A CONTENT ANALYSIS OF ARCHIVAL MARC RECORDS TO MEASURE COVERAGE OF TOPICAL SUBJECT AND FORM/GENRE INFORMATION

by

Lisa C. Stark

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in Library Science

Chapel Hill, North Carolina
August, 1999

Approved by:

____

Advisor

Lisa C. Stark. A Content Analysis of Archival MARC Records to Measure Coverage of Topical Subject and Form/Genre Information. A Master's paper for the M.S. in L.S. degree. August, 1999. 33 pages. Advisor: Helen Tibbo.

This study describes a content analysis of archival MARC records conducted to detect the presence or absence of topical subject and form/genre terms taken from each collection's finding aid. The analysis measures the extent to which some archivists adapt information from finding aids for input into MARC records. Archival standards are discussed, as well as problems found to impede the process of manuscript cataloging using MARC. A sampling of topical subject and all form/genre terms were taken from twenty finding aids. After a search of the corresponding MARC records, less than half of the chosen topical subject terms were found. Additionally, only 43% of the form/genre terms were found in corresponding MARC records, suggesting that cataloging practice is not representing this information from finding aids well.

Subject headings: Archival description

Archives – Cataloging

MARC system – Applications

Standardization – Cataloging

Subject access

Table of Contents

Introduction

Every year, millions of books, magazines, filmstrips, theses, and other materials pass through the hands of library catalogers. The task of describing these materials and making them accessible via computer is monumental, but for decades librarians have used standardization to assist in the undertaking. Catalogers throughout the continent, and now around the world, describe materials in such a way that cataloging records from participating libraries can be placed in large national databases, shared, and understood. The MARC (machine-readable cataloging) standard, appearing in the late 1960s, made this electronic data exchange possible and facilitated economical shared cataloging.

Standardized formats such as MARC facilitate storage and retrieval of information on millions of books, and allow users to view a description of an institution's holdings from anywhere electronic access to the catalog is available. Archivists, however, have long shied away from such standards, attempting to protect the unique nature of their collections and descriptive traditions. Instead of employing cataloging standards, archivists traditionally used extensive written guides, called *finding aids*, to describe their collections. Until recently, these documents were available only at the actual repository site, and only in printed form. Many archivists felt the idea of using the MARC bibliographic standard designed for library material did not fit the evidential and unique nature of their materials. In addition, the lure to save money by sharing cataloging expenses was absent, since all archival collections are unique.

By the late 1970s, archivists and librarians began to see the benefits of universal access via computer networks. Before that time, the National Union Catalog of Manuscripts (NUCMC), a print "database," had been the only central location to index manuscript collections. In the 1970s, the National Information Systems Task Force (NISTF) started work that led to the creation of a MARC format specifically for manuscripts and archives, AMC (Archives and Manuscripts Control).[1] Although limited in space and flexibility, and only reluctantly accepted by many in the archival community, this was a significant step forward in providing national electronic access to archival materials in union databases, such as RLIN (Research Libraries Information Network) and OCLC (Online Computer Library Center), as well as local online library catalogs.

As explained in the next section, the differences between the two collection surrogates, finding aids and MARC records, are significant. The present study attempts to measure the extent to which some archivists adapt information *from* finding aids to *fit into* the library's main form of access, the catalog of MARC records. Topical subject, form, and genre terms were culled from twenty manuscript collection finding aids available on the World Wide Web, and their presence or absence was counted in each collection's corresponding MARC record. After providing definitions of entities discussed, a methodology section describes in detail how the study was conducted. Results of the content analysis and conclusions are then presented.

[1] NISTF's work towards the establishment of AMC was published in David Bearman's *Towards National Information Systems for Archives and Manuscript Repositories* (Chicago: Society of American Archivists, 1987).

Definitions

Manuscript collection usually refers to a group of items of an evidential nature, often papers of some sort, centering on one particular person, family, or organization. For instance, the Robert L. Eichelberger papers might consist of letters Eichelberger wrote as a Civil War soldier to his family in South Carolina, and several ledgers from his business, a lumber mill. A family collection might contain a number of journals kept by an African-American woman living in the segregated South, photographs of her parents and some genealogical information. A *manuscript "collection"* can even be one item, such as an anonymous commonplace book full of French poetry compiled by a young girl in New Orleans. A *series* is a body of materials within a collection arranged with a unified filing system or maintained as a unit by the creator because of some relationship arising out of their creation, form, or function (Daniels & Walch, 1984, p. 342). *Correspondence* is a common series for manuscript collections centered on individuals.

Manuscript collections are not intentional resources for study, but historical records. Some collections are documentary of a person's work, such as the book drafts of a nationally known economics expert; or a publisher's correspondence and book galleys of southern fiction writers. A book from an author and publisher speaks directly to the reader. Manuscript collections, however, contain unprocessed artifacts from which scholars make their own interpretations; they appeal to a general audience only indirectly. Such evidential value is what makes a collection of papers meaningful. While a handwritten draft of a book may have similar or even identical intellectual contents to the published piece, only the draft may hold evidence of the author's writing process such as paragraphs marked out or comments in the margins.

Provenance is the guiding principle for arrangement and description for any manuscript processor. This principle states that records from one unit shall not be mixed in with another, despite any similarities they may have, particularly subject similarities. Such arrangement preserves the organizational context and course of activity that led to the creation of the items (Miller, 1990, p. 25). It tells the researcher something more than the individual pieces alone could tell. An equally important and similar principle in archival arrangement is *original order.* The ordering and grouping of collection items provide valuable information to the user. This context would be lost if the materials were reordered for some reason, such as for subject access, when the creator kept items filed chronologically. *Provenance* refers to a collection or series' relationship with other series or collections, whereas *original order* refers more to the order of items within a series or collection.

