Research Article

The Epic Struggle: Subject Retrieval from Large Bibliographic Databases

HELEN R. TIBBO

Abstract: Archivists have talked at length about the virtue of contributing records to a national bibliographic utility to provide enhanced access to collections. There has been little discussion, however, of the difficulties of finding materials in such large database environments. This article discusses a retrieval study that focused on collection-level archival records in the OCLC Online Union Catalog, made accessible through the EPIC search system. Data were also collected from the local OPAC at the University of North Carolina-Chapel Hill (UNC-CH) in which UNC-CH–produced OCLC records are loaded. The chief objective was to explore the retrieval environments in which a random sample of USMARC AMC records produced at UNC-Chapel Hill were found—specifically, to obtain a picture of the density of these databases in regard to each subject heading applied and, more generally, for each record. Key questions were (1) how many records would be retrieved for each subject heading attached to each of the records and (2) what was the nature of these subject headings vis-à-vis the number of hits associated with them. Findings show that large retrieval sets are a potential problem with national bibliographic utilities and that the local and national retrieval environments can vary greatly. The need for specificity in indexing is emphasized.

This article is based on a paper given at the Society of American Archivists' 1992 annual meeting in Montreal. OCLC supported this research. The author wishes to thank Patricia Haberkern, who did much of the searching.

About the author: *Helen R. Tibbo is presently an assistant professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill. She earned a B.A. in English from Bridgewater State College, an M.L.S. from Indiana University, an M.A. in American Studies from the University of Maryland, and a Ph.D. in Library and Information Science from Maryland as well. She teaches in the areas of reference, on-line information retrieval, and archival studies. Her primary research interests focus on optimizing information retrieval, particularly for information systems that support humanistic and archival research. She is a member of the Society of American Archivists, serving on its Editorial Board and as Chair of the Archival Educator's Roundtable, 1992–94.*

ARCHIVISTS[1] HAVE TALKED AT LENGTH about the virtue of contributing records to a national bibliographic utility such as the Online Computer Library Center (OCLC) or Research Libraries Information Network (RLIN) in order to enhance access to their collections.[2] There has been little discussion, however, of the difficulties of finding materials in such large database environments.[3] Ironically, electronic services such as OCLC and RLIN, which promise vastly improved access to archival collections on a nationwide or even international level over that possible in printed tools such as the *National Union Catalog of Manuscript Materials* (NUCMC), present enormous retrieval problems themselves.[4] As Lester Asheim has noted, "increasing the amount of information and speeding up access to it is more likely to result in information overload and entropy than it is to improve the receiver's ability to benefit from the information."[5]

The user's goal is to find all relevant material and nothing more.[6] As simple as this sounds, it is exceedingly difficult to accomplish, whether the retrieval system is word of mouth, printed format, or an electronic database. As systems grow in size, complexity, and power, they become more inclusive, but barriers to optimal retrieval effectiveness increase as well. This should not be surprising, as information retrieval power is never without its price. The larger and more heterogeneous the database, the more difficult it is to conduct subject or free-text searches effectively. Even known-item searches become slower and potentially more difficult as the search space increases.

Lancaster and his associates observe that the on-line catalog has not improved subject access but may have made the situation worse because it has led to the creation of much larger catalogs that represent the holdings of many libraries.[7] Merging several catalogs into one, when each component catalog provides inadequate subject

[1]*Archives* and *archivists* are being used herein for convenience to indicate both institutional archives and manuscript repositories and archivists and manuscript curators, respectively, unless otherwise noted.

[2]See for example David Bearman, "Archives and Manuscript Control with Bibliographic Utilities: Challenges and Opportunities," *American Archivist* 52 (Winter 1989): 26–39; David Bearman, *Toward National Information Systems for Archives and Manuscript Repositories: The National Information Systems Task Force (NISTF) Papers, 1981–1984* (Chicago, Ill.: Society of American Archivists, 1987); Elaine D. Engst, "Nationwide Access to Archival Information," *Documentation Newsletter* 10 (Spring 1984): 4–6; H. Thomas Hickerson, "Archival Information Exchange and the Role of Bibliographic Networks," *Library Trends* (Winter 1988): 553–71; H. Thomas Hickerson, "Expand Access to Archival Sources," *Reference Librarian* 13 (Fall 1985–Winter 1986): 195–99. James O'Toole has noted that "archivists fulfill only half their responsibility to make records available if they sit and wait for users to come to them. Instead, archivists must be active in *publicizing* their *holdings*. This responsibility implies the necessity of sharing information about what is in each archives," *Understanding Archives and Manuscripts* (Chicago, Ill.: Society of American Archivists, 1990), 67.

[3]Avra Michelson ("Description and Reference in the Age of Automation," *American Archivist* 50 [Spring 1987]: 192–203) has discussed the lack of consistency in archival descriptive practice, especially the assignment of subject headings for MARC AMC records and the implications for retrieval. Matthew Gilmore has noted that the requirement of most bibliographic information systems to include at least one LCSH term in each MARC AMC record "means that archivists frequently must use a very general heading rather than the specific local thesauri," resulting in those materials "disappearing into a void." "Increasing Access to Archival Records in Library Online Public Access Catalogs," *Library Trends* 36 (Winter 1988): 610–11.

[4]Library of Congress, *National Union Catalog of Manuscript Collections* (Washington, D.C.: Library of Congress, 1962–).

[5]Lester Asheim, "Ortega Revisited," *Library Quarterly* 52 (July 1982): 215.

[6]Although it can be argued that a user might only want a subset of all potentially relevant materials, that subset becomes all the items that are situationally relevant for that particular individual at that time.

