
Cornell University Library
Ithaca, NY 14853
jdl8@cornell.edu

October 3, 2005

Exploring the Potential of a Virtual Undergraduate Library Collection Based on the Hierarchical Interface to LC Classification (HILCC)

The authors wish to thank Stephen Davis and Bob Wolven at Columbia University for their assistance with this project, and Karen Calhoun at Cornell for her advice and encouragement, as well as the original idea of testing the feasibility of adapting the Columbia HILCC model for use with a Cornell print collection. They also wish to thank the libraries of Columbia University, the University of Illinois at Urbana-Champaign, Indiana University, and the University of Washington for their quick and gracious response to our request for data, as well as their Cornell colleague, Phil Davis, for his expertise and advice on the use of Pearson correlations.

ABSTRACT: *The Hierarchical Interface to Library of Congress Classification (HILCC) is a system developed by the Columbia University Library to leverage call number data from the MARC holdings records in Columbia's online catalog to create a structured, hierarchical menuing system to provide subject access to the library's electronic resources. In this paper, the authors describe a research initiative at the Cornell University Library to discover if the Columbia HILCC scheme can be used as developed by Columbia, or in modified form, to create a virtual undergraduate print collection outside the context of the traditional online catalog. Their results indicate that, with certain adjustments, a HILCC model can indeed be used to represent the holdings of a large research library's undergraduate collection of approximately 150,000 titles, but that such a model is not infinitely scalable and may require a new approach to browsing such a large information space.*

In 1997, a working group consisting of staff from Columbia University Library's Bibliographic Control Department and Library Systems Office set out to build a hierarchical interface to Library of Congress classification (HILCC). [1] The project's aim was "to assess the potential of using the Library of

Congress classification numbers as provided in standard catalog records to generate a structured, hierarchical menuing system for subject access to resources in the Libraries' electronic collection." [2] The group sought to leverage Columbia's MARC catalog data to permit "Web-based access to the Libraries' electronic resources outside the context of the OPAC." [3] With help from reference staff and selectors, they created a classification mapping table to link discrete ranges in the Library of Congress classification schedules to entries in a three-, and occasionally four-tiered subject tree. Simultaneously, the group developed a Web interface that would give users access to data extracted weekly from catalog records via this multi-level subject hierarchy. In the end, Columbia produced a HILCC model that provided access to some 5,000 electronic resources by way of 541 distinct subject categories. [4]

In his 2002 article, "HILCC: A Hierarchical Interface to Library of Congress Classification," Davis reflected on the challenges of testing HILCC's effectiveness and overall value. He also questioned the project's scalability — "what may seem useful and manageable against a list of 5,000 electronic titles may look quite different when the list has grown to 50,000 or more." [5] Davis invited other institutions to pick up where Columbia left off, to take Columbia's mapping tables and rework them for their own collections and within their own institutional contexts. In 2004, the Cornell University Library accepted Columbia's invitation and began to explore the theoretical possibility of using HILCC to create a virtual undergraduate collection of Cornell's print material. This essay presents the results and conclusions of that investigation.

<1> Applicability of the Columbia Model to Cornell's Undergraduate Collection

Like most large research libraries, the Cornell Library is facing a serious space problem, especially on the central campus where academic real estate is at an all-time premium. The most common means of dealing with this space crunch is to transfer lesser used material to offsite storage facilities, as well as to merge and reorganize those collections that remain onsite. As the Association of Research Libraries (ARL) noted in 1999, "Most ARL libraries already house a significant amount of material in

offsite storage facilities, and the pace of both new construction and renovation of existing structures has accelerated during the past decade." [6] This trend continues unabated and there is no indication that it will lessen in coming decades. On May 18, 2005, for example, the University of Texas at Austin Libraries announced that they are "relocating the Undergraduate Library (UGL) to other discipline-specific campus libraries in their system as the first step in the process of transforming the Flawn Academic Center (FAC) into an integrated, learning commons." [7]

Thus, ARL institutions can benefit from studying ways to create print collections without regard for physical contiguity of the collections' individual elements, collections that transcend the limits of storage locations — that are virtual collections of physical items.

