Research Article

Authority Control Issues and Prospects

DAVID BEARMAN

Abstract: Research on authority control reported in archival, library, and information science literature suggests that efforts to control topical subject terminology are inappropriate and ineffective in an archival setting because researchers are unlikely to use the same terminology as that contained in the documents, and because most users value precision over recall (inclusiveness) in their searching. The author argues that archival retrieval will be enhanced by placing more emphasis on increasing the number of access points and less on achieving consistency in indexing. He describes various kinds of authority files and identifies several (occupation, time period, geographic coordinates, form-of-material, and function) that offer the most promise. He advocates the use of existing reference files and cooperative development of new ones, to be used not only in the traditional authority-control sense but also as valuable information resources in their own right.

About the author: *David Bearman is the editor of* Archives and Museum Informatics *and* Archives and Museum Informatics Technical Reports *and is the founder of Archives & Museum Informatics, a consulting firm in Pittsburgh, Pennsylvania. He has written extensively on issues related to archival automation and archival information exchange.*

OVER THE PAST FEW years, authority control has received increased attention in the archival community.[1] Experience in local automation applications and involvement in national bibliographic databases has heightened archivists' awareness of both system design and data quality issues. Authority control, as currently practiced in archives and library systems, employs a controlled vocabulary to restrict what is entered in the fields of a cataloging database to values contained in external authority files.

The decision to rely on authority files to impose consistency in the database leads to a series of demanding requirements. Somewhere professionals must create and update the authority files. Indexers, catalogers, and others who assign terms for access points must check look-up authority files to be sure that their proposed terms are recognized, either as "preferred" terms or alternative values. Then the indexers may have to do further research to determine whether or not the persons, places, or things in the item or collection being cataloged are identical to those of like name in the authority file. If valid lists are to be maintained, new terms must be confirmed by research in external sources and any conflicts must be resolved and/or reflected in the database. As a consequence of these requirements, many systems that supposedly employ authority control, especially those dependent on manual updating, break down and thereby fail to deliver on their promise.[2]

Given the substantial intellectual, technical, and administrative overhead involved in authority control, it is not surprising that researchers have sought to determine whether vocabulary control works, how it could be made to work better, what alternatives exist, and if it is the best strategy to improve retrieval. Although the conclusions of the research literature are far from uniform, two findings have emerged with great regularity. First, vocabulary control works best when the terminology of the documents and that of the researchers are highly consistent, that is, when the collections and the use of them are relatively homogeneous.[3] Second, limiting access points to authorized terms generally results in lower recall (fewer finds) and higher precision (fewer false finds). In any case, authority control is less effective overall than the introduction of richer lead-in vocabularies, i.e., additional (uncontrolled) terms that point to the appropriate controlled terms.[4]

Taken alone, these conclusions should discourage archivists from any further efforts to employ authority control for topical subject access points, because neither of the necessary situations exists for subject access to archival information systems: use

[1]Much of this literature is discussed, evaluated, and usefully added to by the authors published in Avra Michelson, ed., *Archives and Authority Control, Archival Informatics Technical Report*, vol.2, no.2 (Summer 1988). Individual contributors are: Jackie Dooley, "An Introduction to Authority Control for Archivists," 5-18; Tom Garnett, "Development of an Authority Control System for the Smithsonian Institution Libraries," 21-27; Marion Matters, "Authority Files in an Archival Setting," 29-33; Avra Michelson, "Descriptive Standards and the Archival Profession," 1-4; Richard Szary, "Technical Requirements and prospectus for authority control in the SIBIS—Archives Database," 41-44; Lisa Weber, "Development of Authority Control Systems Within the Archival Profession," 35-40.

[2]Joseph W. Palmer, "Subject Authority Control & Syndetic Structure: Myth & Realities: An Inquiry into certain subject heading practices and some questions about the implications," *Cataloging & Classification Quarterly*, vol.7, no.2: 71-95; Catherine M. Thomas, "Authority Control in Manual vs. Online Catalogs: An examination of 'See' references," *Information Technology and Libraries* 3 (December 1984): 393-8.

[3]Carol Tenopir, "Full Text database retrieval performance," *Online Review* 9 (1985): 149-164; Tenopir, "Searching by Controlled Vocabulary or Free Text," *Library Journal* 112 (15 November 1987): 58-59.

[4]Katherine W. McCain, Howard D. White, and Belver C. Griffith, "Comparing retrieval performance in online databases," *Information Processing & Management* 23 (1987): 539-553.

of subject terminology, both within archival collections and by researchers, is notoriously heterogeneous; and the archival literature has been fairly consistent in arguing that archival users value recall over precision.[5] This hypothesized failure of topical subject-based authority control has been empirically demonstrated by Avra Michelson.[6] It would follow, then, 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.