A *finding aid* is generally a written guide to the physical and intellectual contents of a collection, although it can take on any of a number of different forms. It is a descriptive tool that reflects the arrangement of the collection built on the principles of provenance and original order. A finding aid provides the primary form of access to a collection, and is often created by the archivist who has physically processed the materials. Common components of a finding aid include administrative information such as provenance; arrangement information, such as how the processor may have formed certain groupings or series to give an indication of intellectual connections within a collection; biographical or historical information about the creator; series descriptions; and container lists. More recently, archivists began to include subject headings to be used as index terms in online catalogs, but these are not "native" to archival finding aids.

Finding aids may vary in size from a few pages to hundreds depending on the material and depth of description.

MARC stands for *MAchine-Readable Cataloging* and is the primary access vehicle used to represent library materials, even materials such as software, maps, and government documents. A *MARC format template* is essentially an electronic workform, which has numbered and lettered fields for the archivist to use in the creation of a valid catalog record. The make-up of this workform is dictated by the American National Standards Institute (ANSI). Correct insertions *into* the template mean that the record will display correctly in the library's online catalog, and will permit searching. For instance, all *titles* are to be inserted into the 245 field (see example in Figure 1).

Figure 1. Example of a MARC record for William Styron's *Darkness Visible*

```
ADY-3522   Entered: 06/14/1993   Last Modified: 06/14/1993        DUKE_CATALOG


Type: a Bib l: m Enc l:   Desc: a Ctry: nyu Lang: eng Mod:   Srce:
 Ill:      Audience:   Form:   Cont:      Gvt:   Cnf: 0 Fst: 0 Ind: 0
 Fic: 0 Bio: a Dat tp: r Dates: 1992 1990 Control:

005;   ;  a 19920401000000.0 $
010;   ;  a    91050032  $ o 23941092 $
040;   ;  a DLC $ c DLC $ d NDD $
020;   ;  a 0679736395 (pbk.) : $ c $8.00 ($10.00 Can.) $
043;   ;  a n-us--- $
050; 10;  a RC537 $ b .S88 1992 $
082; 00;  a 616.85/27/0092 $ a B $ 2 20 $
100; 1 ;  a Styron, William, $ d 1925- $                       ←——   Author Field
245; 10;  a Darkness visible : $ b a memoir of madness / $ c William Styron. $  ←——  Title Field
250;   ;  a 1st Vintage Books ed. $                            ←——   Edition Field
260;   ;  a New York : $ b Vintage Books, $ c 1992. $       ←——   Publication Info. Field
300;   ;  a 84 p. ; $ c 21 cm. $                     ←——   Physical Description Field
500;   ;  a Originally published: New York : Random House, c1990. $  ←——  Notes Field
600; 10;  a Styron, William, $ d 1925- $ x Mental health. $
650;  0;  a Depressed persons $ z United States $ x Biography. $          Subject Fields
650;  0;  a Authors, American $ x Biography. $
650;  0;  a Depressed persons $ z United States $ x Suicidal behavior. $
```

PUBLIC DISPLAY

MATERIAL: Book

AUTHOR: Styron, William, 1925-

TITLE: Darkness visible : a memoir of madness / William Styron.

EDITION: 1st Vintage Books ed.

PUBLICATION: New York : Vintage Books, 1992.

DESCRIPTION: 84 p. ; 21 cm.

NOTES: Originally published: New York : Random House, c1990.

SUBJECT: Styron, William, 1925---Mental health.

SUBJECT: Depressed persons--United States--Biography.

SUBJECT: Authors, American--Biography.

SUBJECT: Depressed persons--United States--Suicidal behavior.

In the archival world, a *MARC record* is often a *condensation* of information found in the related finding aid. More details about this condensation are explained in the next section.

In this study, a *topical subject term* is any term or phrase referring to a general subject or class of items, as opposed to a proper noun. Examples include the terms *rabbit*, *movies* or *cooking*. Words or phrases such as *Peter Rabbit* or *Star Wars* are not considered topical since they refer to specific instances of rabbits and movies. *Proper noun subject terms* are very often found in the text of a manuscript item or its documentation. For instance, a manuscript of a book about Noam Chomsky would presumably have that specific string of letters (N-O-A-M- -C-H-O-M-S-K-Y) on the item, making it clear what to include in a finding aid or a catalog record. In contrast, *topical subject terms* are not usually found *on* an item. A royal duke's wedding invitation may have his name inscribed upon it, but it will probably not have the topical subject phrase, *religious rites*. This topic would have to be devised by the archivist for description at a fairly detailed level, whether included as part of a finding aid's scope and content note, or as a subject heading in a MARC record.

Controlled vocabularies provide an important standard for bibliographic subject access as they facilitate the formulation of subject headings for the MARC 6XX fields. The most common one for subject access in library catalogs is the Library of Congress Subject Headings (LCSH). The headings are words or short strings of words used to indicate the subject matter of the materials:

North Carolina – History

Cooking – Southern Style

Dogs -- Fiction

Most begin with a name or topic which is left alone or qualified with one, two, or three additional terms. So that materials on the same topic from different repositories can collocate in a database, consistency in indexing (via LCSH and MARC) is required. For example, a search using *cosmetics* would retrieve all materials a cataloger indexed under the word, but would not retrieve items indexed with other terms such as *makeup*. Controlled vocabularies are excellent tools for controlling large amounts of information, but users must know which terms to use in order to search effectively.

In this study, the term *form* usually refers to the physical format of an item in a collection such as *microfilm*, and *genre* refers more often to the intellectual or artistic form used, such as *poems*, or *short stories*. There are hundreds of examples, but here are a few common to manuscript collections:

Forms: maps, film, cassettes, CD-ROMs, and LPs

Genres: poetry, short stories, speeches, essays, and lectures

Both types of designations may contribute to a catalog record's specificity.