[7]F. W. Lancaster, Tschera H. Connell, Nancy Bishop, and Sherry McCowan, "Identifying Barriers to Effective Subject Access in Library Catalogs," *Library Resources and Technical Services* 35 (October 1991): 388.

access, exacerbates the problem, since the larger the catalog, the more discriminating must be the subject access points provided. In recent years, catalogs have grown much larger without any significant compensatory increase in their discriminating power. Chandra Prabha of OCLC calls large retrievals "a problem of the 1990s."[8] She goes on to note that, "the problem of large retrievals is accentuated in an OPAC [online public access catalog] environment because a majority of users are occasional or casual users."[9] With 30 million records, the OCLC Online Union Catalog (OUC) clearly poses a challenging retrieval environment. Representing archival collections so as to optimize subject retrieval from a large bibliographic utility such as OCLC can truly be an "epic" struggle.

Regardless of the type of material represented—be it books, serials, or archival collections—document retrieval in large bibliographic databases depends on well-constructed document representations or surrogates. The semantic condensation required to represent a 350-page book or a 50-box collection in a catalog entry, or an abstract, or even an archival inventory demands that more is left unsaid than recorded in these surrogates. In the process of semantic condensation, information is necessarily lost. This loss may seem unfortunate, but the remaining distillation, when well selected, becomes a more powerful retrieval tool than the full text of the original. A "good" surrogate eliminates "noisy" information that is found in all full texts and could cause an item to be retrieved when it should not be; a good surrogate also includes information that will facilitate its retrieval in response to appropriate queries.

It is the processor's job to create a surrogate, be it an archival finding aid or a USMARC AMC (Machine Readable Cataloging, Archives and Manuscript Control) record, that captures the most important material in the item represented in as succinct and specific a manner as possible. Of increasing importance in extremely large databases, the surrogate must not merely represent its parent document and/or collection, it must be able to distinguish it from a multitude of other very similar items.

The most subjective elements of MARC AMC records in bibliographic databases, yet certainly some of the most important regarding access, are the subject fields. Many of the other fields, such as collection title, extent, or location, are relatively straightforward.[10] Collection titles can provide some manner of subject access, but for most researchers who want to find collections that contain materials related to a particular topic, a search of the 12 subject fields in a MARC AMC record will be appropriate.[11]

This article discusses a retrieval study that focused on collection-level archival records in the OCLC Online Union Catalog, made accessible through the EPIC search system. I also collected retrieval data from the local OPAC at the University of North Carolina-Chapel Hill (UNC-CH) in which OCLC records produced by UNC-CH are loaded. The chief objective was to explore the retrieval environments in which a random sample of MARC AMC records pro-

[8]Chandra Prabha, "Managing Large Retrievals: A Problem of the 1990s?" in *OPACs and Beyond*, Proceedings of a Joint Meeting of the British Library, DBMIST, and OCLC, OCLC Online Computer Library Center, Inc., Dublin, Ohio, August 17–18, 1988 (Dublin, Ohio: OCLC, 1989), 33.

[9]Prabha, "Managing Large Retrievals," 33–34.

[10]Even with these fields there can be serious retrieval problems, as when institutions just use "Papers" as the full title for a collection.

[11]For a detailed description of these fields, see Harriet Ostroff, "Subject Access to Archival and Manuscript Materials," *American Archivist* 53 (Winter 1990): 100–05. See also Online Computer Library Center, *Archives and Manuscript Control Format*, 2nd ed. with updates (Dublin, Ohio: OCLC, 1986).

duced at UNC-Chapel Hill were found—specifically, to obtain a picture of the density of these databases in regard to each subject heading applied and, more generally, for each record. Key questions were (1) how many records would be retrieved for each of the subject headings attached to each of the records and (2) what was the nature of these subject headings vis-à-vis the number of hits associated with them? I was particularly interested in seeing if the subject headings used at UNC-CH incurred an overwhelming number of postings in the national database and how this related to the number found in the UNC-CH OPAC. I also wanted to compare the number of postings for topical headings and personal names. This type of information is important in assessing how well a database is serving the research community because catalog persistence studies indicate that researchers, even in university settings, rarely are willing to look through hundreds of items in a catalog. Summarizing earlier OPAC studies, Ray Larson notes that users of on-line catalogs frequently find too many items or none at all.[12] If subject headings applied to MARC AMC records incur hundreds of hits in OCLC, even if they work well in the contributing institution's local catalog, it is doubtful that researchers will find the records in the larger national bibliographic environment. To optimize the archival community's investment in providing national access to materials, archivists must explore these large retrieval environments and adjust cataloging and retrieval techniques appropriately.

The EPIC Service

OCLC's EPIC service is a commercially available interactive on-line searching service that provides access to several large databases.[13] The database with which archivists are most concerned is the OCLC Online Union Catalog. If an archives sends MARC AMC records to OCLC, this is the database in which the records will appear. Currently, this database contains well over 30 million records representing information sources in a wide variety of materials and languages. It is growing at a rate of 2 million records per year, or 40,000 records per week. This is OCLC's original database, which library catalogers and interlibrary loan librarians have used for over 20 years for cooperative cataloging and for locating known items for interlibrary loan. The Library of Congress sends an average of 5,000 records per week to OCLC, with other OCLC member libraries contributing about 34,000.

Until the advent of the EPIC search service in 1990 and, more recently, FirstSearch,[14] OCLC provided a search interface designed specifically for catalogers. The classic OCLC search protocol relies on the searcher having a book or other material in hand so that the author, title, publisher, and publishing date are known. The searcher enters parts of the title and the author's name so as to locate any existing cataloging records for that particular item. The system then retrieves any records that match the given known-item specifications. While the cataloger may have sev-

[12]Ray R. Larson, "Managing Information Overload in Online Catalog Subject Searching," *Proceedings of the ASIS Annual Meeting, 1989* (Medford, N.J.: Learned Information, 1989), 129.

[13]For more about EPIC, see Nita Dean, "EPIC: A New Frame of Reference for the OCLC Database," *OCLC Newsletter* (March–April 1991): 21; "The EPIC Service Is Introduced," *OCLC Newsletter* (January–February 1990): 10–16; and Laurie Whitcomb, "OCLC'S EPIC System Offers a New Way to Search the OCLC Database," *Online* 14 (January 1990): 45–50.

