ARCHIVAL ACCESS POINTS: SUBJECTS AND BEYOND
The North American experience

Tim Hutchinson
paper to be presented at
International Seminar on Subjects in Archives
San Miniato, Italy
12-13 Feb. 1998

INTRODUCTION

The issue of subject access to archival materials has provoked a considerable amount of interest and debate during the past number of years. One reason for this may be that the introduction of automated access systems has resulted in the potential to provide more detailed description than was available in traditional card catalogues, and the ability to provide a greater choice of points and methods of access to these descriptions. My role this morning is to relate some of the debate and practice in North America.

First of all, what is a "subject"? The *Library of Congress Subject Headings* (*LCSH*) routinely mixes "topical headings" with headings for form of material, time, and place (to name a few). As Jackie Dooley wrote in "Subject Indexing in Context," "[Richard Lytle] has stated that `requests for records by proper name, geographical area, date, or form may conceal a subject request.'" Dooley contends that

> Such requests do not conceal subject requests, they *are* subject requests. The archival literature often gives the impression that subjects are strictly generic topics--rain forests, football games, railroads, generals, or skyscrapers. In reality, specific named entities, including particular people, organizations, government agencies, geographic places, and events are no less subjects than are generic topics.[i]

On the other hand, Chris Hurley claims that "A function is not a subject. A function is not a subject. A function is not a subject."[ii] The additional access points which Dooley mentions (among others) are extremely important, and I would argue that it is pointless to treat any of them in isolation or as a self-sufficient access point. For clarity's sake, I will try to use "subject" to refer only to topical subjects. Perhaps it would better to say that not everything is a subject, but several access points can lead the way to a subject.

I will begin with a discussion of topical subject headings, both in terms of theoretical aspects and empirical studies; it should become clear that topical subject headings are not enough, on their own, for effective access to archival materials.[iii] I will also discuss relevant

literature from the library and information science field, especially in terms of the retrieval performance of various types of databases. In light of the problems with topical subject indexing, I will also say a few words about authority control and contextual access points -- but I won't have time to treat them in any detail.

Do archival materials even *have* subjects? Several authors have expressed doubts; for example, David Bearman has written that

> Archival material does not have a *subject* per se. Archival material is *of* the material that generates it, but seldom is it consciously authored to be *about* something. ... Archival materials are used to understand the contexts of their creation, and may be exploited for the specific information they contain, but the perspectives brought by users, both to the context of their creation and to the data they may contain are too diverse to support subject indexing.[iv]

Jackie Dooley acknowledges that, unlike books or journal articles, "most original source materials are written with no conscious subject or thesis in mind."[v] She claims, however, that "this essential difference ... hardly implies that subject terminology cannot be usefully be applied to archival descriptions." She gives this example:

> Despite the fact that the soldiers who penned [Civil War diaries] did not set forth format theses, draw particular reasoned conclusions, or neatly package their work with a table of contents and for ease of consumption, their writings are *about* certain things: the soldiers themselves, life in their regiments, specific places and times, particular Civil War battles, their hometowns, and their thoughts on life and death, to mention only a few obvious possibilities.[vi]

Even so, as Dooley concedes, this does not mean that subject indexing of archival materials is easy. As Avra Michelson puts it, archival records are often "heterogeneous collections that require many more index terms than those used to describe monographs."[vii] For example, in a study by Michelson (to which I'll return), the average number of index terms used by participants was 13; and Helen Tibbo founds that in a sample of OCLC catalogue records the average was 8, while some RLIN records have more than 200 index terms. (OCLC and RLIN are two national union databases in the United States.)

These problems seem to be widely acknowledged. Archival materials tend not to be easily described by a limited number of subject headings. Also, subjects headings may depend on how the materials are to be used, which is difficult to determine and can change over time. (An oft-cited example of this is the "discovery" of social history; archival collections were not indexed in a way that facilitated finding relevant material.) David Bearman is one of the most vocal critics of attempts at topical subject access in archives, claiming that "archivists should stop wasting their time on the effort to control topical subject terminology and instead should look for findings that can lead to more strategic approaches to vocabulary control."[viii] In my view, archivists should not abandon topical subjects, but there are certainly compelling arguments for not limiting indexing to topical subjects.