RLIN and *OCLC* are two large national databases (also referred to as *bibliographic utilities),* to which most libraries in the U.S. contribute catalog records. The *RLIN* database was created by The Research Libraries Group, whose members are top academic and research institutions in the U.S. and abroad. It was originally conceived for use by librarians to manage their collections, but has broadened in use to include researchers and other lay users. Thousands of records for library materials are contributed to the database daily (Hannon, 1998, p. 2). *OCLC* began in a similar fashion in the early days of shared cataloging, and receives records from many academic institutions (Smith, 1998, p. 251). Both databases are used by librarians and patrons alike, to retrieve needed catalog records, to facilitate interlibrary loan requests, and for other information needs. They have now become the primary databases used to search for archival materials as well. Each database contains millions of records, sometimes making search retrieval sets extremely large and unwieldy.

A *search* refers to any formatted command a user enters into such an online database for finding relevant materials. There are a variety of types of searches, including the one most pertinent here, the *subject search*:[2]

S=West Virginia

S=Jogging

These searches would retrieve books or other cataloged materials indexed under the search terms, here *West Virginia,* or *Jogging*. A *known-item* search aims to find a specific instance of an item usually by title:

T=Gone with the Wind

T=Darkness Visible

Both *subject searches* and *title searches* are examples of *field searches*. For instance, a *title search* such as "T=Dreams," will search the 245 (title) fields of all records in the database. In some databases, if the exact text string entered is found in the beginning of a particular 245 field, that record will be retrieved. A *keyword title search* would find the string anywhere in the 245 field. Rather than searching a specific field, *keyword searching* searches nearly all fields likely to contain topical information. A *keyword*

[2] Specific search syntax and labeling may vary in different databases.

search for a specific string would retrieve all records containing that string in the title field, a note field, a subject field, or other field depending on the database design.

Retrieval is the action of "bringing up" or selecting (*retrieving)* records by using a search command. In large databases, retrieval results may be very different from results in small ones due to the sheer volume of records. In RLIN or OCLC, there is a great potential for retrieving too many records when searching topical subjects or keywords. For example, one common LC subject heading -- United States – History – Civil War, 1861-1865, -- presently retrieves 66,532 items from OCLC's *WorldCat* database. If the subject fields were searched with this particular heading, many online catalogs would only retrieve any book, software, manuscript, or other item with that entire subject heading assigned. Some search engines, however, will search for each word of the heading independently rather than look only for those items with the entire string. For instance, a user looking for information on the slave trade in Georgia, may enter this subject search:

S=Slave trade – Georgia

This command would not only retrieve records with the LC subject heading of *Slave trade – Georgia*, but depending on the retrieval system, might also find records with these two subject headings:

Slave names

Peach trade – Georgia

This kind of retrieval allows for broader retrieval sets, but also for more items that are irrelevant. The potential for retrieving a great deal of irrelevant materials grows greater with the size of the database. Smaller databases also have these problems, but to a lesser extent. (Tibbo, 1994a, p. 314).

In some databases, it is possible to *limit searches by format*. Some national utilities such as OCLC allow the searcher to limit retrieval by specifying a format (e.g., "bks" entered after the find command in OCLC's Union Catalog will limit the retrieval to books only). Until recently, searches could be limited to retrieve manuscripts if the "amc" code was used. Since all formats were integrated in 1996, manuscript collections are now cataloged on a variety of format templates. This means that limiting a search to retrieve only manuscripts is no longer possible.[3]

Literature Review

How well have archivists been able to provide topical subject access to their collections using the MARC standard? The history of archives and standards may help to answer this question. The next section will first look at the National Information Systems Task Force (NISTF), created in the 1970s to plan a national system for access to the country's manuscript collections. Then discussed are the creation of AMC (the MARC format developed specifically for manuscripts), and the reluctance of archivists and librarians to mix their respective materials in a single computer system. How MARC accommodates archival information and the difficulties of subject analysis are then reviewed, followed by the importance of topical subject terms, and the difficulties of retrieval. Finally, possible solutions to these problems are presented.

[3] MARC format integration has made retrieval of archival and manuscript materials from large databases such as OCLC and RLIN more difficult. The advances NISTF encouraged with AMC were overturned,

The Development of Archival Standards

In the 1970s, the Society of American Archivists (SAA) created a group, the National Information Systems Task Force (NISTF), to look at the larger picture of creating a national information "system" for archives and manuscripts. At that time, their thinking was to keep archival information separate from bibliographic materials (Lytle, 1984, pp. 354-360). NISTF's chair, Richard Lytle, explains in one article how he narrowed the group's role to something more practical, and finally, NISTF was able to establish "preconditions" for archival information exchange. Their focus became the standardization of archival "data elements," and a data element dictionary was created (1984, p. 360).

As Steven Hensen describes, the divergent descriptive practices of librarians and archivists made the development of the MARC/AMC format a slow one. Archivists felt that MARC was too bibliographically oriented to accommodate their unique items (1988, p. 539), and librarians, who feared MARC would need to be altered for archives, felt that such changes would defeat the purpose of a standard (Bearman, 1990, p. 243). Thus, technical considerations and political tensions were linked as both librarians and archivists slowly accepted each other's presence in the discussion on standards. Librarians wanted to maintain the integrity of the strict MARC standard, but the nature of the information about manuscripts was nothing like bibliographic information, which fits nicely into MARC fields.