[14]According to OCLC, "FirstSearch is an interactive searching system for library patrons" that allows them "to search a variety of bibliographic databases. . . . By following on-screen instructions, patrons can search successfully without special training." Online Computer Library Center, Inc., *The FirstSearch Catalog* (Dublin, Ohio: OCLC, 1992), 1.

eral variant records to look through, they will all be for the particular title in hand (different editions perhaps), or, if only author information is entered, they will all represent works by that individual. Despite the size of the database, a searcher can very quickly locate items via this system because all searches are based on specific, concrete information such as titles, authors' names, and International Standard Book Numbers. The OCLC Online Union Catalog has always held subject information in the form of subject headings (usually Library of Congress Subject Headings [LCSH]) for each record, but it was not until the development of the EPIC service that OCLC provided a means by which to do subject searching, thus using these existing access points.

The EPIC search service complements the original OCLC search engine by providing keyword, phrase, and subject searching. A searcher can use Boolean, proximity, and range searching features as well as truncation and index scanning.[15] The EPIC command interface, the search language, is based on the NISO *Common Command Language for Interactive Information Retrieval (Z39.58)*. The EPIC search interface is extremely powerful, but this does not mean that users will easily be able to produce good searches. The more simplistic FirstSearch system designed specifically for end users also presents serious retrieval problems because the main problems lie not with the searching front ends but with the OCLC OUC database itself. While this enormous database works extremely well for cataloging and interlibrary loan, where the searcher has a specific title or author in mind, it is a relatively unexplored morass for subject searching. The most evident problems revolve around the size of the database and the use of broad, precoordinated Library of Congress Subject Headings for postcoordinate retrieval. These problems are not restricted to archival searching and MARC AMC records; indeed, producing manageable and complete subject search results for monographs in such a system is potentially even more difficult.

In an effort to adapt LCSH terms for electronic retrieval, OCLC takes each subject heading assigned to a book, archival collection, or other material and breaks it apart. This is very useful as it eliminates the need for users to construct lengthy LCSH strings in order to do subject searches and allows more flexible searching.[16] To retrieve items assigned the heading "North Carolina—History—Civil War, 1861–1865—Personal Narratives, Confederate," a searcher would enter a statement with the following elements in any order connected by the Boolean *and*: *find su=(North Carolina and History and Civil War 1861–1865 and Personal Narratives Confederate)*. The *su=* tells EPIC to look only through subject headings but does not limit retrieval to only records with this particular subject heading string. For example, an item with the following combination of subject headings would also be retrieved: "United States—History—Civil War, 1861–1865—Personal Narratives, Confederate" and "North Carolina—Description and Travel *and* 19th Century." Unfortunately, there is no mechanism by which the searcher can just receive items with a particular subject heading string, nor can the searcher browse complete subject strings in the scan mode and see how many items are posted to each.

[15]Index scanning does not work well with the subject fields, as the subject strings, common to LCSH, are broken into constituent parts and do not appear in any scannable index as complete strings.

[16]Many individual library OPACs require users to enter full LCSH strings with correct syntax in order to retrieve items on a topic.

Information Overload

In general, the two primary purposes of subject control are (1) to allow the user to find material on a subject, and (2) to collocate a repository's materials on a subject at one point in the catalog, thus giving the user a summary of what is contained in that collection on the given topic. National union catalogs, such as OCLC's OUC, go one step further. Because the OCLC Union Catalog is a national database that employs LCSH, it collocates topical materials from around the country at each subject heading. Richard Smiraglia further notes that "when LCSH is used to supply subject headings for AMC formatted records, the archival materials will collocate with published materials on the same topic in an integrated bibliographic system (network or local), thus giving a user an opportunity to browse bibliographic records for both published works and primary source material under a topical heading."[17]

While this is theoretically a wonderful research opportunity that might well bring new researchers into archival repositories because they find archival materials next to books in the catalog, such collocation works best, or perhaps only works at all, with relatively small collections. OCLC's Union Catalog, with over 30 million records, hardly fits into the "relatively small" category. If 15,000 records collocate at a subject heading, or Boolean combination of terms—not an unheard-of retrieval in EPIC—the chance that the researcher will view any one of the records is greatly diminished; indeed, it becomes a chance event dependent on when the search is done, when the record was entered into the database, and how users deal with information overload.

Researchers are not without resources to deal with information overload. Joel and Mary Jo Rudd list several ways in which library users turn a potential information overload into a manageable load.[18] They explain that in addition to using Herbert Simon's principle of "satisficing" (acquiring a "satisfactory" subset of available information), researchers faced with cognitive and temporal limitations on information acquisition frequently just "skim off the top," looking only at the first few items they find in a catalog or on the shelves. Because most bibliographic databases present retrieval sets in last-in, first-out (LIFO) order, any given record collocated at a subject heading may fall victim to the "Andy Warhol" phenomenon, wherein each record is famous for its 15 minutes until it sinks into the morass of the database as newer records pile on top of it. The problem here, of course, is that the most appropriate records, particularly in fields such as history, where information does not go out of date quickly, may be at the bottom of the pile. Indexing consistency becomes important for only the most comprehensive searches and tenacious database searchers, but distinction drawn among items comes to the fore. Ortega y Gasset's 1934 definition of a librarian as "a filter interposed between man and the torrent of books" can now apply to the archivist and the on-line catalog or on-line bibliographic systems.[19]

Stephen E. Wiberley, Jr., Robert A. Daugherty, and James A. Danowski conducted a "users' persistence" study in

[17]Richard P. Smiraglia, "Subject Access to Archival Materials Using LCSH," in *Describing Archival Materials: The Use of the Marc AMC Format,* edited by Richard P. Smiraglia (New York: Haworth Press, 1990), 64.