PROBLEMS WITH SUBJECT ACCESS:

Interindexer consistency:

Avra Michelson conducted an experiment of the archival repositories that in 1986 were contributing to the Research Library Group's (RLG) Research Library Information Network (RLIN). Participants in the study assigned topical index terms (using *LCSH*) to the same three descriptions of collections, using their own descriptive procedures.[ix] Good consistency was expected, "because survey respondents performed this exercise with the equivalent of an identical card catalog description in hand, preventing many of the opportunities for divergence that arise in drafting descriptions from the beginning."[x] However, this did not occur. For the first description, for example, "21 indexing repositories assigned 162 different access points. ... No term was assigned by all indexers, resulting in an indexing consistency rate of zero."[xi]

David Bearman has taken this study to demonstrate the "failure of topical subject-based authority control."[xii] This may be overstating the case somewhat. It is important to remember that Michelson's study used subject headings from *LCSH*, whose problems have been well-documented, and not only for archives.[xiii] It can be argued that *LCSH* is even more problematic for archival materials, however, especially since it was developed for books. Helen

Tibbo suggests that

> Effective subject access in ... large, heterogeneous databases may require the development of more specific, subject-oriented thesauri, such as the *Art and Architecture Thesaurus*. Smaller, more subject-specific databases, akin to the specialized bibliographies humanists have long used, may also provide better control of materials.[xiv]

It is important to note that studies of interindexer consistency of library cataloguing and journal indexing have generally yielded results similar to those of Michelson.[xv] That is, it's not a problem confined to archival materials. It might be interesting to do an interindexer consistency experiment with archives who use a thesaurus designed for a smaller group of archives (or even a single archives), such as the Public Archives of Alberta Subject Headings (PAASH).[xvi] To a great extent, PAASH is a small subset of *LCSH*, potentially better controlled. However, anecdotal evidence would suggest that even a smaller vocabulary does not guarantee better consistency. For example, the United Church of Canada Archives uses an in-house controlled vocabulary, based on *LCSH* and, less frequently, *Canadiana Subject Headings*. An archivist heading a project there reported the same sorts of problems.[xvii]

Subject access from large databases:

Avra Michelson's study highlighted the problem of a lack of consistency in indexing. The opposite problem also seems to exist. Helen Tibbo studied the success of subject retrieval in the OCLC Online Union Catalog, by choosing a "random sample" of 59 MARC AMC records describing collections in one repository, then searching the entire database for occurrences of the subject headings found in those 59 records. Restricting to manuscript materials, the mean number of postings per term was found to be approximately 60, with the median closer to 45.[xviii] For all records in the database (library and manuscript materials), these numbers ranged from 196 to 229, and 79 to 101, respectively.[xix] The latter finding is particularly significant if OCLC

and similar bibliographic utilities are to be used to retrieve materials regardless of format, but even the numbers corresponding to manuscript materials are high. A user study in an academic library found that although 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."[xx]

For librarians, the most important uses of OCLC are shared cataloguing and inter-library loan; that is, this database was originally intended for known-item searches rather than subject searches. According to Tibbo's study, for library materials represented in OCLC, the average number of postings per subject heading is extremely high (higher, indeed, than for archival materials). Thus Tibbo's study not only indicates that for large bibliographic databases, *LCSH* subject headings may inappropriate for archival materials (as they are too general), but that subject access in *general*, at least on its own, may not be suitable to retrieve catalog records from large bibliographic databases. It's hard to divorce studies about subject access from the vocabulary list being used, and there are indeed problems, not only for archival materials, with *LCSH*. However, it appears that *LCSH* (or subject indexing in general) and its application are problematic in large bibliographic databases; this is not a case of something failing for archival materials which succeeds for books. This is especially troubling for access to archival materials, however, since archivists or researchers seeking materials in remote repositories are unlikely to have the information necessary to facilitate traditional provenance-based access.[xxi] Or rather, it highlights the fact that systems supporting provenance-based access should be developed.