Bearman writes that the larger architecture of databases like OCLC and RLIN still presumed the presence of only bibliographic information. For instance, primary

since choice of MARC format is now made according to physical format. Thus a reel of film, regardless of searching choices (title, author, etc.) still were based upon how access was provided to books, and not manuscripts. He called this subtle "threat" a symptom of "bibliocentricism" (1989a, p. 32). Published materials vastly outnumbered all other formats cataloged, then and now.

Although some archivists still consider descriptive standards, and even wide access, inappropriate for manuscripts, both librarians and archivists came to realize that the large bibliographic utilities of OCLC and RLIN could be useful conduits of access to manuscripts. "The two principal obstacles to making the quantum leap from reluctance to acceptance were the lack of a MARC-compatible format that fully met the needs of archival description and conversely, the lack of a system of archival description that was truly MARC-compatible" (Hensen, 1988, p. 540). In 1983, NISTF created the *AMC* MARC format tailored for use by archivists to exchange data. AMC was a format flexible enough to hold archival information, but still a standard that could bring archival materials into the national bibliographic utilities and into local online catalogs. Of importance to archivists was the fact that searches could be conducted to retrieve *only* manuscript records, thereby keeping their materials distinct. Perhaps more significant was the fact that thousands of collections would now have basic online subject access. A patron could submit a subject search on a single topic and retrieve not only books on that subject, but manuscripts under that topic as well.

Cataloging Archives and Subject Analysis

With the creation of AMC (1983), the profession now had a machine-readable format and records began to flow into national databases and local library online catalogs.

its archival status, must be cataloged using a film MARC format template.

Archivists, previously used to free-text description, now were faced with preformatted templates, coding, and controlled vocabularies. How was the archivist to proceed with MARC and create useful records? As explained in *Archives, Personal Papers, Manuscripts, and Archives* (*APPM*), finding aids are created to be a summary of the most important information from a manuscript collection. Following this, the MARC record is to become an even briefer summary or condensation of information from the finding aid. Helen Tibbo discusses a possible danger in the cataloging process:

> 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. (1994b, p. 312)

The information lost is likely to contain topical subject matter. She goes on to say, however, that if the information left out is done selectively, the MARC surrogate could become a powerful tool for retrieval, eliminating extraneous information. Such "noise" might contribute to the retrieval of excessively high numbers of records, and information only tangential to the search subject.

The task of the archivist then is to identify what is most important from the finding aid to put into the MARC record. The unique nature of manuscripts can make this extremely difficult to discover. Manuscript collections large and small can have tremendous diversity of both physical formats *and* subject matter. The materials could easily contain dozens of *different topics* of interest to scholars. For example, a senator's papers might easily have information on his or her home state, national elections, and fly fishing, the senator's favorite pastime. Such variety makes analysis difficult, but it is up to the archival processor to discover the essential topical subjects. This problem of multiple topics, the difficulty in topic discovery or *aboutness* (manuscript collections are not usually written *about* anything), and the ramifications for users are discussed next.

Particularly in the humanities, as Tibbo writes, diversity and complexity of collections inherent in a group of artifactual materials can challenge archivists in subject analysis. Besides coming in a variety of forms, materials are unpublished and contain no unifying bibliographic information (1994a, pp. 608-11), making subject access all the more important. In her study of manuscript subject access, Dooley agrees, especially regarding particular searches for visual archival materials, like sketches or drawings (1992, p. 347). These may have no textual clues for the archivist at all. Unlike book cataloging, there is usually no title page, CIP[4] cataloging, and sometimes even no text from which to work. Bearman writes that it is not only the diversity of the materials being cataloged, but also the diversity of user perspectives that complicates the process (1989b, p. 289). Archivists must know the materials well, but, like book catalogers, they must also know how the user may be trying to access them.

Archival materials are so fundamentally different from books (as the diversity of formats and subjects reveals), we must return to the concept of provenance for guidance. In her discussion on archival standards, Jackie Dooley notes that some archivists rationalize that provenance leads, albeit indirectly, to subject content, thereby relieving the archivist of any other sort of subject indexing needed for the MARC standard (1992, p. 345). Pugh also notes a similar rejection of standards by archivists:

[4] Cataloging in Process, or minimal cataloging from the Library of Congress often found in the preliminary pages of new books.

> Archivists respond instead to the unique, organic, and activity-based quality of records. Basic to archival arrangement is the canon that records cannot be arranged according to an enumerative scheme but must be arranged according to the principles of provenance and original order, reflecting the processes that created them. (1982, p. 34)

Provenance and original order have long been the archivist's primary concerns, according to which all processing and cataloging is completed. For help in discovering subject access points, Hensen contributes the idea of searching for a "bibliographic identity:"

> It is only insofar as these materials provide a record of major and minor historical events that they assume value and interest as tools of research. In order to provide access to this research potential, the manuscripts must be assigned a bibliographic identity. With published materials this identity is *prima facie*, deliberate, and straightforward, with most of the data that defines this identity provided clearly and explicitly, usually on the title page. With unpublished materials, however, this identity must be created through a process of formulating and extracting the elements of bibliographic description from the content and context of the manuscripts. (Hensen 1988, p. 543)

Archivists must often go beyond the work of a library cataloger, and look for suggestions not only in the collection itself, but also from its history. A travel journal may directly mention nothing of its creator, but an archival cataloger may look for clues about the age of the creator and geographic origins.

This goal of extracting what is most important means the archivist must consider what a collection is "about." On a higher level, how might an archivist discover what an abstract drawing, or a list of purchases for a farm, or a group of landscaper's files is "about"? Dooley writes that discovering the subjects "is the idiosyncratic and specific problem for archivists to solve" (1992, p. 348). Since different people would have different opinions as to what something is "about," cataloging suffers with inconsistencies.