[18]Joel Rudd and Mary Jo Rudd, "Coping with Information Load: User Strategies and Implications for Librarians," *College and Research Libraries* 47 (May 1986): 315–22.

[19]Jose Ortega y Gasset, *The Mission of the Librarian,* translated by James Lewis and Ray Carpenter (Boston: G.K. Hall, 1961).

1987.[20] They looked for what David Blair calls the "anticipated futility point."[21] Blair defines this as the number of documents a researcher will be willing to begin to browse through. Karen Markey has called this user "perseverance."[22] Wiberley and his colleagues adapted Blair's definition to the number of references in an on-line catalog that users were willing to scan in discretionary information-seeking situations. Subject searching fits into this discretionary type of information seeking in that the user never knows the extent of information available and thus feels no compulsion to search out a particular fact or title. Wiberley, Daugherty, and Danowski studied user persistence or perseverance with an academic library OPAC that contained more than 425,000 records. They studied user transaction logs and questionnaires. The median response to the question "How many postings would you consider to be too many?" was fifteen. The transaction log data indicated a sharp drop-off in persistence with more than 30 postings and a great drop-off after sixty. More specifically, they found that while a majority of users "displays all general records for searches that retrieve between eleven and thirty postings, when searches retrieve more than thirty postings, a majority of users displays no records."[23]

OPAC users, such as those in the Wiberley, Daugherty, and Danowski study, may tolerate fewer citations than on-line–search service clients, who may turn to commercial on-line databases only when they want an exhaustive search. The searching literature and vendors such as DIALOG Information Services generally hold that very few on-line-search clients are willing to look through more than 100 citations, with many people willing to scan only 50 or fewer items. This information holds serious implications for archival researchers using on-line databases such as OCLC's OUC and locally or Internet-available library catalogs. To understand how best to represent documents or collections of materials in these contexts, we need first to explore these retrieval environments.

Methodology

In February 1992 I selected a random sample of 60 MARC AMC records representing collections held in UNC-CH's Southern Historical Collection from the OCLC Online Union Catalog. A graduate assistant searched the subject headings attached to each of these records in OCLC as well as in the university's on-line catalog in March 1992. For example, "Merchants—North Carolina—History—19th century" retrieved 67 items in the UNC-CH on-line catalog and 106 items in the OCLC OUC in March 1992. In August 1992 and June 1993 I again searched all headings in the on-line catalog, the entire OCLC database, and the manuscripts portion of the OCLC database. In comparing the data I discovered that because one record was such an outlier it distorted the picture for the mean number of hits per search term and per record. In this case, one heading—Sermons—received 54,904 hits in OCLC in August 1992. I eliminated this record from the sample, thus bringing the usable population to fifty-nine. I also discovered that the graduate student had

[20]Stephen E. Wiberley, Jr., Robert A. Daugherty, and James A. Danowski, "User Persistence in Scanning Postings of a Computer-Driven Information System: LCS," *Library and Information Science Research* 12 (October–December 1990): 341–53. See also Stephen E. Wiberley, Jr., and Robert A. Daugherty, "Users' Persistence in Scanning Lists of References," *College and Research Libraries* 49 (March 1988): 149–56.

[21]David C. Blair, "Searching Biases in Large Interactive Document Retrieval Systems," *Journal of the American Society for Information Science* 31 (July 1980): 271.

[22]Karen Markey, *Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs* (Dublin, Ohio: OCLC, 1984), 67–71.

[23]Wiberley, Daugherty, and Danowski, "User Persistence," 352.

Table 1. Mean Number of Postings per Term

EPIC—June 1993	229
EPIC—August 1992	207
EPIC—March 1992	196
EPIC/mss—June 1993	67
EPIC/mss—August 1992	59
Local Total—June 1993	42
Local Total—August 1992	39
Local Specific—June 1993	29
Local Specific—March 1992	20

Table 2. Mean Number of Postings per Term per Record

EPIC—June 1993	252
EPIC—August 1992	235
EPIC—March 1992	220
EPIC/mss—June 1993	69
EPIC/mss—August 1992	63
Local Total—June 1993	45
Local Total—August 1992	41
Local Specific—June 1993	29
Local Specific—March 1992	20

Table 3. Median Number of Postings per Term

EPIC—June 1993	101
EPIC—August 1992	93
EPIC—March 1992	79
EPIC/mss—June 1993	46
EPIC/mss—August 1992	43
Local Total—June 1993	26
Local Total—August 1992	24
Local Specific—June 1993	21
Local Specific—March 1992	14

Table 4. Median Number of Postings per Term per Record

EPIC—June 1993	128
EPIC—August 1992	120
EPIC—March 1992	105
EPIC/mss—June 1993	44
EPIC/mss—August 1992	40
Local Total—June 1993	27
Local Total—August 1992	24
Local Specific—June 1993	21
Local Specific—March 1992	16

searched the local records in a different manner, so that the local data for March 1992 are not able to be compared to the August 1992 results but are comparable to one set of the June 1993 findings.

Findings

Table 1 shows the mean number of hits or postings for the 519 subject headings associated with the 59 records. The mean number of subject headings per record was 8.8, with a median of 8.0. The first EPIC search retrieved an average of 196 postings per heading. Only five months later this number rose by 11 points, and nine months after that it went up another 22 points. Keeping Wiberley, Daugherty, and Danowski's findings in mind, these results should be alarming.