The question the authors sought to answer was the following: can Cornell use Columbia HILCC mapping to represent Cornell's current undergraduate print collection of approximately 150,000 titles — that is, would the Columbia mapping be transferable from one research library's e-resource collection to another research library's print collection and, perhaps more importantly, would it scale?

The HILCC mapping tables identify alpha-numeric call number ranges in the LC classification schedules that correspond to subject categories in the system's user interface. Each unique subject string is assigned a numeric label in the table (here called a "subject code"). Although classification ranges are mapped to one and only one unique subject string, these hierarchical subject categories are often the product of more than one LC classification range.

[INSERT TABLE 1]

Table 1 presents an excerpt from Columbia's HILCC charts that deals with LC's "A" schedule and Columbia's "General" subject categories. Each row is numbered and indicates a unique range in the LC classification. The range itself is recorded in the "CLASS…" columns. The HILCC subject string is recorded in the "CATEGORY…" columns. The last column contains the subject code, which links single or multiple rows to a single HILCC subject string, as applicable. For example, rows 342 and 343 in the excerpted table represent the LC classification ranges AC0 through AC799 and AC900 through AC1100,

both mapped to the Columbia HILCC subject string "General—Collections & Series (General) and numerically labeled with the subject code 1270.

In order to apply the Columbia mapping scheme to the titles in Cornell's undergraduate library, the authors extracted all the call numbers from the undergraduate collection, using a program written by Peter Hoyt from the Cornell University Library Systems Office. They then wrote a Perl script to match each call number to a range in the Columbia HILCC tables to derive the corresponding HILCC subject string and increment the count. [8] The authors then output the results as a delimited file for analysis.

The results of running the Cornell undergraduate library's call numbers against the Columbia HILCC scheme were not promising. Although, in many cases, the number of titles assigned to individual subject strings represented what might be construed as a reasonably manageable retrieval set, in other cases the number of titles assigned to individual subject strings was quite high.

[INSERT TABLE 2]

Table 2 illustrates the incremental breakdown of titles per subject string for both Cornell and Columbia. The first column in the table lists hit rate ranges, from no titles retrieved for a given subject string to a maximum range of 10,001 to 15,000 titles retrieved. The second column records the number of subject strings that fell into these retrieval spans for the Cornell undergraduate collection. The third column gives this number as a percentage of the total number of HILCC categories. The fourth and fifth columns provide the same data for Columbia's e-resource collection. Although the number of titles in several of the retrieval sets were similar to those extracted by Columbia for their e-resource collection, Cornell's use of Columbia's HILCC scheme resulted in several subject categories that comprised more than 1,000 titles, 10 subject categories that yielded more than 2,500 titles, and 2 subject categories that, if searched in a live database, would return more than 10,000 title hits. At the same time, some 42 percent of all Columbia HILCC subject categories yielded 10 or fewer titles, with 121 categories retrieving no results at all. This histogram suggests that using the Columbia HILCC scheme, as is, would not lead to optimal results if applied to Cornell's undergraduate print collection — at least not with that collection's

current content. Imagining an effective interface, given current technology, that would accommodate retrieval sets of this size using a structured hierarchical menu system is difficult.

The authors did speculate, however, that Cornell might be able to modify Columbia's HILCC scheme to better fit the test case. Before starting to edit the mapping tables, though, they investigated the potential applicability of a revised HILCC scheme for other libraries' undergraduate collections. They solicited data from four other ARL libraries (the Columbia University Library, the University of Illinois at Urbana-Champaign Library, the Indiana University Libraries, and the University of Washington Libraries), ran their call numbers through the Perl scripts described above, and derived a Pearson correlation of the results. The correlation shows the relative similarities in number of titles per subject category among these library collections, using a count of the number of titles that map to a given HILCC subject category for each of the sample libraries and comparing the numbers. The tendency was similarity between institutions; that is, a given category that had a high number of titles at one institution was likely to have a high number of titles at other institutions. Based on the results of this analysis, the authors went forward with the revision of Columbia's HILCC scheme, confident that the retailored tables would be generally useful to other libraries who wish to build further on Columbia's or Cornell's work.