NORTH AMERICAN PRACTICE

A little over a year ago, I conducted an informal survey of subject indexing practices (conducted through the Archives and Archivists listserv[xxii]). 32 of 35 respondents indicated that their repository attempts to provide topical subject access. Of those, most (26) use *Library of*

Congress Subject Headings to some degree, but only seven use it exclusively. Ten use it along with an in-house vocabulary list (including controlled lists based on *LCSH*); four use it along with another published thesaurus (such as the *Art and Architecture Thesaurus*); there are also five respondents using a combination of the above and/or natural language terms. The survey also indicates that most (24) use a keyword-searchable database. I emphasize that this survey is extremely unscientific and should not be taken to generalize any group of archives (the information about availability of automated databases, especially, should be considered in light of the fact that this survey was available only to those archivists with access to the Internet).

However, the results relating to the use of *LCSH* are not surprising and probably fairly representative of current practice; *LCSH* is the most widely-available vocabulary list, but since it was not created for archival materials it is not entirely suitable, and archivists are trying to develop and adopt alternatives, such as PRESNET for the U.S. Presidential Libraries, the *Art and Architecture Thesaurus*, and the Public Archives of Alberta Subject Headings (PAASH).[xxiii] Another reason that *LCSH* is so widely used is that for the past dozen years, archival descriptive standards in the United States have largely been based on MARC -- that is, they've grown out of the library tradition.

EMPIRICAL STUDIES
PERSPECTIVES FROM LIBRARY AND INFORMATION SCIENCE

Controlled and free-text searching

I'd like to take a few minutes to talk about research in the library and information science community. With the advent of full-text databases and natural language searching capabilities, one reaction is to say, do we really need controlled indexing anymore? While library databases are of course different than archival databases, many of the principles of retrieval remain the same, so I think we can learn something by considering research in this area.

A great deal of research has dealt with free-text and (more recently) full-text searching, their effectiveness in searching, and their impact on the recall and precision of a search. A

distinction needs to be made between *free-text* and *full-text*; free-text (i.e. "natural language") searching does not need to involve the full text of a document.

To quote one researcher, "[C]onventional wisdom holds that free-text terms contribute to precision by virtue of being more specific and more current than controlled vocabulary terms. It also holds that a controlled vocabulary, by virtue of its classing functions, serves primarily to promote recall."[xxiv]

When we consider full-text databases, the situation appears to be reversed. At least two studies indicate that full-text searching leads to higher recall and lower precision than controlled-vocabulary searching.[xxv] This should not be surprising. As Jennifer Rowley notes, a "characteristic of full-text databases is the number of access points. Typically, with a very large database of full text it will be even more difficult to achieve acceptable recall at tolerable precision. Full text should give greater recall, but lower precision than a database of less than full text."[xxvi]

Generally, it is acknowledged that a combination of free-text and controlled-vocabulary searching is necessary. Jennifer Rowley undertook an extensive review of the literature in this area, and concluded:

> Despite much debate extending over more than a century, together with a range of research projects, information scientists have failed to resolve the debate concerning the relative merits of controlled and natural languages. There is general recognition that controlled language and natural language should be used in conjunction with one another, and there is some agreement as to the relative merits of each of these systems. This is based, however, on practice and experience rather than proved and tested research.[xxvii]

In an archival setting, a non-full-text database searchable by free-text terms might be compared to a fonds- or series-level description which includes a number of fields including administrative history and scope notes, analogous in many ways to an abstract. It is not clear, however, which model--full text or less than full text--is best (if at all) suited for analysis of the situation for archival materials. Clearly an archival catalogue record (a MARC record, for example) is not a "full-text" document, in the sense that it is a surrogate for a set of materials.