Even when the manuscript records are separated from the other materials in OCLC, the average retrieval was 67. News is better for the local catalog, with an average of 42 hits in June 1993 and 39 in August 1992. These numbers represent the total figure given for an entry such as "Virginia—Civil War." The UNC-CH catalog provides this figure for this term and all subdivisions, such as "Correspondence" or "Stores and supplies" before listing any brief titles on the screen. The searcher in March had gone to the second step of looking in the index—the actual list of subject headings used in the catalog—and had recorded the number of items specifically attached to the broader term (e.g., "Virginia—History—Civil War, 1861–1865"), but she did not include figures for any of the subtotals. Thus, the March figure, which took more searching expertise to derive, is as conservative as possible and is still twenty. This number rose to 29 by June 1993. Table 2 provides data on the

Table 5. Postings per Record, EPIC, June 1993

9	152	525	841	1,497	4,589
16	179	536	866	1,565	4,723
20	198	552	966	1,808	4,778
20	257	646	971	1,813	5,749
46	299	651	974	1,846	6,320
67	355	652	1,155	1,867	6,453
80	360	658	1,206	2,877	9,454
110	406	668	1,306	3,161	13,622
125	422	735	1,357	3,728	16,666
150	500	810	1,482	3,918	—

Table 6. Mean Number of Posting per Record, EPIC, June 1993

2.25	25.00	52.50	129.14	186.57	338.91
4.00	27.18	52.75	130.40	194.20	351.22
5.00	31.25	67.50	131.60	205.10	489.75
5.30	33.83	76.57	144.33	243.50	526.67
7.67	41.67	83.17	147.00	259.00	567.58
15.20	42.83	96.60	150.75	265.44	668.00
15.23	44.75	107.67	156.50	302.58	859.46
15.71	46.72	116.69	162.75	314.87	1,290.60
16.00	50.18	123.50	169.63	319.67	4,166.50
16.75	51.43	128.33	177.50	327.79	—

Table 7. Posting per Record, Local Total, June 1993

6	47	141	217	352	955
9	78	148	224	406	961
10	84	151	254	425	1,022
14	85	154	296	450	1,097
17	97	161	299	465	1,122
22	106	164	311	496	1,133
23	108	178	344	532	1,207
30	114	178	346	699	1,233
35	122	189	348	762	1,700
38	124	207	348	829	—

mean number of postings per term per record. The results are even worse with this method of calculation.

The median number of postings per term (table 3) and per term per record (table 4) may represent a more accurate picture of the data due to a few extremely heavily posted terms that distorted the means. Although it contains lower figures, table 3 shows a 22-point increase in the OCLC figures over the 14-month period.

Enumerations of the number of postings per record and the mean number of postings per term per record show the range in postings (tables 5 through 8).

Table 9, showing the greatest number of hits per term, indicates how useless a subject heading can become in a database of 30 million records. "United States—History—Revolution—1775–1783" retrieved 16,393 items in the June 1993 complete OCLC search and 1,628 items from the

Table 8. Mean Number of Postings per Item per Record, Local Total, June 1993

1.50	7.60	16.44	27.18	55.11	76.00
1.80	8.31	18.50	30.20	56.25	81.20
2.50	8.40	19.00	32.00	57.33	86.50
4.25	9.64	19.50	34.56	59.63	86.82
4.67	11.50	20.33	35.42	60.94	92.11
4.70	12.00	20.67	36.29	63.50	109.73
5.00	12.13	22.00	38.83	68.13	120.13
5.00	13.50	24.86	40.25	69.60	224.20
5.75	14.83	25.43	46.50	70.40	425.00
7.08	15.40	27.13	51.38	70.50	—

Table 9. Greatest Number of Postings Per Term

EPIC—June 1993	16,393*
EPIC—August 1992	15,641*
EPIC—March 1992	15,001*
EPIC/mss—June 1993	2,213**
EPIC/mss—August 1992	1,797**
Local Total—June 1993	1,628*
Local Total—August 1992	1,438*
Local Specific—June 1993	1,012†
Local Specific—March 1992	962†

*United States—History—Revolution—1775–1783
**North Carolina—History
†World War, 1914–1918—France

Table 10. Terms with Only 1 Posting

EPIC—June 1993	132	25%
EPIC—August 1992	144	28
EPIC—March 1992	145	28
EPIC/mss—June 1993	154	30
EPIC/mss—August 1992	167	32
Local Total—June 1993	181	35
Local Total—August 1992	190	37
Local Specific—June 1993	189	36
Local Specific—March 1992	241	46

Table 11. Mean Number of Postings per Multiple-Hit Terms

EPIC—June 1993	307
EPIC—August 1992	286
EPIC—March 1992	272
EPIC/mss—June 1993	95
EPIC/mss—August 1992	87
Local Total—June 1993	64
Local Total—August 1992	60
Local Specific—June 1993	45
Local Specific—March 1992	37

UNC-CH catalog. "North Carolina—History" retrieved 2,213 just from the manuscripts in OCLC.

The number of headings incurring only one hit—the vast majority of these being personal names—indicates that the remaining topical subject headings received more postings on average than shown above (see table 10). The picture becomes bleaker and bleaker for the use of topical subject headings in such a large database, especially when we realize that many of the headings analyzed are used predominately by archivists, and archivists have been contributing to OCLC only for a few years! Even the UNC-CH catalog now averages over 60 postings for the multiple-hit headings (table 11). These figures would be even higher if subject headings with two and three hits (still mostly names) were added to those with just one.

Table 12 shows the number of hits on individuals' names and the average number of postings on the names as a whole. There were 108 individuals' names included as subject access points in these records. In comparison, entries for 29 families, such as "Rogers," "Smith," and "Erwin,"

Table 12. Total and Average Number of Postings for Individuals' Names

Date of Search	Number of Postings	Average Number of Postings
EPIC—June 1993	450	4.2
EPIC—August 1992	432	4.0
EPIC—March 1992	408	3.8
EPIC/mss—June 1993	189	1.8
EPIC/mss—August 1992	186	1.7
Local Total—June 1993	225	2.1
Local Total—August 1992	214	2.0
Local Specific—June 1993	200	1.9
Local Specific—March 1992	183	1.7

yielded many more postings, particularly in the OCLC OUC (table 13).