This "aboutness" dilemma is also treated by other authors including Smiraglia. He notes that letters *to* someone may not be *about* them, but the actual collecting of the letters *is by* the person. "[C]an't a scholar learn something about a subject by perusing the correspondence received by that person? So in some sense the collection of letters is, in its entirety, *about* [italics added] the recipient" (Smiraglia, 1990a, p. 10). The intellectual differences between the subject matter indirectly received, and overt naming or description are evident and lie in the principles of provenance and original order. Again, the meaning of the collection lies in its evidential "unique" nature.

In his study of how library science indexing standards can be applied to museum objects, Steven Shubert discusses aboutness as having two senses. He claims that the "intensional" type of aboutness is influenced by the topic's current environment, varying regularly. The second sense of aboutness is more stable:

> A certain core of aboutness may be recognized as the explicit, universally valid, context-free, inherent subject of an item. This "extensional" concept of aboutness may be used in indexing and classification. (Shubert, 1996, p. 83)

Here, he refers to a main theme, or summary topic. It is specifically this type of aboutness, however, that Smiraglia and others feel does not exist. Looking at family letters, election posters, and a ledger from the oldest son's book sales, how could an archivist assign one specific theme? Smiraglia attempts to guide the archivist in certain general directions to determine such aboutness, but feels that broad subjects or themes, such as Shubert's concept of extensional aboutness, renders subject access meaningless. Many times, a theme that would fit an entire collection could be so broad that searching it would retrieve thousands of records. For example, southern repositories hold many letters written by soldiers during the Civil War. Each collection could be retrieved with a keyword search for *Civil War*, but the overall number of records retrieved would be too large to sort through effectively. Tibbo feels that indexing significant collection *series*, and not just collection level concepts, would assist in alleviating this problem (personal communication, July 1999).

For materials on similar topics to collocate in a database, archivists must consistently apply subject headings. Avra Michelson conducted a study showing high rates of inconsistency in archival indexing.[5] Dooley cites this study in arguing for more description uniformity within the profession (1992, p. 347).

In addition to the MARC format, other limits are placed upon archival cataloging (Helen Tibbo, personal communication, July 1999) that might contribute to inconsistency. OCLC's union catalog has field length and character limits as do smaller institutional online catalogs. Long abstracts or scope notes cannot be squeezed into MARC records in these databases. Only a small number of 520 fields, free-text note fields, are allowed. Assigning controlled vocabulary subject headings can also be limiting. Library of Congress Subject Headings must be ordered in very specific ways to allow collocation of materials on the same subjects. Poor choice of search terms on the user's part is also a hindrance for precise retrieval, even when the archivist has included appropriate cataloging.

The User's Environment

Subjects, and topical subjects in particular, have become essential access points for manuscript collections. They become the *most* important access points when researching broad topics, as opposed to specific people, places, or things. This is especially true for users new to a field. At a time when special collections libraries are attempting to attract more undergraduates as users, topical term indexing is crucial. A freshman in an introductory ecology class is much more likely to try a topical search term such as *deforestation*, than he is to look for and search on specific scholars prominent in the field.

Jackie Dooley notes that even sophisticated users still do subject searching, especially as fields become more interdisciplinary. In addition, scholar's interests and methods can change over time, such as the move to studying the masses instead of the elite. This means more general subject indexing, as fewer names searched are famous (Dooley, 1992, p. 351). Indeed, many innovations in online catalogs such as boolean searching and online index browsing, came about as studies affirmed that topical subject access was more common than had been thought (Drabenstott & Vizine-Goetz, 1994, p. 124). Karen Markey Drabenstott and Diane Vizine-Goetz cite one study by Marilyn Ann Lester (1989) in which over 70% of the queries in an online catalog were topical (as cited in Drabenstott & Vizine-Goetz, 1994).

A subject entered into the MARC record's subject fields (6XXs) can be searched directly. For instance in the Styron example, this search:

S=Depressed persons

[5] Michelson's study is documented in his 1987 article "Description and Reference in the Age of

would retrieve Styron's book, since that phrase is precisely what is entered in the 650 field. This same phrase placed in the 500 field (free-text note field), however, would only be found by a keyword subject search. In either case, a search for a general topic (e.g. rabbits, gunfire, bookends…) is risky in a large database, since there is potential of retrieving all records containing these words. Depending upon the database, even those records that contain either *depressed* or *persons* might be retrieved as well for the *depressed persons* search.

Possible Answers

There are ways archivists can lessen problems associated with subject analysis and use of the MARC record. Shubert discusses one way called *facet analysis* (1996, p. 362) relating to indexing at the series level (Helen Tibbo, personal communication, July 1999). For museum objects especially, he feels that looking for an item's *set* of topics (or facets), can result in better description and indexing, than if only one central theme were indexed (1996, p. 362). Thesauri can help with consistency in this matter. For some time now *Art & Architecture Thesaurus* (*AAT*), has been serving this purpose in museums, and in many libraries and archives as well. Tibbo (1994a, p. 314), and Dooley (1992, p. 348) note that increased use of such thesauri will promote consistency among archivists and will result in analysis that is more detailed. More detailed analysis, however, and summarizing could be prohibitively time consuming and expensive for many repositories.

A second suggestion for archivists is to include form or genre terms in MARC records, thereby increasing retrieval precision. Users would then be allowed to limit searches by form or genre, and avoid excessively high retrieval sets. OCLC allocates two

Automation," *American Archivist* 50 (Spring): 192-208.

MARC fields specifically for such terms, 655 (Index Term—Genre/Form), and 755 (Added Entry—Physical Characteristics). These fields, however, do not yet have wide usage. Tibbo, Smiraglia, Lytle, Bearman, Shubert, and Dooley again all mention thesauri, authority files, form/genre terms or other kinds of lists from which all archivists could work, increasing consistency and retrieval that is more precise.