It is not necessary to dismiss all types of authority control for all archival purposes; the research literature suggests that certain kinds of authority control, if correctly implemented, can benefit specific types of uses. Rather than emphasize the headings-management aspects of name-authority control, archivists should focus substantially greater efforts on building cooperative reference files for occupation, function, geographic-coordinate, time-period, and form-of-material terms; these will provide access to people, organizations, places, events, and records, respectively.

When Does Authority Control Work?

The first Cranfield experiments and similar large retrieval experiments of the early 1960s demonstrated that simple controlled vocabularies with normalized word endings (dropping suffixes such as *ing* and *ed*) and synonymy performed better than full vocabulary control.[7] Subsequent studies comparing natural language and controlled vocabulary searches revealed differences in retrieval described only as differences, and recommended searching both controlled and uncontrolled terminology. The most influential of these studies reported that full text searches provide better recall but poorer precision.[8]

The most comprehensive review of subject authorities in library systems over the past twenty years, conducted by a partisan of vocabulary control, Elaine Svenonius, recently concluded that subject authority control isn't working.[9] Svenonius assigns the blame for the failure of authority control to a combination of what she calls "intrinsic variables," such as the quality of the vocabulary and the nature of the discipline that it represents, and "external variables," such as the skills of indexers and searchers and the criteria for retrieval evaluation. She particularly finds that the problem lies in lack of distinction between types of controlled vocabularies and concludes that "it would seem that the vocabulary of a discipline should precede any attempt to control that vocabulary for information retrieval."[10] Her conclusions are not unlike the findings of Bhattacharyya, and the earlier Cranfield II experiments, which demonstrated that the greater the terminological consistency in the field, the greater the benefit of vocabulary control.[11]

Subject based authority control in archives also fails in applications because

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

[6]Avra Michelson, "Description and Reference in the Age of Automation," *American Archivist* 50 (Spring 1987): 192-208.

[7]Cyril W. Cleverdon, *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. (Cranfield, England: College of Aerodynamics, October 1962).

[8]Karen Markey, Pauline Atherton, and Claudia Newton, "An analysis of controlled vocabulary and free text search statements in online searches," *Online Review* 4 (1980): 225-236.

[9]Elaine Svenonius, "Unanswered Questions in the Design of Controlled Vocabularies," *Journal of the American Society for Information Science* 37, no.5, (September 1986): 331-340.

[10]Ibid., 336.

[11]K. Bhattacharyya, "The Effectiveness of Natural Language in Science Indexing and Retrieval," *Journal of Documentation* 30 (1974): 235-254; Cyril W. Cleverdon, Jack Mills, and Michael Keen, *Factors Determining the Performance of Indexing Systems*, 2 vols. (Cranfield, England: College of Aeronautics, 1966).

subject analysis of archival materials is extremely problematic. Archival material does not have a *subject* per se. Archival material is *of* the activity that generates it, but seldom is it consciously authored to be *about* something. For instance, the records of elementary school matriculation are often used to answer demographic questions about immigration, family size, or life-expectancy, but this is not their "subject." Nor is their subject "elementary education," although this term is most often assigned. Their subject, insofar as they have one, is the registration procedures of governmental agencies, but the data they contain will shed little light on these procedures. 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.

If the research literature leads archivists away from control of subject terminology in favor of other access points, its lessons do not end there. Archivists have typically implemented authority control with the aim of increasing consistency in a database, because greater consistency would presumably improve retrieval. Studies of library systems show, however, that when implemented, authority control doesn't improve retrieval overall as much as simply assigning more terms. Ann Schabas found that adding natural language from titles improved recall by either PRECIS or Library of Congress Subject Headings (LCSH) access; oddly, though, the two very different approaches yielded otherwise similar results for retrieval effectiveness.[12] Anne Piternick demonstrated that end-user thesauri, rather than thesauri used by indexers, could dramatically improve access by user terminology.[13] Together, these studies confirm a long known information retrieval principle that, the more terms assigned to a document, the better the chances of its being retrieved.

It seems that when controlled vocabularies augment retrieval precision, they do so because they increase the size of the lead-in vocabulary by providing more terms that can match user queries and point to records in the database. Archivists should note that a vocabulary can play this role, whether or not it is also used to "control" headings in the database. They can capitalize on these findings by emphasizing cross-references, broader and narrower terms, and synonymy in the development of vocabularies and focus their attention on the construction of "front-end" systems, whether manual or automated, that assist users to expand their vocabulary and refine their search terminology appropriately. Archivists should add terminology and enrich the links between terms rather than attempt to validate terms, increase inter-indexer consistency, or enforce rules for choice of headings. But first, they need to determine which access points to focus on, and what kinds of authority lists are most appropriate to each.