Until the EAD is more widely adopted, one will not normally search an entire finding aid either, but even a finding aid is not really "full-text." On the other hand, it could be argued that the finding aid is the "full text" and the catalogue record is the "surrogate," because a catalogue record is normally created from the finding aid, not from the archival records themselves. Indeed, the important characteristic of full text may not be whether the text is a surrogate for a more complete document, but rather how extensive the text is.[xxviii] That is, even though an archival description may not technically be a "full-text document," if the administrative history and scope and content notes are lengthy, the difficulties noted above by Rowley may still occur. One answer would appear to be that the situation for archives cannot be completely generalized from information science--in addition to the ambiguities just mentioned, there are theoretical difficulties about even applying subject index terms to archival materials--and that more studies of this nature need to be carried out specifically for archival materials.

Ribeiro: Controlled vs. uncontrolled index terms

I'm aware of only one similar study in an archival context. A study by Fernanda Ribeiro compared controlled and uncontrolled indexing languages. (I'm cheating a bit here -- the study was carried out in Portugal as part of a British dissertation, so it's actually outside the "North American" scope of my talk.) Indexing was done at the series level for three different record groups in a city archives. In the first database, the one with uncontrolled index terms,

> The search dictionary contain[ed]:
> - reference codes of each record;
> - complete names of the archival entities and each of the words that appear in these;
> - series titles and each of the words in the titles;
> - dates recorded in appropriate fields;
> - words marked between diamond brackets, in different fields.[xxix]

In the second database, the last category was replaced by controlled index terms; that is, the derived index terms were translated into authorized terms.

Unfortunately, Ribeiro excluded from the study the very type of records which critics of archival subject indexing view as the most problematic:

> [T]here are some series that, even with homogeneous document types, cover such a large range of subjects that content analysis is impracticable. The enormous variety of subjects dealt with in these series made it impossible to establish any objective criterion for content analysis or to identify the concepts. So, *these were not indexed*.[xxx]

Based on a calculation of precision, one conclusion was that the database with controlled subject terms (database B) "present[ed] a 13.6% better performance than" the database with uncontrolled terms (database A).[xxxi] However, it is also worth considering how well the databases work together. The "incremental advantage measure ... quantifies the advantage (or disadvantage) that would be obtained by adding to the records retrieved from a database, the records retrieved from the other database analysed in comparison."[xxxii] An analysis of this value allowed Ribeiro to conclude that:

> *The two databases are complementary*, because total overlap occurred in the retrieval for only 7 questions. In the great majority of cases, each database's retrieval showed an advantage when added to the other's. ... In view of these considerations, it must be concluded that combining uncontrolled subject indexing language with a controlled one, in the same database, is the most effective means to achieve better performance.[xxxiii]

This result is consistent with the findings in library and information science.

CONCLUSION

To conclude, I think it's clear that subject indexing is an important aspect of any archival retrieval system. With the introduction, for example, of the EAD standard (for encoding full-text finding aids), it's tempting to discard traditional tools such as subject indexing and authority control. However, research and experience seem to show that these tools need to be integrated into new systems.

Still, topical subjects are not enough on their own. Archival materials, of course, are arranged by provenance, and thus provenance has been a standard way of gaining access to them.

A well-known study by Richard Lytle, published in 1980, suggested that a combination of provenance-based access and subject-based access would be the most effective. Several archivists have been calling for provenance-based access to be more fully integrated into automated retrieval systems. For example, David Bearman and Richard Lytle have suggested that archivists need to

> view provenance information as a provider of retrieval access points; emphasize form of material and function in retrieval systems; establish provenance authority records; and integrate archival processes from records creation through records appraisal to records description.[xxxiv]

Provenance-based access has always been an important method of access for the mediated use of archives. That is, materials are located by figuring out which agency might have created relevant records -- and this is often carried out by a reference archivist who is very familiar with the institution. With an increased Internet presence for many archives, and the ongoing development of national union databases to be accessed by users without the mediation of archivists, it will be important to try to build in as many access points as possible -- but still make it as easy as possible for the average user to understand the retrieval system. These access points include function and form of material, and could be related to authority records (the International Council on Archives, for example, has adopted the International Standard Archival Authority Record, or ISAAR). This raises many issues, which I'm not able to explore in detail. But I hope I've managed to give you some idea of the situation in North America.