<1> The Cornell Model

In order to create a more usable HILCC scheme for Cornell's test collection, the authors needed to make some assumptions. First of all, how many titles encompassed by a single HILCC subject strings are too many? Second, at what point does the hypothetical user interface contain too many subjects — too many branches and hierarchical levels on the subject tree — if modifying a HILCC scheme requires splitting the strings into further categories and sub-categories? The authors decided on two mutually supportive and presumably manageable parameters to address these concerns: editing Columbia's HILCC tables in such a way that no subject string would apply to more than 1,000 or fewer than 10 titles in Cornell's undergraduate library. Reducing the higher hit rates would necessarily result in the creation of

more subject categories. However, combining categories that originally returned very few or no title hits would decrease the number of subject categories and, hopefully, counterbalance the effects of this expansion. As Columbia had done, the Cornell researchers opted to use the LC classification schedules for guidance on how to split or merge categories logically. When new subject category names were required, they would, for efficiency's sake, assign whatever seemed most appropriate without consultation with colleagues from other departments (though actual implementation of a revised HILCC scheme to build a live user interface would call for broader input along the lines of Columbia's implementation model). Finally, the authors decided to examine the physical material *in situ*, rather than through automated methods, since they had only a general notion about how best to approach the reorganization of the tables and no clear idea about how easy or difficult it would be to slice and dice extremely narrow classification ranges containing hundreds of titles.

The work of restructuring the Columbia HILCC tables for use with Cornell's undergraduate collection took roughly 65 real-time hours and resulted in a revamped scheme of 500 subject strings (8 percent fewer than Columbia HILCC) in a five-level subject tree (one level deeper than Columbia HILCC). No subject string encompassed more than 1,000 titles and only 4 strings retrieved fewer than 10 titles. The four subject strings that retrieved fewer than 10 titles were left as is because there seemed to be either no reasonable way to combine them with other categories or a clear expectation that the hit rate would increase gradually over time. For example, the string History & Archaeology—Regions & Countries—United States—Local History—Territories, Protectorates, Etc., though yielding only 2 titles, cannot be logically associated with any of the other fifth-level regional categories associated with U.S. local history. The string Languages & Literatures—English—English Literature—Individual Authors—2001- , retrieving only 7 titles, is an example of a subject category that is expected to grow. The two Columbia subject categories that required the greatest adjustment were those for Languages & Literature—English—American Literature, and Languages & Literatures—English—English Literature, yielding 13,906 and 13,173 title hits respectively, using the unedited Columbia categories. The authors broke these subject strings down into 32 and 33 new categories respectively, in a five-level structure.

Table 3 shows 9 of the 33 categories into which the authors split the original subject string, Languages & Literatures—English—American Literature, in order to bring the retrieval set to within the 10 to 1000 hit range for each remapped subject string. Thus, hypothetical users of Cornell's revised HILCC scheme for American literature would be able to specify more precisely the subject categories they wished to browse and retrieve a more manageable number of titles with each search, though they would need to drill down an additional 2 levels in the subject tree to do so.

[INSERT TABLE 3]

Tables 4 and 5 illustrate a segment of Columbia HILCC in which more than one subject string covered fewer than 10 titles, an area revised according to the 10 to 1000 hit range parameters of the project. This reorganization of the "General" category resulted in a 25 percent reduction in the number of categories for that subject area.