Authority Control Issues in Archives

User Queries and Access Points. Logically, user queries should be the point of departure for defining a strategy to augment access. Incredibly, archivists have no published literature of user-query analysis with which to begin. They do not even have a research literature that reports empirically on user expectations in use of archives.

Existing archival literature has been fairly consistent in declaring that users of archives value recall (inclusiveness) over

[12]Ann H. Schabas, "Postcoordinate retrieval: A Comparison of Two Indexing Languages," *Journal of the American Society for Information Science* 33, no.1 (1982): 32-37.

[13]Anne B. Piternick, "Searching vocabularies: A developing category of online search tools," *Online Review* 8, no.5 (1984): 441-449.

precision, but no empirical evidence has been offered.[14] If this is true, authority control will not work in archives since studies regularly show that authority control either has no effect or decreases recall. Until this assumption is shown empirically to be wrong, however, personal experience would suggest that, with the exception of genealogists and biographers, archival visitors are interested in seeing exemplary, rather than exhaustive, documentation. Furthermore, we know that archivists are themselves the largest users of archives, followed closely by records creators,[15] both of whom almost certainly value precision over recall, whether they are seeking a known item or a report of all instances of a specific kind of record.

Therefore, authority control over some access points might well contribute to archival practice. To identify the most appropriate access points we should look more closely at users' questions. What kinds of terms appear in queries? How specific or general are they? Do they invoke a place, a time, a type of record, a person, or a function that generates documentation? Are our users seeking archival records or are they more interested in information that would be in our *authority* files rather than in *bibliographic* records? How sophisticated are they about the usages contemporary to the period they are researching?

In the absence of appropriate research on user queries, we can only look at each potential archival access point and consider the extent to which authority control for each might improve a variety of different kinds of retrievals undertaken for different purposes. Through a logical process, we need to determine what access points appear to offer the greatest promise for authority control and then establish criteria by which to decide whether they should, or should not, be controlled.

Potential Vocabularies. If we begin with the National Information Systems Task Force (NISTF) data dictionary and the fields of the MARC AMC format, we can readily identify only fifteen fields other than topical subjects that are: (1) likely to be searched by a researcher, or (2) likely to be used as the basis of an administrative report.[16]

These fields include:

action	corporate name
creator	event
form	function
genre	geographic name/ place
language	medium
method of action	occupation
personal name	relator
status	

We know that vocabulary control performs best for retrievals in which high precision is valued, as in the case of day-to-day administrative retrievals to support activities such as records scheduling or appraisal. In this, as in most administrative retrievals, finding a relevant precedent is more important than an exhaustive search. One of the arguments advanced in favor of controlling form-of-material and function vocabularies is that adopting controlled vocabularies for form-of-material and function will support access to records across jurisdictions for which similar appraisal

[14]Pugh, "The Illusion of Omniscience;" Michelson, "Description and Reference."

[15]Paul Conway, "Research in Presidential Libraries: A User Survey," *The Midwestern Archivist* 11 (1986): 35-56.

[16]National Information Systems Task Force, "Data Dictionary," comp. and ed. David Bearman, in Nancy Sahli, ed., *MARC for Archives and Manuscripts: The AMC Format* (Chicago: SAA, 1985). The list of candidates for authority control does not differ greatly from that which would be proposed for library materials, except that a librarian would immediately notice that it does not include title, or such absolute identifiers as ISSN or ISBN and LC Card Number. Indeed, the absence of transcribed titles or unique extra-institution identifiers for unpublished materials is one strong argument for authority control in some other access points.

considerations will apply.[17] Unfortunately, the vocabularies we have for these access points are still underdeveloped. They need to be enriched with substantially more lead-in vocabulary and then tested.

Another area in which controlled vocabularies might be powerful is the retrieval of names of entities—persons, corporations, events, and geographical places—that are unambiguously involved in the creation of records. Admittedly the names of people involved in creation of archival records are less well known than the names of authors of publications, and lists such as the Library of Congress Name Authority file barely begin to meet the needs of archivists; still, a collocation of individuals who are involved in the records creation process is a worthy goal. We also know that researchers will often approach archives seeking persons with specific biographical characteristics, such as occupations, places of birth, or political or religious affiliations, rather than people with known names. For this reason, archivists need to pay greater attention to building biographical reference files than to aligning the form of name used for an individual through vocabulary control and headings management.