Discussion

What do these data tell archivists, who both create records for national databases and help researchers locate materials in them? First and foremost, it is important to realize that what may work locally will not necessarily work in a 30-million-record database. This is not to say that the use of such databases for national access is not a good idea. Rather, archivists have to understand the nature of the environments into which they are sending their records and do all they can to help them compete. In database terms this means providing access points that will help the records to be retrieved and read when they are relevant items. Both local and national concerns must be balanced. A "good" record is a little bit like the proverbial good child: It should speak only when spoken to—that is, present itself for retrieval when it is relevant to a researcher's needs, but otherwise be silent. As with children, it is often difficult to make bibliographic records behave.

To extend the analogy, most child experts will tell you that environment, as well as genetics or specific parental teachings, plays a role in how children behave. Such is the case with bibliographic records. A bibliographic record that does not use standardized subject access terms may never be found in a national database. Such practice will lead to low-recall searches. At the same time, a seemingly excellent record with standardized subject headings that represents a collection very well may find itself buried in other seemingly excellent records if there is much material on that topic in a large database. In this scenario, document discrimination and search precision become overriding concerns. The record and its creator must adapt to this environment or risk oblivion. The same record may work well "at home," where there are relatively few items on this topic in the on-line catalog. Conversely, the local catalog may require augmented local subject headings that make sense in that environment. Not only must archivists consider collection and user characteristics in providing subject access, they must also consider the environment into which the records will be sent. This may mean sending one record off to a national utility while placing another record, perhaps with local subject headings and location information, in the home OPAC. There is no reason, other than additional processing costs, why the two records must be identical.

Representing Archival Collections

Most important in making records behave

Table 13. Total and Average Number of Postings for Family Names

Date of Search	Number of Postings	Average Number of Postings
EPIC—June 1993	2,963	102.0
EPIC—August 1992	2,684	92.5
EPIC—March 1992	2,621	90.4
EPIC/mss—June 1993	590	20.4
EPIC/mss—August 1992	277	9.6
Local Total—June 1993	165	5.7
Local Total—August 1992	149	5.1
Local Specific—June 1993	139	4.8
Local Specific—March 1992	112	3.9

in any bibliographic environment is the archivist's responsibility for capturing the key concepts of the materials in their finding aids. As David Bearman argues, consistency of topical headings is not so important if we provide very good searching tools such as switching vocabularies and "intelligent" front ends (and that is a big "if").[24] Selecting and representing key concepts are highly subjective and difficult tasks, and those selected will not always fit the needs and visions of future users. This work will never be scientific, but it will always be important, just as archival processing has been important in the past. The great service here is to reduce the bulk of information to be searched in a meaningful and rational manner, keeping in mind that it is better to do this work now than to wait for perfection that will never come. Representing materials completely and succinctly, while differentiating them from a multitude of similar documents, lies at the heart of any information storage and retrieval system. As with the MARC AMC format, archivists now need to focus on the types of subjects to be documented. They need to build a subject access framework to identify what subjects in archival collections should be represented in subject indexing, as Bearman and others have pointed out.[25]

Beyond ensuring that the truly significant material in a collection is represented, appropriate indexing language is central to creating good bibliographic records. Regardless of the item, be it an entire collection or a series, the specificity and exhaustivity of the indexing language are important. If these elements are appropriate to the material being represented, some subject-indexing consistency should follow, with strict authority control being left to more specific forms of information, such as personal, corporate, and geopolitical names, as well as collection forms and functions. Database producers have long recognized the importance of appropriate indexing languages for their materials. Thus, databases such as MEDLINE, ERIC, and Psychological Abstracts all have their own controlled vocabularies and thesauri.

[24]David Bearman, "Authority Control Issues and Prospects," *American Archivist* 52 (Summer 1989): 288.

[25]Bearman, "Authority Control," 286–99; Helena Zinkham, Patricia D. Cloud, and Hope Mayo, "Providing Access by Form of Material, Genre, and Physical Characteristics: Benefits and Techniques," *American Archivist* 52 (Summer 1989): 300–19. Helen Tibbo extends this notion to a framework for abstracting in *Abstracting, Information Retrieval and the Humanities: Providing Access to Historical Literature* (Chicago: American Library Association, 1993).

Broad, undifferentiated topical headings, common to LCSH, do not appear to work well for retrieval from large electronic databases. If repositories collecting in similar areas work together on authority lists, appropriate index terms, and user thesauri, these efforts could increase the consistency with which the institutions with key collections represent their materials without sacrificing necessary specificity to a monolithic indexing language. This would also allow archivists to retain much of their "rugged individualism"[26] while cooperating with related institutions. Archivists could then coordinate and disseminate such vocabularies nationally.

Avra Michelson sent a common description of an archival collection to several repositories and discovered a total lack of consistency in descriptive practice, especially in the assignment of subject headings.[27] While no conclusions regarding indexing consistency can be drawn from the present study, it is clear that archivists from across the United States are applying the same subject headings, even quite lengthy and complicated LCSH strings, to hundreds and thousands of records. In many cases they use terms that librarians also select quite frequently. We do not know if archivists are consistently applying these terms to the same concepts, but we do know that large numbers of postings are accruing at certain topical headings, even when these are delimited by geographic location and date. Because archivists in different institutions never index the same collections, more context-sensitive studies of indexing consistency may be necessary if we are to judge accurately the extent of indexing consistency. What is clear is that better document discrimination, possibly through more specific, appropriate, and exhaustive indexing languages, is necessary as databases continue to grow.

Another representation issue is determining the best archival level on which to base MARC AMC records. It is important to recognize that these records serve to describe collections in only a minimal fashion. The primary function of MARC AMC records is to lead the searcher to a finding aid, which in turn documents and describes in detail the collection and parts thereof.[28] As such, MARC AMC records cannot fully describe a collection, nor should they. Having said this, I should add that it is probably best to provide collection-level access in MARC AMC records, as the introductory information in an inventory serves as an umbrella for the series and folder descriptions. Certain situations, however, can make the creation of just collection-level records arbitrary. A particular series, or even an individual item, may outweigh the value of the rest of the collection. If this is the case, and if the general terms that best describe the collection as a whole do not provide optimal specificity for the important part of the collection, a separate MARC AMC record would help to facilitate access. Such a record, however, would have to lead the researcher to the collection-level record or provide enough provenancial context so that the researcher could locate the collection.