[INSERT TABLES 4-5]

Through this two-pronged strategy, the Cornell researchers were thus able to restructure the Columbia HILCC scheme, developed to provide the underpinning for a Web-based hierarchical menuing system for subject access to that library's collection of electronic resources, into a revised scheme to provide (theoretically) the basis for a similar menuing scheme for subject access to Cornell's undergraduate collection — all with a moderate amount of human intellectual effort. They also demonstrated (once again, in theory) that such a framework is scalable up to approximately 150,000 titles.

This menuing scheme is unlikely to be infinitely scalable, however. Mapping the entire Cornell University Library against the revised HILCC categories, using the same 10 to 1000 title hit range, would require over 12,000 categories, with significantly more hierarchical tiers, to represent the complete Cornell collection of some 4.4 million titles (the number of Cornell's total holdings in late summer 2004). The authors derived this projection by examining the results of their first attempt to map Columbia HILCC, as is, against Cornell's undergraduate holdings, then calculating the average number of additional subject categories required to bring the retrieval sets of 1,000 titles or more into the desired target range of 10 to 1000 hits per subject string (see Table 6). Then they extracted the call numbers for all 4.4 million

titles held by Cornell, mapped them to the revised HILCC scheme, and used the conversion factor derived from the calculation above (2.87) to estimate how many additional subject strings would be necessary to retrieve 1,000 titles or less for each and every string in the scheme (assuming that when mapping the entire Cornell collection, no fewer than 10 hits would be represented in any subject category — an assumption that proved to be correct). An excerpt from the chart representing some of the large, medium, and small retrieval sets revealed in this exercise appears in Table 7.

[INSERT TABLES 6-7]

Thus, although modifying the Columbia HILCC scheme to create a menuing system for a typical ARL undergraduate collection should be possible, HILCC's scalability is limited. Databases of approximately 150,000 titles may be approaching the limits of a HILCC scheme's effectiveness.

<1> Next Steps: Browsing and Visualization

As previously noted, the decision to limit retrieval sets in the Cornell remapping of Columbia HILCC to 1,000 titles or less, was somewhat arbitrary. The presentation of that many search results using a conventional library catalog interface (or a Google-type interface, for that matter) would be decidedly unwieldy. Even limiting the results to no more than 500 titles or to as few as 200 titles would present obstacles to quick and easy browsability. Before pursuing further work with HILCC schemes, researchers should consider questions of browsability and visualization of search results. How many search results are too many? In a comprehensive investigation of University of California's (UC) MELVYL Systemwide Library Catalog use over 479 days in 1998 and 1999, Michael D. Cooper found that users of the system, which at the time included the catalogs of nine campuses plus other institutions and some additional citation databases, on average displayed 4-5 citations per session, or 2 - 3.5 citations per 100, depending on the database searched. [9] Cooper highlights the fact that the amount of time users allocated to displaying results was steady across databases, between 30 - 40 seconds per session, and speculates that "one explanation is that irrespective of the database, there are certain motor limits in place when an

individual scans citations on a screen that keep the time relatively constant." [10] Jansen, Spink, and Saracevic discovered from Excite search engine data in 2002 that 58 percent of users look at only the first page of 10 results, 19 percent look at the second page and 9 percent will go to the third page. [11] Only a small percentage continue browsing beyond that. They concluded: "any search result beyond the tenth position in the list would be meaningless for 58% of Web users." [12] What, then, is the threshold of usefulness for conventional displays of large retrieval sets or, perhaps more precisely, what is the threshold of their usability? Are the standard modes of presentation of hierarchical menus the best choice for HILCC and similar subject schemes that aim to deliver user-friendly access to large library collections?

The proper response to this last question should be a definitive "no." Although online access to catalog data has sped up and improved users' ability to find and use information about library collections, computer interfaces have, in some ways, reduced the capacity to browse these collections by limiting one's sense of the overall contents of a library. Scrolling through screen after screen of surrogate data is not always a good substitute for moving freely through library stacks, where one's eyes may catch a broad peripheral glimpse of dozens of items at a time, while honing in on particular pieces for one reason or another. While one could argue that browsing a collection through the mediation of a computer monitor, keyboard, and mouse merely calls for a different approach to browsing — a reorientation of browsing techniques, if you will — technology should permit a scope that is at least as broad as the traditional library browsing space.