Corporate names are even more problematic, because archival corporate entities come and go with astonishing frequency. Collocating corporate entities is again a worthy aim, but archivists will be less well served by investments in authority control over the names of organizational sub-structures, such as divisions, departments, and task forces of corporate entities, because those names are only meaningful in their particular organizational context. Such names cannot be used across organizations to locate similar functions or records, which correspond to the kinds of research questions brought by users. Controlling such terms fails, in the same way that efforts to assign concrete "levels" to organizational units fails, to support the intellectual move "from syndetic structure of records to syndetic structure of entries."[18] This is what happened in the design of SPINDEX. There is no inherent property of organizations that makes a third tier unit in one corporation equivalent to a third tier unit in another organization (or even another third tier unit in the same organization). Nor are there properties in records systems that dictate that a subdivision of one record system will resemble that of another in its scope.[19]

Even though events, ranging from groundbreakings to wars, are important in the creation of records, events are known by many different names, and any significant event is comprised of numerous lesser events, demanding significant research. Lists of named events are not readily available. Instead of using names as an access point, archivists would be wise to exploit an attribute of events that can be represented uniformly by indexers and users: *when* they took place. Archivists should pay greater attention to chronological access, both to events authority files that use date ranges and to records that use named time periods.

While users bring many geographical perspectives—geological, geocultural, and geolinguistic—to archives, geopolitical terminology is shared by users and records-

[17]David Bearman, "'Who about what' or 'From Whence, Why and How': Intellectual Access Approaches to Archives and the Implications for National Information Systems" in *Archives, Automation & Access*, ed. Peter Baskerville and Chad M. Gaffield, (Victoria, BC: University of Victoria, 1986); David Bearman and Richard Szary, "Beyond Authorized Headings: Authorities as Reference Files in a Multi-Disciplinary Setting," in *Authority Control Symposium*, ARLIS Occasional Papers #6, ed. Karen Muller (Tucson: Art Libraries Society of North America, 1987), 69-78.

[18]Robert H. Burger, *Authority Work: The Creation, Use, Maintenance and Evaluation of Authority Records and Files* (Littleton, CO: Libraries Unlimited, 1985), 7.

[19]Max Evans, "Authority Control: An Alternative to the Record Group Concept," *American Archivist* 49 (Summer 1986): 249-261.

creating institutions. For terminology that differs depending on the disciplinary perspective of the users, means are now being developed to provide graphical interfaces that dramatically increase the power of lead-in vocabularies to identify overlapping geographical regions despite different user vocabularies.[20] Coordinate-based searching, which lies at the heart of these graphical approaches, is based on defining the geographic coordinates of terms so that a variety of types of geo-terms from different disciplinary perspectives can be transformed into coordinate expressions. For example, because the coordinates of a town may fall within the coordinates of a valley, researchers studying settlement of the valley should also retrieve items indexed by the name of the town. Geographical coordinates also provide ways to solve problems created by the movement of geo-names over time.[21] Overall, geographical access and control over the coordinate expression of geographical locations offers considerable promise.

Types of Authorities & Implementations

Correct identification of fields that are candidates for authority control in archives is only the first step. Once fields that might be controlled are selected, archivists need to identify the type of authority file most appropriate to each field and the most strategic implementation for each.

Authority files are defined by three salient characteristics: type of file (term list or reference file), presence or absence of structure, and presence or absence of scope notes. The combination of these variables yields the following eight types of authority files:

- *term lists*, with and without scope notes
- *term lists*, with and without syndetic structure
- *reference files*, with and without scope notes
- *reference files*, with and without syndetic structure

Term lists are lists of single words or compound terms. In addition to the authorized terms themselves, term lists may include data about the entities referred to when such additional information is required to *uniquely identify* the entity (birth dates to prevent confusion between persons with the same name, for example). Library name authority files and subject headings are term lists.

Reference files are databases in which data supplied for authorized terms goes beyond what is required to distinguish between like terms. Thus, a reference file for persons might include educational affiliations and degrees, honors and awards, and important life events. Reference files for events would name participants, discuss consequences, define the time of occurrence, and identify related events.

Scope notes are not about the entity named by the term, but about how the term is used by indexers. They define the conditions under which an authorized term may be assigned as an access point within a given application.

Lists in which the terms are arranged in an order other than one based on their meaning, such as alphabetical, or chronological, lack syndetic structure. *Syndetic structure* refers to hierarchical and other associative relationships between terms, reflecting their linguistic significance. Syndetically structured lists are sometimes called taxonomies or thesauri; they may or

[20]James Ross, "Geographic Headings Online," *Cataloging & Classification Quarterly* 5 (Winter 1984): 27-43.