Dooley writes about how form terms can even help indicate certain subjects (1992, p. 348), thereby contributing to a collection's "aboutness." For instance, even without looking at the content, a ball invitation could be assigned a subject heading such as "Social customs." In some databases, *form* terms may be searched directly as "subjects," as well as part of any keyword search.

Methodology

To learn more about how well archivists are able to represent topical and form/genre information in MARC records, a test was conducted. First, a sample of finding aids with corresponding MARC records was selected. The finding aids were drawn from the complete lists posted on the web-sites of both Duke University's Rare Book, Manuscript, and Special Collections Library, and of the University of North Carolina at Chapel Hill's Manuscripts Department. From those, a purposive sample was taken of ten finding aids from each repository. The selections spanned a range of finding aid lengths, a range of finding aid authors, and a variety of materials creators and subject matter. In this fashion, the sampling was felt to be representative of the variety of finding aids which Duke and UNC post, and will hopefully have some generalizable characteristics for other finding aids from similar institutions.

Certain characteristics disqualified a collection finding aid from the experiment.

- If a chosen finding aid had no corresponding MARC record, another similar one was chosen to take its place.
- "Preliminary" finding aids were disqualified since they are likely to be altered or rewritten completely.
- Finding aids with no named author (such as "Staff") were excluded.
- For simplicity, finding aids with more than one corresponding MARC record were also excluded.

In the next step, all topical noun phrases in each finding aid were identified, counted, and the finding aid section where each appeared was noted. The list of terms and phrases was further narrowed in these ways:

- Those terms or phrases that occurred less than five times in a finding aid were excluded. This limited the test to topical terms commonly mentioned, those more likely to be candidates for a search.
- Those terms considered unreasonable possibilities for a search were excluded. Primarily this meant pronouns, terms which often refer to particular instances of a topic, and which have no meaningful content themselves. Terms such as *materials* were also excluded for this reason.
- Form/genre terms were excluded since they primarily refer to physical and intellectual form instead of content. Additionally, their "subject-ness" was to be tested in the second part of the analysis.

From the remaining terms in each collection's set, five were randomly chosen. One of the smaller collections only had three terms appearing in its finding aid five or more times, which is why the total number of terms equaled ninety-eight. Examples from the final sample include these: *revisions, cotton*, *health*, *overseers, saxophone*, *copyist, bank, government*, *nuclear weapons*, *alliance*, *silver*, *canvassing*, *positions*, *societies*, *salesman*, *tour*, and *taxes*.

Each of the final set of terms was searched in its collection's corresponding MARC record using Microsoft Word's "Find" command.[6] Each singular and plural instance of each term was counted as one occurrence. Otherwise, only exact matches were counted as well as the MARC field in which each appeared. For instance, the terms *mother* and *mothers* were counted as one occurrence each, but *motherly* was not counted at all.

Finally, all form or genre terms were identified in each finding aid, and subsequently searched in the corresponding MARC record. What constituted a form or genre was interpreted broadly. Anything that could be construed as a physical form or intellectual form was counted, with one exception. The word *materials* was not considered a physical or intellectual form. It conveys no meaning as to what it represents other than representing some *physical or intellectual thing or matter* that is in a collection.

Findings

Of all ninety-eight topical subject terms, about 40% were *not found at all* in the corresponding MARC records, reflecting cataloging practice that has represented information in finding aids more often than not, but not represented it well. Fifty-nine terms (or 60%) were found *somewhere* in their respective collection's MARC record *at least once* (see Table A) and would most likely have been retrieved with a keyword search, but the topical information represented by the other 39 terms was lost.

[6] Microsoft Word 97 was used in the study.

For two collections, the MARC records contained *only one* of the five topical subject terms selected from the finding aids. Neither of these collections is very small, one having around 5,000 items and the other 15,500, again suggesting that a great deal of information was completely lost in the condensation process from finding aid to MARC record.

Among all twenty collections, only two had MARC records containing *all* of the selected topical terms; in these cases, condensing information from the finding aids meant no loss of information with regard to these terms. The MARC records contained specific information, varied enough in subject matter to hopefully be discriminating surrogates, but standardized enough to fit within the MARC structure. One of these two collections has about 1,000 items, which is relatively small, but its MARC record contains three 520 (free-text notes) fields. The other collection only has one item. *APPM* states that the step of creating a finding aid may be unnecessary for a small collection (1989, p. 4). In these two cases, finding aids were most likely created first, and then the MARC record was created second in order to represent the collection online. The finding aids, however, do not provide much more access than the MARC record.[7]

Other percentages varied and reveal inconsistencies in comprehensive subject analysis: Five collections each had 40% of its terms found, five had 60%, and six had 80%.

[7] Recent advancements with SGML and the finding aid document type definition EAD mean more and more finding aids will be represented online, providing even more precise access to specific components of a finding aid.

Table A Number and Percentage of Term Occurrences in MARC Record

Collection Name	Number of Topical Subject Terms Selected	Percentage Found in MARC	Number Found in MARC	Number Not Found in MARC
Ammons	5	40%	2	3
Ballard	5	20%	1	4
Burke	5	80%	4	1
Craven	5	40%	2	3
Currency	5	40%	2	3
DMBB	5	40%	2	3
Guthrie	5	20%	1	4
Hanks	5	80%	4	1
Harley	5	60%	3	2
Howland-McIntosh	5	60%	3	2
Ker	5	100%	5	0
KKK	5	80%	4	1
Mahone	5	60%	3	2
McLeod	5	80%	4	1
Morgenstern	5	60%	3	2
Pescud	3	100%	3	0
Sellers	5	80%	4	1
Southeastern	5	60%	3	2
Walser	5	40%	2	3
Wootten	5	80%	4	1
Totals	**98**	**60%**	**59**	**39**

Of all of the terms found in the MARC records, less than half (49%) were found in the subject fields (6XXs, See Table B). A keyword search would find all of these terms, but if a user were to enter one of these terms in a subject search, the record would not be retrieved and possibly more information then would be lost to the user.