In a 2004 contribution to *D-Lib Magazine*, Dushay introduced a prototype for just such an online mechanism. [13] Developed for use with the National Science Digital Library, the NSDL Virtual Book Spine Viewer addresses the "focus + context problem:" That is, it optimizes the utility of the browsing software by allowing the user to focus on details, without sacrificing the "larger context of the information space." [14] "Figure 4. Virtual Spine Viewer" in Dushay's paper shows how such a browser would work. The frame at the left of the screen contains a subject-based, hierarchical menu that is smaller, but still similar to that of HILCC. The middle frame represents the subject category space in

which the "book spines" are arrayed so that they can be seen at a glance. Within what is technically known as a "scatter plot ZUI" (or Zooming User Interface), the titles are "scattered" along horizontal and vertical axes over which the user can position a kind of virtual magnifying glass to select potentially useful titles in greater bibliographic detail. The upper boxes in the frame at the right allow users to customize the layout of items on the two scatter plot axes by prioritizing two search variables. The lower box in the right frame displays additional bibliographic information pertaining to the "book" selected. Dushay's virtual book spine viewer provides both focus and context, and creates a browsing environment that more resembles the "traditional information space" than other online browsing tools. Dushay's viewer is one example of a next generation browser that might allow for better and more useful access to retrieval sets that currently fill more than two or three results screens. Dushay and others doing research into browsing information systems could benefit from an examination of the extensive pre-World Wide Web research conducted on browsing online library catalogs, summarized by Kurth and Peters. [15]

<1> Conclusion

Although the research described in this paper stopped short of actually building and implementing an interface with which to provide bibliographic access to a virtual undergraduate collection, a library might put the Columbia and Cornell HILCC schemes to other uses. A complete histogram of hit rates across all HILCC categories, like the one developed in this study, might be a useful tool for analyzing the subject scope of an existing collection — whether destined for undergraduates or delimited by other user or subject criteria. Further, if holdings in a given collection were mapped periodically against the same HILCC categories, the results might paint a revealing picture of how recent additions reflect (or do not reflect) the pre-existing or presumed subject orientation of that collection and how its subject focus might be changing. HILCC tables could also be used to analyze interlibrary loan trends and patterns (a use to which Columbia has reportedly begun to apply Cornell's revised HILCC

scheme). One could even envision the merger of all three of these data sets into a single graph to track the direction of collection growth and demand over time.

There are clearly several directions in which further research might go. It is not yet clear whether the Cornell University Library will implement a HILCC-based interface to provide bibliographic access to its physically collocated current undergraduate collection or, in the future, to a physically dispersed collection, if space concerns on the central campus demand such redistribution of the physical material. Like the Columbia University Library, however, the authors invite other institutions to pick up where they've left off, to adopt and customize Cornell's modified HILCC scheme for use in their own collections and within their own institutional contexts, and to explore further the possibility of using scatter plots ZUIs, hyperbolic trees, and other information visualization techniques to present HILCC data optimally to the end user.

At a philosophical level, though, the problems with hierarchical classification, even with an improved display, run deep. Shirky recently argued in "Ontology is Overrated: Categories, Links, and Tags" that the evolution of the Web itself shows the inherent brittleness of all attempts at authoritative classification for very large domains. [16] The most ambitious attempt at classifying the Web may be Yahoo's, but who really uses their classification now? Rather than trying to impose ordered classification on users what may work better is to allow users themselves to classify things any way they see fit and to create order and paths of exploration derived from the raw data. The success of Google searching, based as it is on links across Web sites, is the best example of such an approach. Those who pursue further research on the creation of virtual collections will need to focus on this trend as well.

References

1. Stephen Paul Davis, "HILCC: A Hierarchical Interface to Library of Congress Classification," *Journal of Internet Cataloging*, 5, no. 4 (2002): 19-49.