[21]Oreste Signore and Rigoletto Bartolli, "Controlling Geographic Descriptions: A Case Study for Historico-Geographical Authority," *Proceedings of the International Conference on Terminology Control for Museums*, 22-24 September 1988, (Cambridge, Museum Documentation Association, in press).

may not include scope notes and may or may not incorporate information that goes beyond what is required to distinguish one authorized term from another.

The fifteen MARC AMC fields that might provide access points do not all require the same kind of authority control. Several, such as action, language, medium, method of action, relator, and status, might be adequately controlled by relatively short term lists without scope notes or syndetic structure; such lists are often called *value tables*. The criterion for determining if a value table will suffice is whether users can achieve high precision searches when the system they are using shows them their choices. Generally, short lists of unambiguous and discrete concepts are adequately controlled by value tables.

Other access points, such as events and geographical places are best retrieved by dates and geographical coordinates which are controlled by *conventions*. Dating conventions translate different ways of expressing the same date into a common representation. If a searcher can translate a query into dating conventions, or the system can do so automatically, then the name of the event need not be controlled in an authority. Similarly, if a geographical or geological term can be looked up in an authority file that translates it into coordinates, and the database can be searched on coordinates, then control over the terms used to describe geographic places is exercised by expanding the lead-in vocabulary in the authority file, rather than by restricting the values in the database. Conventions can also be useful to achieve greater consistency between indexers and users in formulation of personal and organizational names, although name formulation conventions, such as those dictated by AACR2, will not assure anywhere near the degree of commonality that dating conventions or geographical-coordinate-based location identifiers do.

The remaining data elements, form, genre, function, and occupation, are attributes of authority files for records, corporate entities, and persons. They have more open-ended vocabularies than those subject to value-table control, and they do not lend themselves to stylistic conventions. To decide whether users would benefit from the control of these elements, we need to distinguish among a variety of implementation choices that determine how authority control will actually operate within a given system.

Authority controls can be implemented in many ways, but three fundamental distinctions may help archivists to appreciate the way implementation affects success, and the reasons why authority control is perceived differently by users of different systems:

- Some systems employ substitution of terms and others do not. *Term substitution* replaces values entered by users in data entry or in searching with authorized values from an authority file so that only authorized values reside in the database and/or are searched.
- Some systems use pre-coordinated terms while others assume post-coordination.
- Some systems assume one consistent authority source; others permit multiple, inconsistent authorities to coexist.

Term Substitution. Most archivists are best acquainted with the typical library authority control system, which is designed for the management of pre-coordinated headings from a single authority source and is experienced as a substituting implementation in data-entry mode only. In this type of system, the only terms permitted in authority-controlled fields of bibliographic records are the authorized versions. Alternative vocabulary is posted only in the authority file. Users searching on an unauthorized term must know to look in the

authority file to retrieve records that will match on authorized terms. In data entry, on the other hand, the authorized term replaces any unauthorized term entered in the bibliographic database.

This implementation has several disadvantages. For staff describing archival records, the same choice of term is imposed regardless of the term that was current at the time the records were created. Thus distinctions that were significant to contemporaries are often lost. At the same time, researchers who are sensitive to such differences and employ terms from the period of the records or the vernacular of the region will not automatically be shown records that satisfy their query; they must know to look in the authority file for appropriate related terms. Concepts with different meanings in a number of disciplines cannot be controlled by authority files unique to specific domains because multiple and potentially conflicting authorities are not recognized.

Implementations that do not substitute terms at data entry are better suited to archives. Even though data-entry term replacement is the simplest implementation tactic for imposing commonality on language, it violates subtleties of terminology by enforcing a "preferred" term, erasing differences in regional language and changes in the meaning of terms over time. In the national library system, users are expected to search for books with a vocabulary from our own time. This strategy makes sense when searching for contemporary non-fiction books and articles (fiction is not "subject" indexed). But in archives, where users are seeking to locate evidence of precisely the kinds of subtle social and cultural shifts reflected in the usages of the past, data-entry term substitution is not an acceptable strategy.[22] In systems that "switch" terms on searching, the language actually in the database is the natural language of the record, for example, the functions of a political office as they were described at the time the records were created, even if such functions are now obsolete or subsumed by different functions in our contemporary parlance. If a user employs any equivalent term in searching, the term used in the query will look up its variants in the authority file and switch to any of them so as to collocate records with like indexing concepts. Archivists need to design search-term substituting authority systems in which the indexer puts in what is in the record, the user puts in what is in his or her head, and the system finds all the appropriate matches and gives the user what he or she wants.[23] When vocabulary control is implemented without data-entry term replacement, retrieval is heavily dependent on the use of vocabularies that are rich in lead-in terminology. Fortunately, many designers of online library catalogs are also prescribing such robust front ends.[24]

Coordination. A similar failing of library catalogs for archives becomes apparent when we examine the pre-coordinated headings used in the Library of Congress Subject Headings (LCSH). Despite many failings exhaustively documented by librarians,[25] LCSH serves relatively well for

[22]David Henige, "Library of Congress Subject Headings: Is Euthanasia the Answer?" and "Library of Congress Response," *Cataloging & Classification Quarterly* 8: 7-19.