Table B Number and Percentage of Term Occurrences in Specific MARC Field

MARC Field	Number of Terms in Field	Percentage of the Total Found in MARC
110	2	3%
351	2	3%
500	5	8%
520	51	86%
544	1	2%
545	18	31%
600	2	3%
610	17	29%
650	15	25%
651	4	7%
6XX fields	**29**	**49%**
Other fields	**57**	**97%**

Note. Some terms occurred in multiple fields

In most cases, the topical subject was either directly represented in the 6XXs, or not represented at all. For instance, *trombone* was searched in a MARC record that only contained the following subject headings:

600	Burke, Sonny, 1914-
650	Big band music
650	Big bands -- United States
650	Jazz musicians
650	Jazz -- 1941-1950
650	Jazz -- 1951-1960
610	Decca Records (Firm)

No synonym for *trombone* was found either. Terms such as *jazz*, while related, convey nothing of the topic *trombone*. This complete absence of the topic *in the 6XXs*, was the case with 64 of the 98 terms.

Including *trombone*, thirty terms had *no exact match* in the 6XXs, but one or more of the present 6XX fields could conceivably be seen as a general heading for that term. For instance, it is not surprising to find the term *trombone* in a collection with the subject *jazz music* attached to it. The reference is very broad, but the absence of the term *trombone* could be considered a detail lost to collection level or series level cataloging. Other examples in the sample include these in Table C:

Table C Examples of Terms and Corresponding General Headings Found in MARC Record

Term	Series or Collection Level Headings in MARC
overseer	*Slavery, Slave trader*
instrumentation	*Jazz, Big band music*
denomination	*Paper money*

Also in a number of cases (twenty-two), the sample term was found to match one word of a proper noun or topical subject phrase in the 6XXs. See Table D:

Table D Examples of Terms and Corresponding Proper Nouns or Topical Subject Phrases Found in MARC Record

Term	Proper Noun or Topical Subject Phrase Found in MARC
university	*Princeton University*
family	*Brownson family*
forest	*Knights of the Green Forest*
life	*social life and customs*
construction	*construction equipment*
education	*education, cooperative*

In these cases, some subject keyword searches would find these records.

Table E Number of Term Occurrence by MARC Subject Field

Field	Count
600	78
610	63
630	1
650	161
651	43
655	8
Total	354

Among the 6XX fields (see Table E), the topical subject field, 650, was the most common. Most of the time, however, the subjects they contained did not include the sample term (all terms that occurred five or more times in the finding aid). This suggests that while archivists include topical information in a MARC record, they do not consistently input the information into the fields specifically designed for it.

Table F shows a breakdown of the term occurrences by finding aid section. Of the terms found in the MARC records, the higher number came from either the scope notes/abstracts or the series descriptions, the richest parts of a finding aid, where free-text description is found. Highlighted is the fact that nonstandard description nearly always holds richer language than standardized description with controlled vocabulary and length restrictions. The same holds true for an interviewer who might receive more depth and richness of testimony by asking open-ended questions as opposed to giving a multiple choice questionnaire.

Table F Number and Percentage of Term Occurrences by Finding Aid Section

Finding Aid Section	Number of Terms from Finding Aid Section	Percentage of Terms from This Section Found in MARC	Number Found in MARC	Number Not Found in MARC
Abstract/Scope & Content Note	70	79%	55	15
Administrative Information	2	100%	2	0
Biographical/Historical Note	39	69%	27	12
Brief Description	3	100%	3	0
Other	6	50%	3	3
Series Descriptor/List of Series	71	62%	44	27
Container List	40	50%	20	20

A second analysis was conducted to see if form or genre terms found in a collection's finding aid were also included in the collection's MARC record. All form terms found anywhere in each finding aid were searched in the corresponding MARC record (see Table G). Examples of the form terms searched include the following: *correspondence*, *speeches*, *letters*, *ledgers*, *daybooks*, *commonplace books*, *account books*, *fliers*, *broadsides*, *licenses*, *muster roles*, *scores*, *photographs*, *tintypes*, *ambrotypes*, *deeds*, *titles*, *appraisals*, *poems*, *short stories*, *scrapbooks*, *daguerrotypes*, *stereographs*, and *posters*.

Table G Number and Percentage of Form/Genre Term Occurrences in MARC Record

Collection	Number of Form/Genre Terms Found in Finding Aid	Percentage Found in MARC	Number Found in MARC	Number Not Found in MARC
Ammons	23	57%	13	10
Ballard	23	35%	8	15
Burke	7	43%	3	4
Craven	45	31%	14	31
Currency	11	27%	3	8
DMBB	42	38%	16	26
Guthrie	19	42%	8	11
Hanks	39	62%	24	15
Harley	37	43%	16	21
Howland-McIntosh	37	49%	18	19
Ker	22	50%	11	11
KKK	10	70%	7	3
Mahone	22	32%	7	15
McLeod	31	45%	14	17
Morgenstern	33	27%	9	24
Pescud	2	100%	2	0
Sellers	14	57%	8	6
Southeastern	22	18%	4	18
Walser	26	23%	6	20
Wootten	18	83%	15	3
Totals	**483**	**43%**	**206**	**277**

Of the 483 form/genre terms found in finding aids, only 206 (43%) were found in the corresponding MARC record. Increased use of these terms in the future might increase specificity of searches found, since many form terms are unique to manuscript collections. Keyword searches using a form term or limited by a form term may retrieve sets that have a higher percentage of manuscript collections than book titles. This unique feature can perhaps help keep manuscript records distinct since the AMC format and search label are no longer available for this purpose.