2. Davis, "HILCC," 19.

3. Davis, "HILCC," 21.

4. Columbia University Library,"E-Journals Subject Browse." Accessed May 27, 2005, www.columbia.edu/cu/lweb/eresources/ejournals/subjects.

5. Davis, "HILCC," 40. As noted in Table 2 of the current article, Columbia's e-resource holdings have now in fact exceeded 50,000 titles.

6. Association of Research Libraries, *Library Storage Facilities, Management, and Services, SPEC Kit*, 242 (Washington, DC : ARL, May 1999).

7. University of Texas Libraries, "Relocation of UGL Book Collection to Begin." Accessed May 27, 2005, www.lib.utexas.edu/about/news/fac-bookrelocation.html.

8. For more on these scripts and the way they interact, see Section II, Step 3 of the Cornell HILCC project log. Accessed May 27, 2005, www.library.cornell.edu/cts/browseandextend/.

9. Michael D. Cooper, "Usage Patterns of a Web-Based Library Catalog," *Journal of the American Society for Information Science and Technology*, 52, no. 2 (January 15, 2001): 145.

10. Michael D. Cooper, "Usage Patterns of a Web-Based Library Catalog," 145.

11. Bernard J. Jansen, Amanda Spink, and Tefko Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," *Information Processing and Management*, 36 (March 2000): 207-227.

12. Jansen, Spink, and Saracevic, "Real Life, Real Users, and Real Needs," 214.

13. Naomi Dushay, "Visualizing Bibliographic Metadata: A Virtual (Book) Spine Viewer," *D-Lib Magazine*, 10, no. 10 (2004). Accessed May 27, 2005, www.dlib.org/dlib/october04/dushay/10dushay.html.

14. Dushay, "Visualizing Bibliographic Metadata."

15. Martin Kurth and Thomas A. Peters, eds., *Browsing in Information Systems: An Extensive Annotated Bibliography of the Literature* (Ann Arbor, Michigan: Pierian Press, 1995.

16. Clay Shirky, "Ontology is Overrated: Categories, Links, and Tags." Accessed May 27, 2005, www.shirky.com/writings/ontology_overrrated.html.

Table 1. Excerpt from Columbia University Library's HILCC mapping table (from "Columbia University Digital Library Projects: Hierarchical Interface to LC Classification, Arranged by Class Number Range 03/05/04." Accessed Sept. 2, 2005, www.columbia.edu/cu/libraries/inside/projects/metadata/hilcc/newfiles/class.html)

ROW_ID	CLASS1A	CLASS1N	CLASS2A	CLASS2N	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	SUBJECT CODE
342	AC	0.0000	AC	799.9990	General	Collections & Series (General)			1270
233	AC	800.0000	AC	899.9990	General	Dissertation Indexes			1004
343	AC	900.0000	AC	1100.9990	General	Collections & Series (General)			1270
344	AE	0.0000	AE	90.9990	General	Encyclopedias (General)			1271
345	AG	0.0000	AG	600.9990	General	Dictionaries (General)			1272

Table 2. Number of titles per HILCC subject: Cornell and Columbia

TITLES PER SUBJECT	CORNELL UNDERGRADUATE	PERCENTAGE	COLUMBIA E-RESOURCES	PERCENTAGE
0	121 subjects	22%	166 subjects	31%
1-10	108 subjects	20%	118 subjects	22%
11-50	102 subjects	19%	111 subjects	21%
51-100	56 subjects	10%	39 subjects	7%
101-250	57 subjects	11%	52 subjects	10%
251-500	35 subjects	6%	20 subjects	4%
501-1000	27 subjects	5%	22 subjects	4%
1001-2500	25 subjects	5%	10 subjects	2%
2501-5000	7 subjects	1%	2 subjects	<1%
5001-10000	1 subject	<1%	1 subject	<1%
10001-15000	2 subjects	<1%	0 subjects	0%
Total Titles	150,200		64,830	
Mean Titles/Subject	210		120	
Median Titles/Subject	16		8	
Standard Deviation	844		171	
No HILCC Hit	7993		570	