[23]Michael Gorman in Mary W. Ghikas, ed., *Authority Control: The Key to Tomorrow's Catalog. Proceedings of the 1979 Library Information Technology Association Institute* (Phoenix: Oryx Press, 1982).

[24]Marcia Bates, "Subject Access in Online Catalogs: A Design Model," *Journal of the American Society for Information Science* 37, no.6 (1987): 357-376; Charles R. Hildreth, *Online Public Access Catalogs: The User Interface* (Dublin, OH: OCLC, 1982); Walt Crawford, *Patron Access: Issues for Online Catalogs* (Boston: G. K. Hall, 1987).

[25]Pauline A. Cochrane, *Redesign of Catalogs and Indexes for Improved Online Subject Access: Selected Papers of Pauline A. Cochrane* (Phoenix: Oryx Press, 1985), 475 pp.; Karen Markey and Francis Miksa,

books because books are written with a consistent level of specificity and have a predominant subject matter. Thus, we may find books about relatively specific topics, such as a biography of a person or a history of a village or company, and we may find books about relatively global concepts like pollution, ideology, or authority control. In either case, a subject heading constructed in the manner of Library of Congress headings, such as *topical subject—place—time period*, will place the book into a subject category with similar materials.

A single book does not contain materials relating to the history of viticulture in California, damages caused by volcanoes in Turkey, *and* evidence of labor union pension frauds in the way that the records of a civil court might. Also the context of the creation of the book and its form are not relevant to the meaning of its text in the way that contexts of creation and form-of-material are relevant to understanding archival records. In our hypothetical civil court records series, one archival researcher may be interested in depositions, another in the acts of a particular judge, and a third in legal arguments. If archivists adopt pre-coordinated strategies, they are likely to select terms to describe records that will not match the unpredictable perspectives of users.

Implementations need to support searches for discrete facets of post-coordinated headings so that users can look for broader and narrower terminology on any dimension, exploiting facilities provided by hierarchical thesauri. Users could then search independently on different facets in order to hone their search results. Even though some library catalogs provide for character-string searches or component-term searching of pre-coordinated headings, they lack the functionality to broaden and narrow searches independently by facet. This capability is essential in order to realize the expectations of the architects of the newly adopted MARC field 654, designed for headings composed of Art and Architecture Thesaurus (AAT) and Medical Subject Headings (MeSH) terms. Such an implementation requires new systems functions, however, because it is highly unlikely that a user would otherwise invoke a term constructed of precisely those facets selected by an indexer without substantial assistance.[26]

Multiple, Independent, and Conflicting Authorities. It is possible to have several discrete authority files, reflecting different perspectives and disciplines, linked to a single field. Elsewhere, Richard Szary and I have advanced a number of theoretical arguments in favor of such databases, which we call Cultural Information Systems, and Szary has since expanded on the requirements for an archival authority control system, which supports such multiple independent authorities.[27] In subsequent articles, I have discussed the problems of implementing multiple independent and potentially conflicting authorities in a logical design for an art historical database, and acknowledged the difficulties facing users of a system in which all knowledge, and not just specialist knowledge, is subject to conflict.[28] Nevertheless, implemen-

"Subject Access Literature, 1986," *Library Resources and Technical Services* 31 (October 1987): 334-54.

[26]For greater context to this debate, see the report of discussions following papers presented at the AAT session at the ARLIS/NA meeting in 1987 cited in *Archival Informatics Newsletter* 2, no.1 (Spring 1988): 7; David Bearman and Toni Petersen, "Searching Databases Indexed Using the AAT," submitted to *Art Documentation*.

[27]Bearman and Szary, "Beyond Authorized Headings;" Szary, "Technical Requirements and Prospectus." See also Richard Szary, "Design Requirements for Archival Authority Systems," (Paper delivered at the fifty-second annual meeting of the Society of American Archivists, Atlanta, 2 October 1988).

[28]David Bearman, "Buildings as Structures, as Art, and as Dwellings: Data Exchange Issues in an Archi-

tations supporting multiple independent authorities are well suited to use in a setting in which users approach the records with such heterogeneous interests, and they are attractive to archives because they open up the possibility of importing authority files from a variety of external contexts.[29]

Where Should We Go From Here?