In OCLC, with a limited number of subject headings per record, the archivist frequently cannot assign enough headings to index appropriately both the entire collection and its significant parts. In RLIN,

[26]Janet Gertz and Leon J. Stout, "The MARC Archival and Manuscripts Control (AMC) Format: A New Direction in Cataloging," *Cataloging and Classification Quarterly* 9 (1989): 5.

[27]Michelson, "Description and Reference."

[28]In *Archives, Personal Papers, and Manuscripts*, 2nd ed. (Chicago, Ill.: Society of American Archivists, 1989), Steve Hensen notes that "The chief source of information for archival materials is the finding aid prepared for those materials" (p. 9) and not the materials themselves unless there is no finding aid or provenance or accession records. Thus, the cataloging record is derived from the finding aid as the finding aid is based on the collection.

where any number of headings can be used, excessively long lists and long records can discourage researchers from looking at the items they retrieve. In both cases, separate records that provide access to the collection as a whole and to specific parts would provide better access to this material than does one inadequate or overly long record. If different names, organizations, or institutions are prominent in various series of a collection it may be a good idea to make linked MARC AMC records for each relevant series and to index these with the prominent names. Subject access common to all series in a collection should be kept with the main record so as not to replicate the topical headings for the collection several times within the database. This is not to say that all series, folders, or items need to be represented just because a few are deemed to be important. What might appear to be uneven representation of the collection in terms of a finding aid could provide optimal access to key elements. In this way, cataloging and access become intricately tied to appraisal. It is important to remember that in a database all records, whether they represent important or relatively insignificant materials, become equal in the retrieval game. Responsible appraisal of what should be represented in the database becomes a powerful retrieval tool.

As Bearman notes, a large number of subject headings per record gives that record a better chance to be retrieved.[29] When repeated by everyone, the practice of applying more and more subject headings will serve primarily to increase the size of the database and will result in overwhelmingly large retrieval sets and long records. This is already the case in RLIN, where records may go on for 12 screens and have over 200 subject headings.[30] The best policy is to select important material and represent it accurately and precisely. As with appraisal, selection is critical. It is irresponsible to "pollute" a retrieval environment with extraneous or repetitive postings to terms just to increase the likelihood that a given record will be retrieved. We do not want to clog up our databases any more than our shelving or backlog areas.

Retrieving Archival Materials Reference archivists must become expert searchers of national bibliographic systems and on-line catalogs that are available on the Internet if they are to provide their clients with the highest possible level of service. Since both OCLC and RLIN must be employed for comprehensive searches, and since many archival researchers want high-recall searches,[31] archivists must become well versed in both systems. This means becoming familiar with the searching languages and capabilities and, more important, with basic information-retrieval principles and strategies. Today's electronic information-retrieval systems are deceptively easy to use, so much so that even the novice searcher can find something on most topics. At the same time, it is often very difficult to do a good search that optimizes recall and precision. This is particularly true in large databases. Archivists must be prepared to do searches for clients and to assist clients in conducting their own searches. Indeed, there is a large role for user education, particularly with CD-ROM products and Internet-available on-line catalogs. Searching guides and instructional classes will become necessary if clients are to do their own searching.

[29]Bearman, "Authority Control," 289.

[30]Kathleen Roe discussed the problems related to lengthy RLIN records at the 1992 SAA Annual Meeting in Montreal in a paper titled, "Autonomy vs. Community: Life in an Archives Database Commune."

[31]Mary Jo Pugh, "The Illusion of Omniscience: Subject Access and the Reference Archivist," *American Archivist* 45 (Winter 1982): 33–44.

Archivists must not only learn how best to apply subject headings, they must also turn this knowledge into searching expertise. Librarians are coming to see the difficulty of using a precoordinate indexing language, such as LCSH, for postcoordinate retrieval, and hopefully there will be significant improvements in future LCSH versions in the age of OPACs. OCLC has recognized the precoordinate problem, and thus breaks each LCSH heading and subheading apart to facilitate more flexible retrieval. Because of this, the searcher does not have to worry about matching the syntax of lengthy LCSH strings in EPIC, although this may still be the case with on-line catalogs available locally and on the Internet. Reference archivists must become skilled in searching all of these tools. They must know how to construct LCSH strings for searching OPACs and realize that the breadth of many LCSH terms, even when combined with other terms and delimiters in the OCLC or RLIN OUC, may prohibit precise retrieval.

When searching individual OPACs via the Internet, it is important to remember Avra Michelson's study. Archivists tend to use different terms (even when restricted to a controlled vocabulary) to describe the same things. Thus, when searching someone else's catalog, we should remember that it is important to use a number of synonomous search terms to ensure high recall (if that is the objective). It is always easier to search our own catalog wherein we know the terms local staff members tend to use over and over. It would be a great service to the field if institutions with like collections cooperated in building "common term" lists and then made these available to other institutions and clients, complete with examples on how to make searches as specific as possible. These could even be mounted on Internet gopher servers for easy access.

Headings divided by geographic and temporal elements—facets found to be important to historians' information-seeking methodologies—work well to distinguish items that are topically related.[32] Jackie Dooley also notes the importance of space and time delimiters for providing more refined subject access.[33] Such delimiters, however, provide only a partial answer. As can be seen from examples given above, even when subject headings contain locales and date ranges, a large number of records may be retrieved, and thus the actual topical subject terms must also be specific. Conversely, many items may be omitted from date- or place-restricted retrieval sets if processors failed to include all possible specific delimiters and subheadings. When a collection covers several geographical areas and years, processors may be forced to use broader terms because they are restricted in the number of more specific designations they can make. Reference archivists should advise clients searching OCLC or similar databases to use geographical and temporal elements in search strategies, but clients should also be aware that many relevant records will not be retrieved with these limitations. Processors must assign geographic and temporal subheadings to LCSH when these notions are central to the collection being represented, and reference archivists must explain the realities and limitations of database searching to clients.