Table 3. Cornell HILCC Subject Categories for American Literature, with Cornell Undergraduate Title Count (Excerpt)

SUBJECT CODE	CLASS1A	CLASS1N	CLASS2A	CLASS2N	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	CATEGORY5	COUNT
1203.1	PS	1.0000	PS	144.9990	Languages & Literatures	English	American Literature	History & Criticism	General	462
1203.2	PS	147.0000	PS	195.9990	Languages & Literatures	English	American Literature	History & Criticism	Special Classes of Authors & Subjects	234
1203.3	PS	201.0000	PS	228.9990	Languages & Literatures	English	American Literature	History & Criticism	19th-20th Centuries (General)	142
1203.4	PS	241.0000	PS	286.9990	Languages & Literatures	English	American Literature	History & Criticism	Special Regions & States	69
1203.5	PS	301.0000	PS	379.9990	Languages & Literatures	English	American Literature	History & Criticism	Poetry, Drama & Prose (General)	517
1203.28	PS	3550.0000	PS	3553.9990	Languages & Literatures	English	American Literature	Individual Authors	1961-2000, A-C	763
1203.29	PS	3554.0000	PS	3559.9990	Languages & Literatures	English	American Literature	Individual Authors	1961-2000, D-I	767
1203.30	PS	3560.0000	PS	3564.9990	Languages & Literatures	English	American Literature	Individual Authors	1961-2000, J-N	779
1203.33	PS	3600.0000	PS	3626.9990	Languages & Literatures	English	American Literature	Individual Authors	2001-	45

Table 4. Columbia HILCC Subject Categories for General Works, with Cornell Undergraduate Title Count

SUBJECT CODE	CLASS1A	CLASS1N	CLASS2A	CLASS2N	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	COUNT
1270	AC	0.0000	AC	799.9990	General	Collections & Series (General)			202
"	AC	900.0000	AC	1100.9990	General	Collections & Series (General)			*****
1004	AC	800.0000	AC	899.9990	General	Dissertation Indexes			0
1271	AE	0.0000	AE	90.9990	General	Encyclopedias (General)			14
1272	AG	0.0000	AG	600.9990	General	Dictionaries (General)			25
1273	AI	0.0000	AI	122.9990	General	Indexes (General)			13
1274	AM	0.0000	AM	500.9990	General	Museum Publications			11
1275	AN	0.0000	AN	9999.9990	General	Newspapers (General & Popular)			2
1276	AP	0.0000	AP	272.9990	General	Periodicals (General & Popular)			94
1277	AS	0.0000	AS	945.9990	General	Academies & Learned Societies Publications			14
1278	AY	0.0000	AY	2001.9990	General	Almanacs, Directories & Yearbooks (General)			7
1279	AZ	0.0000	AZ	999.9990	General	History of Scholarship & Learning			54
1294	Z	1001.0000	Z	1199.9990	General	Bibliography	Bibliography (General)		53
1295	Z	1200.0000	Z	4999.9990	General	Bibliography	Bibliography (National)		46
1296	Z	5000.0000	Z	7999.9990	General	Bibliography	Bibliography (Subject)		45
1297	Z	8000.0000	Z	8999.9990	General	Bibliography	Bibliography (Personal)		22

Table 5. Cornell HILCC Subject Categories for General Works, with Cornell Undergraduate Title Count (Excerpt)