Notes

[i]Jackie M. Dooley, "Subject Indexing in Context," *American Archivist* 55 (Spring 1992): 348; emphasis in the original.

[ii]Chris Hurley, "What, if Anything, is a Function?" *Archives and Manuscripts* 21, no. 2 (1993): 212.

[iii]This discussion will be restricted to textual records (and, to an extent, corporate rather than personal records). There is a fair amount of literature about subject indexing of images (in the information sciences community in general), but that is unfortunately beyond the scope of this paper.

[iv]David Bearman, "Authority Control Issues and Prospects," *American Archivist* 52 (Summer 1989): 289.

[v]Dooley, "Subject Indexing in Context," 347-348.

[vi]Dooley, "Subject Indexing in Context," 348; emphasis in the original.

[vii]Avra Michelson, "Description and Reference in the Age of Automation," *American Archivist* 50 (Spring 1987): 199.

[viii]Bearman, "Authority Control Issues and Prospects," 288.

[ix]Michelson, "Description and Reference in the Age of Automation," 194.

[x]Ibid.

[xi]Ibid.

[xii]Bearman, "Authority Control Issues and Prospects," 288.

[xiii]Two comprehensive literature reviews relating to this are Pauline A. Cochrane with Monika Kirtland, *Critical Views of LCSH--The Library of Congress Subject Headings: A Bibliographic and Bibliometric Essay* (Syracuse, New York: ERIC Clearinghouse on Information Resources, 1981); and Steven Blake Shubert, "Critical Views of LCSH--Ten Years Later: A Bibliographical Essay," *Cataloging and Classification Quarterly* 15, no. 2 (1992): 37-97.

[xiv]Tibbo, "Indexing for the Humanities," *Journal of the American Society for Information Science* 45 (September 1994): 616.

[xv] For example, for *LCSH*, Yasar Tonta studied interindexer consistency between Library of Congress (LC) and British Library (BL) cataloguers, using 82 titles in the field of library and information science. It was found that "[b]y applying Hooper's equation for exact matches, the average indexing consistency value between BL and LC catalogers was found to be 16%. ... For

both exact and partial matches, the average indexing consistency value between BL and LC catalogers was found to be 36%." Similarly, Lois Mai Chan compared cataloguing done by LC cataloguers and cataloguers from other libraries that contribute to OCLC, and found 15% total consistency. Other studies have indicated that interindexer consistency is widely varied, and that there is a direct relationship between interindexer consistency and retrieval effectiveness. Chan notes, though, that many "earlier studies ... suffer from the lack of reliable means of measurement because of the presence of uncontrolled variables."

[xvi]See Jean E. Dryden, "Subject Headings: the PAASH [Public Archives of Alberta Subject Headings] Experience," *Archivaria* 24 (Summer 1987): 173-180.

[xvii]Ruth Dyck Wilson, "A Conversion Experience in the United Church Archives," *Archivaria* 35 (Spring 1993): 138.

[xviii]Helen R. Tibbo, "The Epic Struggle: Subject Retrieval from Large Bibliographic Databases," *American Archivist* 57 (Spring 1994): Tables 1 and 3, p. 317.

[xix]Ibid.

[xx]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), 352; cited in Tibbo, "The Epic Struggle," 316.

[xxi]This assumes that the researcher is not looking for a particular set of materials (of a particular person, family, or corporate body) and that a topical subject request has been presented. However, even searches of the former type may be problematic, if the name being sought is common or if the correct form of the name is not known. This illustrates the importance of other access points such as occupation and function.