The shortcomings of library-system-based authority control for retrieval of archival materials should encourage us to look to other options. When we realize that we need to focus more on the benefits of expanded lead-in vocabulary and less on achieving consistency of indexing, we will be attracted to reference files instead of term lists and will place greater value on the benefits of authority control from the searchers perspective as a means to identify alternative access points than on its value for headings management.

Reference files enable us to extend the functions of authority control beyond headings management because they contain information beyond that required to distinguish terms from each other. The concept of extending term lists into reference files grew out of comparing the practices of catalogers involved in headings management and those of researchers compiling scholarly databases. The sole function of the library catalogers' independent authority files for people, organizations, and geographical places is to control the terms in the headings of bibliographic records. Catalogers may build substantial value-added databases, but users can only exploit them for the limited purpose of heading validation. By recasting such bibliocentric databases using a relational data model, it is clear that records, persons, organizations, events, and places should occupy separate files, with defined relationships between the files, which support researchers whose information needs can be satisfied by data from any of these sources.

Because reference files contain information beyond that required for headings management, they can act in authority control roles without data-entry term substitution. Each record includes all the variant terms that might be used in place of the authorized name, thereby serving as the locus of lead-in vocabulary, and, at the same time, each record contains numerous dimensions that can lead users out of the reference file, to other reference files, and to archival records. For example, a record about a person leads out to organizations with which that individual was affiliated, the place of the person's birth, others with whom he was associated, and his publications and talks. By searching for records of others with whom the subject of our query was affiliated, we may find information about that subject that would not have been indexed in even a detailed description of the other records.

The most common reference-authority files in cultural repositories contain data about persons or organizations. Extensive reference files are also created around records, whether accessioned as holdings, scheduled for appraisal, or destroyed. Other reference files include data about cultural, political, and intellectual events. Another contains information about spaces, whether these are buildings, archaeological sites, locales of collecting, or geological, geographical, and geo-cultural places. It is critical for archivists to recognize that many of the terms that are used as access points for records, such as roles, functions, and occupations, are attributes of entities other than records, and that these must be linked to bibliographic files.

tectural Information Network," in *Databases in the Humanities and Social Sciences-4*, ed. Lawrence J. McCrank (Learned Information: Medford, NJ, 1989), 41-48.

[29]Carol A. Mandel, *Multiple Thesauri in Online Library Bibliographic Systems. A Report prepared for the Library of Congress Processing Services* (Washington, DC: Library of Congress, 1987).

The entities about which reference files are constructed are real-world things such as people and events, which are of considerable interest to non-archivists as well as being useful ways to gain access to records. Therefore, the reference files built by cultural repositories are valuable databases to others. For example, organization history reference files built by archives contain information about the authority, mission, structure and function of organizational units that is essential to archival administration but also represents a valuable resource both for the organization itself and for outsiders. B.A.G. Fuller of the Mystic Seaport Museum reports: "the concept of separating information into entity files and authority files [is] immensely useful and versatile... We are using the format developed for our vessel authority file as the basis for a union list of all watercraft in American museums."[30] In the process of providing access to the union list, Fuller is constructing a reference file on historical watercraft that is a historical database in its own right.

Not all reference files used in archival information systems need to be created de novo by archivists. Reference files in an archival information system may be the primary databases of the discipline or organization that created them. Archivists and the builders of other cultural information systems only need identify databases that contain information which could be linked to records, and then import such databases into their systems. Thus, geographical reference files could come from a geological survey, corporate data from the SEC or the local Chamber of Commerce, data about individuals from *Who's Who*, and data about the U.S. Government from the *Federal Register*.[31] Because reference files can be acquired from other disciplines, especially from the computerized databases of the parent organization of the archives, employing a deep network of reference files for one application need not be prohibitively costly. It does, of course, require implementations that can support multiple, independent, and conflicting authorities.

It is time to implement a database of independent reference files supporting archival description and information retrieval. Several years ago, Richard Lytle and I described the appropriate relationship among such files.[32] Max Evans subsequently elaborated how data about organizations could be converted into authority records that exercise control over provenance data elements in the description record.[33] For several years now the participants in the Research Libraries Information Network have been elaborating a vocabulary on functions of organizations.[34] It is time to put these together with the appropriate retrieval system.