SUBJECT CODE	CLASS1A	CLASS1N	CLASS2A	CLASS2N	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	CATEGORY5	COUNT
1270	AC	0.0000	AC	799.9990	General	Collections & Series (General)				202
"	AC	900.0000	AC	1100.9990	General	Collections & Series (General)				*****
1298.1	AC	800.0000	AC	899.9990	General	Directories, Indexes, Information Resources (General)				27
"	AI	0.0000	AI	122.9990	General	Directories, Indexes, Information Resources (General)				*****
"	AY	0.0000	AY	2001.9990	General	Directories, Indexes, Information Resources (General)				*****
"	ZA	3038.0000	ZA	5199.9990	General	Directories, Indexes, Information Resources (General)				*****
1271	AE	0.0000	AE	90.9990	General	Encyclopedias (General)				14
1272	AG	0.0000	AG	600.9990	General	Dictionaries (General)				25
1274	AM	0.0000	AM	500.9990	General	Museum Publications				11
1275.1	AN	0.0000	AN	9999.9990	General	Newspapers & Periodicals (General & Popular)				95
"	AP	0.0000	AP	272.9990	General	Newspapers & Periodicals (General & Popular)				*****
1277	AS	0.0000	AS	945.9990	General	Academies & Learned Societies Publications				14
1279	AZ	0.0000	AZ	999.9990	General	History of Scholarship & Learning				54
1294	Z	1001.0000	Z	1199.9990	General	Bibliography	Bibliography (General)			53
1295	Z	1200.0000	Z	4999.9990	General	Bibliography	Bibliography (National)			46
1296	Z	5000.0000	Z	7999.9990	General	Bibliography	Bibliography (Subject)			45
1297	Z	8000.0000	Z	8999.9990	General	Bibliography	Bibliography (Personal)			22

Table 6. Derivation of scaling factor to estimate number of subject categories required to map all of Cornell's holdings to HILCC (excerpt)

SUBJECT CODE	TITLE COUNT	SUBJECT CODES (SPLIT)	SUBJECT CODES (SPLIT) PER 1000 TITLES	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	CATEGORY5
1203	13906	33	2.37	Languages & Literatures	English	American Literature		
1130	13173	34	2.58	Languages & Literatures	English	English Literature		
…	…	…	…	…	…	…		
1113	1115	2	1.79	Social Sciences	Social Welfare & Social Work	Criminology, Penology & Juvenile Delinquency		
1084	1059	3	2.83	Business & Economics	Economics	Industries		
TOTAL (ALL SUBJECT CODES WITH >1000 TITLES)	**94823**	**272**	**2.87**					

Table 7. Application of scaling factor to estimate number of subject categories required to map all of Cornell's holdings to HILCC (excerpt)

SUBJECT CODE	TITLE COUNT	SCALING FACTOR	TOTAL NUMBER OF SUBJECT CODES (SPLIT) -- ESTIMATE	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	CATEGORY5
1035.1	126156	2.87	362.07	Music, Dance, Drama & Film	Music			
1225.1	119708	2.87	343.56	Art, Architecture & Applied Arts	Fine Arts			
1267.1	112593	2.87	323.14	Languages & Literatures	East Asian & Ural-Altaic Languages & Literatures			
1536.1	103292	2.87	296.45	Sciences	Agriculture & Animal Sciences			
…	…	…	…	…	…	…	…	…
1503	1053	2.87	3.02	Law, Politics & Government	Military & Naval Science	Military Engineering		
1549.1	1019	2.87	2.92	Law, Politics & Government	Government (Non-U.S.)	Government (Canada)		
1258	998	2.87	1.00	Philosophy & Religion	Religion	North & South American Religions		
1221.11	976	2.87	1.00	Philosophy & Religion	Philosophy	Renaissance		
…	…	…	…	…	…	…	…	…
1532.1	135	2.87	1.00	Law, Politics & Government	Military & Naval Science	Space Warfare & Surveillance		
1100.2	119	2.87	1.00	Social Sciences	Recreation & Sports	Auto Travel & Racing		
1253.1	64	2.87	1.00	Social Sciences	Psychology	Psychotropic Drugs & Other Substances		
1009.6	30	2.87	1.00	History & Archaeology	Regions & Countries	United States	Local History	Territories, Protectorates, Etc.
TOTAL (ALL SUBJECT CODES)	**4462395**		**12772.91**					