Table H Number and Percentage of Form/Genre Term Occurrences in Specific Field

MARC Field	Number of Form/Genre Terms Found in Field	Percentage of total
245	20	10%
300	18	9%
351	17	8%
500	11	5%
520	167	81%
544	10	5%
546	1	0%
555	1	0%
610	2	1%
650	15	7%
655	7	3%

Note. Some terms found in multiple fields.

Although form/genre terms were sought and counted in all MARC fields, including the controlled vocabulary subject fields, the vast majority of those found were in the 520 fields, free-text content note fields. This again reveals how archivists may tend to rely on the free-text fields instead of dealing with controlled vocabulary. Only 3% of the form terms found were in the 655 field, specifically created for them (see Table

H). None were found in the 755, a similar field for physical characteristics. Again, subject searches for these form/genre terms would have retrieved only the seven terms found in the 655 field.

Conclusion

Manuscripts, while fundamentally different from books, have found their way into large bibliographic databases with the help of the MARC standard. The archivist's profession has advanced greatly since NISTF began their work in the 1970s; OCLC and RLIN now contain thousands of MARC records from archival institutions, and cataloging has become a regular part of many archivists' work. Impediments to archivists' use of the bibliographic format, however, remain. In this study, even those terms represented repeatedly in a finding aid were not detected in the MARC record 51% of the time. Some of the selected MARC records contained all of the chosen terms and some contained only one. These results confirm inconsistencies in cataloging.

Despite the impediments caused by the MARC standard and the unique nature of manuscripts, archivists must try to capture the most important concepts of collections, and to convey them succinctly so as not to retrieve other titles or collections tangential to the subject desired. They must capture as many different subjects in a collection as possible, and include form and genre information.

In this sampling, clearly there is more representation that could be included. The findings suggest that archivists should find ways to incorporate more topical and form/genre information into MARC records, and thus increase access for library and archive patrons.

References

Art & Architecture Thesaurus (Version 3.0) [Database]. (1999). Available http://shiva.pub.getty.edu/aat_browser/ [Los Angeles]: J. Paul Getty Trust.

Bearman, D. (1989a). Archives and manuscript control with bibliographic utilities: Challenges and opportunities. *American Archivist, 52,* 26-39.

Bearman, D. (1989b). Authority control issues and prospects. *American Archivist, 52,* 286-299.

Bearman, D. (1990). Can MARC accommodate archives and museums? Technical and political challenges. In T. Petersen & P. Molholt (Eds.), *Beyond the Book: Extending MARC for Subject Access* (pp. 237-246). Boston: G.K. Hall.

Daniels, M. F. & Walch, T. (Eds.). (1984). *A Modern Archives Reader: Basic Readings on Archival Theory and Practice.* Washington, DC: National Archives and Records Service, U.S. General Services Administration.

Dooley, J. M. (1992). Subject indexing in context. *American Archivist, 55,* 344-354.

Drabenstott, K. M., & Vizine-Goetz, D. (1994). *Using Subject Headings for Online Retrieval: Theory, Practice, and Potential*. San Diego, CA: Academic Press.

Hannon, H. (1992). *Discovering RLIN: an Introduction to the Research Libraries Information Network*. Mountain View, CA: Research Libraries Group.

Hensen, S. L. (1989). *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies, and Manuscript Libraries* (2nd ed.). Chicago: Society of American Archivists.

Hensen, S. L. (1988). Squaring the circle: The reformation of archival description in AACR2. *Library Trends,* (Winter), 539-552.

Lester, M. A. (1989). Coincidence of user vocabulary and Library of Congress Subject Headings: Experiments to improve subject access in academic library catalogs (Doctoral dissertation, University of Illinois at Urbana-Champaign, 1989). *Dissertation Abstracts Online, 50-07A,*1838.

Lytle, R. H. (1984). An analysis of the work of the National Information Systems Task Force. *American Archivist, 47,* 357-365.

Miller, F. M. (1990). *Arranging and Describing Archives and Manuscripts.* Chicago: Society of American Archivists.

Pugh, M. J. (1982). The illusion of omniscience: Subject access and the reference archivist. *American Archivist, 45,* 35-56.

Shubert, S. B. (1996). Subject access to museum objects: Applying the principles of the subject approach to information from library and information science to the documentation of humanities museum collections (Doctoral dissertation, University of Toronto, 1996). *Dissertation Abstracts Online, 57-08A*, 3309.

Smiraglia, R. P. (1990a). Introduction: New promise for the universal control of recorded knowledge. In R. P. Smiraglia (Ed.), *Describing Archival Materials: the Use of the AMC Format* (pp. 1-15). New York: Haworth Press.

Smiraglia, R.P. (1990b). Subject access to archival materials using LCSH. In R. P. Smiraglia (Ed.), *Describing Archival Materials: the Use of the AMC Format* (pp. 63-90). New York: Haworth Press.

Smith, K. W. (1998). OCLC: Yesterday, today, and tomorrow. In K. W. Smith (Ed.), *OCLC, 1967-1997: Thirty Years of Furthering Access to the World's Information* (pp. 251-270). New York: Haworth Press.

Tibbo, H. R. (1994a). The epic struggle: Subject retrieval from large bibliographic databases. *American Archivist, 57,* 310-326.

Tibbo, H.R. (1994b). Indexing for the humanities. *Journal of the American Society for Information Science, 45,* 607-619.