If only primary materials are desired, limiting a search to the manuscripts segment of the OCLC database seems a good strategy to limit set size. Examination of records in the larger OCLC sets, however, reveals that many archival materials have been entered in the MARC book format. Thus, searches restricted to manuscripts will not retrieve all relevant items. Furthermore, such searches will not collocate

[32]Tibbo, *Abstracting, Information Retrieval, and the Humanities.*

[33]Jackie Dooley, "Subject Indexing in Context," *American Archivist* 55 (Spring 1992): 348.

published and unpublished sources, which may be what the researcher wants. This strategy should be used quite carefully and explained to the client.

Subdivision by form is a useful retrieval strategy, but headings such as "Sermons" or "Diaries" by themselves get lost in the shuffle. It is very important in large databases to combine form headings with other topical, temporal, or geopolitical headings. Along this line, headings such as "Brown Family," while they may work in our local catalogs where there is only one Brown family, produce quite undesirable results in a national catalog. Ideally, each subject heading is supposed to denote only one concept. Although there may be linkages among the over 200 records in OCLC with the heading "Brown Family," in many cases individual families that are in no way related are represented. This indicates a total lack of authority control and results in excessive postings because separate concepts (different families) are represented by the same term.[34] Searchers should usually try to limit queries with family names to particular geographic locations.

Entry of specific personal or corporate names, which can be expected to have very few hits even in large databases, seems to be one way to provide specific access without running the risk of unwieldy retrieval sets. Without time-consuming name authority work, however, names may provide only partial access to relevant materials. Fortunately, searchers may be able to overcome many variations in names with truncation and other search tactics, but total pseudonyms will remain invisible to a searcher unless a link is made in the database.[35] The primary drawback to retrieval by personal name is that the researcher must know the key players in the area being studied before finding the material. While names and institutions provide a type of subject access, they augment rather than replace topical access points.

The Future At this time, we just do not know enough about how researchers attempt to look for archival materials in national databases or in local on-line catalogs. This information should drive the design of our information systems and our document representations. In its absence, the cardinal rule of indexing—"Index at the most specific level possible"—should always apply, but this edict is often ambiguous. Even more problematic is the searcher's analog: "Search at the most specific level possible." Richard Pearce-Moses raised valuable questions in this regard in a posting to the ARCHIVES listserv in December 1992:

> Fixing up LCSH and MARC may be the last steps we should be worrying about. Maybe we need to define some common research strategies based on patron needs: What are patrons asking of our materials? and What tools do we need to match our material to those requests?[36]

In addition to user studies, much more research into the nature of retrieval from large bibliographic databases is needed. This work would benefit all players in the information community, as most databases

[34]There are two ways in which authority control (i.e., use of a control vocabulary) can be violated: (1) the same concept can be represented by different terms, and (2) different concepts can be represented by the same term. The former case is most often considered, but the latter may be more difficult to overcome from a retrieval point of view, particularly when large numbers of records are retrieved.

[35]Actually, a sophisticated search system would be able to retrieve pseudonyms of any name entered without the searcher ever being worried with the matter, if so programmed. This is not the reality of major search services today and the upkeep cost of such a service makes it unlikely in the near future.

[36]Richard Pearce-Moses, "LCSH—Summation and Opinions—Sources—1992," ARCHIVES listserv (15 December 1992).

are growing at an alarming rate. Retrieval studies comparing OCLC, RLIN, and Internet-available OPACs are also needed. Because the RLIN record structure is more felicitous to archival information, most archivists believe it is the information system of choice for archival materials. Only research will substantiate this belief. If researchers know which repositories hold the materials they want, searching individual catalogs via the Internet may produce the best retrieval results once users deal with all the OPAC search variations. This approach is the electronic equivalent to writing individual archivists to see what their collections hold in a given area. Many interesting studies wait to be conducted.

In this day of information gluttony and those surfeited years that surely lie ahead, responsible appraisal and provision of access to significant materials are central to the archivist's function. We know we cannot save everything. Now we must learn that only a portion of what we do save will merit specialized avenues of access. If we do not practice such restraint and temperance, the national bibliographic databases will grow to useless proportions and our processing backlogs will overwhelm us. We need to represent those materials deemed worthy with as much specificity as possible to stem the tide against the meaninglessness of massive retrievals from electronic systems. As noted earlier, catalogs need to describe works and collections while distinguishing them from a myriad of others. To achieve the former without the latter will produce databases that are both enormous and brutally random. They will become the archivist's, and the librarian's, Moby Dick: an obsession to maintain with an overwhelming whiteness and lack of meaning and direction. Lester Asheim has observed that "the rich store of information to which librarians can now provide access has a tremendous potential for good—to the individual and to the society." He continues by noting that, "as collectors, librarians have contributed to the information overload which inhibits rather than promotes achievement of the goal we had in view." He asks librarians if they do not "have an obligation now to provide a solution to the problem [that they] have helped to create."[37] Is it not time that archivists started to face the problem of information overload and stopped being lulled into a false sense of security offered by national databases and the allure of superficial subject access?

Some call for scrapping the information systems we now have and starting over, but this will not solve all the problems. There will never be a "perfect" information storage and retrieval system for archival materials, even if archivists design it from scratch specifically to meet their needs, because language and the human mind are the real problems. Subject retrieval—or for that matter, any form of text representation—will never be perfect. Archivists must recognize this and move forward, balancing local and national needs and building systems that are useful and possible. In the long run, there is no substitute for well-selected index terms that represent the primary aspects of a collection. This is never easy, but the less effort put into representing materials in a database, the more difficult retrieval will be. Archivists must decide on which side of the retrieval equation they wish the effort and cost to fall.

[37]Asheim, "Ortega Revisited," 225–26.