[xxii]Tim Hutchinson, Archives and Archivists listserv, "Subject access to archival materials," 14 October 1996.

[xxiii]See, for example, William H. McNitt, "Development of the PRESNET Subject Descriptor Thesaurus," *American Archivist* 52 (Summer 1989): 358-364; Toni Petersen, "Developing a New Thesaurus for Art and Architecture," *Library Trends* 38 (Spring 1990): 644-658; and Jean E. Dryden, "Subject Headings: The PAASH [Public Archives of Alberta Subject Headings] Experience," *Archivaria* 24 (Summer 1987): 173-180.

[xxiv]Elaine Svenonius, "Unanswered Questions in the Design of Controlled Vocabularies," *Journal of the American Society for Information Science* 37 (September 1986): 334.

[xxv]Carol Tenopir, "Full Text Database Retrieval Performance," *Online Review* 9 (1985): 149-164; and R.S. Ro, "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full Text Retrieval I: On the Full Text Retrieval," *Journal of the American Society for Information Science* 39 (1988): 73-78; cited in Jennifer Rowley, "The Controlled versus Natural Indexing Languages Debate Revisited: A Perspective on Information Retrieval Practice and Research," *Journal of Information Science* 20 (1994): 113.

[xxvi]Rowley, "The Controlled Versus Natural Indexing Languages Debate Revisited," 113.

[xxvii]Rowley, "The Controlled Versus Natural Indexing Languages Debate Revisited," 116-117.

[xxviii]See, for example, C.W. Cleverdon, *A Comparative Evaluation of Searching by Controlled Language and Natural Language in an Experimental NASA Data Base*, Space Documentation Service, European Space Agency, 1977; cited in F.W. Lancaster, "The Perspective--Natural Language Versus Controlled Language: A New Examination," *Perspectives in Information Management 1* (London: Butterworth & Co., 1989): "it was found that the natural language searches gave a significantly higher recall and differed little in precision from the controlled term searches. Cleverdon concluded ... that it was the length of the abstract that was largely responsible" (p. 15).

[xxix] Fernanda Ribeiro, "Subject Indexing and Authority Control in Archives: The Need for Subject Indexing in Archives and for an Indexing Policy using Controlled Language," *Journal of the Society of Archivists* 17, no. 1 (1996): 35.

[xxx]Ribeiro, "Subject Indexing," 30; emphasis added.

[xxxi]Ribeiro, "Subject Indexing," 38. This figure was derived from the fact that the average precision for database A and database was calculated to be 43.6% and 57.2%, respectively; see Appendix 4, p. 52. No attempt was made to calculate recall, unfortunately, so this study sheds no light on the effectiveness of topical subject retrieval.

[xxxii]Ribeiro, "Subject Indexing," 40. In particular, the "incremental advantage of A given B retrieval" is calculated to be n/b, where 'n' is the number of records retrieved in database A, but not in database B; and 'b' is the number of records retrieved in database B.

[xxxiii]Ribeiro, "Subject Indexing," 40-41; emphasis in the original.

[xxxiv]David A. Bearman and Richard H. Lytle, "The Power of the Principle of Provenance," *Archivaria* 21 (Winter 1985-86): 42. These issues, and others, have been further explored and developed in articles including David Bearman, "Documenting Documentation," *Archivaria* 34 (Summer 1992): 33-49; Terry Cook, "The Concept of the Archival Fonds: Theory, Description, and Provenance in the Post-Custodial Era," in *The Archival Fonds: From Theory to Practice*, ed. Terry Eastwood, 31-85 (Ottawa: Bureau of Canadian Archivists, 1992); Hugo Stibbe, "Implementing the Concept of Fonds: Primary Access Points, Multilevel Description and Authority Control," *Archivaria* 34 (Summer 1992): 109-137.