The major reference file in this model that is still missing is that for "form-of-material." Ever since the NISTF data dictionary introduced the concept, it has been difficult to explain the concept of a cultural abstraction corresponding to a kind of record without encountering skepticism about whether such a property of records truly exists. Similar properties have been advanced by students of diplomatics, but until recently these have not been carried for-

[30]B.A.G. Fuller, personal letter to author. See also "In Search of A System: A Report on Mystic Seaport Museum's Experience as it Begins Computerization of its Collections," *Spectra* 15, no.3 (Fall 1988): 12-14.

[31]David Bearman, "The National Archives and Records Service: Policy Choices for the Next Five Years," *For the Record*, December 1981.

[32]David Bearman and Richard H. Lytle, "The Power of the Principle of Provenance," *Archivaria* 21 (1986): 14-27.

[33]Evans, "Authority Control: An Alternative."

[34]David Bearman, "Archives and Manuscript Control Within Bibliographic Utilities: Challenges and Opportunities," *American Archivist* 52 (1989): 26-39; Kathleen Roe and Alden Monroe, "The Role of Function in Archival Practice," (Fall 1988, unpublished).

ward into modern records.[35] Recently, research by information scientists studying electronic records has confirmed the existence of document formalisms of the kind suggested by the concept of form-of-material.[36]

The underlying argument is that archivists and historians employ rules for interpreting the probable content of documents based on formal properties of records. For instance, we expect that marriage certificates will provide information about the birth place of parents, thus serving as a source of information about migration patterns; and we expect military enlistment records to include health and mental profiles, serving as evidence of the population distribution of diseases and intelligence. We recognize the form-of-material of a record quite independently of its content, thus knowing immediately when we see a memorandum, an award, or an order form, without having to read its contents.

Research into the nature of documents is now demonstrating not only that "document formalisms" (structural features of records that signal their contents to the culturally attuned) exist, but also that machines can be taught to distinguish between document types. Computers can parse documents for their internal components and "mark" them with such document-marking languages as Standard Generalized Markup Language (SGML), creating a sort of electronic "fingerprint" of a form-of-material.[37] Some of the elements of these files are now becoming clear. They look a bit like records schedules without dates or names of offices; they contain a field for SGML-like "fingerprints" and fields for data elements typically found recorded in this type of record. Further work must be done before we can construct useful form-of-material reference files, but what makes this work exciting to archivists is that, if we are to achieve reasonable quality retrievals from huge full-text databases, it is critical to be able to limit searches by form-of-material and internal structural components.[38]

Such fingerprints might also form the controlled vocabularies that link reference files of document types to databases of archival records. Researchers knowing only that they are interested in smallpox vaccination might match their search on terms in the description of school matriculation and military conscription records in the form-of-materials reference file and navigate across the SGML fingerprints to records in the archives matching the formal type.

Conclusions

The research literature in library and information retrieval systems suggests what empirical research in archives has confirmed, that authority control over topical subject terminology in archives is ineffective. Other access points, including form

[35]Luciana Duranti, "Diplomatics: New Uses for an Old Science," *Archivaria* 28 (1989): 7-27.

[36]David M. Levy, Daniel C. Brotsky, and Kenneth R. Olson, "Formalizing the Figural: Aspects of a Foundation for Document Manipulation," (Systems Sciences Laboratory: Xerox Palo Alto Research Center, 31 August 1988, unpublished).

[37]Michael B. Spring, "Copymarks," (SLIS Department of Information Science: University of Pittsburgh, September 1988, unpublished).

[38]David C. Blair, "Full-text Retrieval: Evaluation and Implications," *International Classification* 13 (1986): 18-23; Jung Soon Roo, "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. I. On the Effectiveness of Full-Text Retrieval," *Journal of the American Society for Information Science* 39, no.2 (March 1988): 73-78; "II. On the Effectiveness of Ranking Algorithms on Full-Text Retrieval," ibid., 39, no. 3 (May 1988): 147-160; T. Saracevic, P. Kantor, A. Y. Chamis, and D. Trivison, "A Study of Information Seeking and Retrieval: I. Background and Methodology," ibid., 39, no. 3 (May 1988) 161-176; T. Saracevic and P. Kantor, "A Study of Information Seeking and Retrieval: II. User, Questions, and Effectiveness," ibid., 117-196; T. Saracevic and P. Kantor, "A Study of Information Seeking and Retrieval: III. Searches, Searches and Overlap," ibid., 197-216.

and genre, function, and occupation, appear to be more promising. These access points are attributes of records, organizations, and people, and could be recorded in authority files for those entities if archivists expanded the concept of an authority file from a term list to a reference file. If such reference files were implemented with searcher-term switching rather than data-entry-term switching, researchers would get the benefits of authority control without sacrificing the historical accuracy of descriptions. Some concrete suggestions can be made for the structure of person, organization, and records files in relational databases in which each file serves as an authority to the others.