The Power of the Principle of Provenance

by **DAVID A. BEARMAN** and **RICHARD H. LYTLE**

The task of managing information in organizations is becoming more challenging as the organizations become larger and more complex, and as information technologies and general societal developments increase the volume and sophistication of available information. This task can best be met by the careful study of how these organizations create, use, and discard information. A *practical* understanding must be gained of organizations as living cultures or organisms which create and use information; upon this foundation, sound information management can be developed.

How many archivists will recognize the preceding comments as relevant to them and the archival profession? Do archivists have much to contribute to the management of information in large organizations? Will archivists bring their knowledge of how organizations create and use information to bear on modern information management problems? Will the archival profession consequently make a transition to the modern information culture, or will it remain behind as a keeper of paper and electronic relics?

The key to the archivists' contribution to information management lies in their unique perspective provided by the principle of provenance as it concerns organizational activity, especially how organizations create, use, and discard information. Despite the insights provided by provenance, however, archivists have not exploited its potential for retrieval in traditional archival applications, and have not even attempted its wider application to the management of all information within their organizations.

This article offers a critique of the application of the principle of provenance in traditional archival environments and proposes its expansion in a more powerful application to information management. The article also advocates a much more aggressive leadership role for archivists in the wider management of information resources.[1]

1 During its lengthy gestation period, many persons helped by commenting on the paper. Alice Prochaska, Eddy Higgs, and John Walford of the Public Records Office provided comments back in 1982. Several interested staff of the U.S. National Archives helped about the same time, at least once in a luncheon session called for the purpose. More recently, Terry Eastwood at the University of British Columbia, Tom Nesmith, General Editor of *Archivaria*, and Terry Cook, former General Editor of *Archivaria*, provided very useful critique. Fred Stielow, University of Maryland, provided some very useful comments on a late draft. We are especially appreciative of the interest shown by our Canadian colleagues. Their encouragement motivated us to complete the paper.

1. Theoretical Background

In work over the past several years, the authors have noted increasingly persistent problems about the distinctive value of provenance for the retrieval of archival materials. Lytle, comparing the power of provenance as a retrieval tool with library-oriented content-indexing techniques, described weaknesses in provenance as archivists use it in retrieval, while pointing out its greater potential if more rigorously applied.[2] Bearman has been systematically defining data elements and information flows in archival information systems, and has found that some of the problems in archival retrieval systems result from a failure to distinguish between provenance information about organizations and descriptive data about the records themselves.[3] In the course of writing and distributing several drafts of this paper to archival colleagues, it became apparent that problems with traditional provenance-based arrangement and description as a tool for retrieval were widely perceived, but that neither the sources of the problems nor solutions had emerged.

Exploration of provenance as the predominant means of archival retrieval reveals it as much more than a principle for the arrangement and description of archival materials. The more aggressive application of provenance to retrieval is apparent in Lytle's earlier definition:

> The Provenance or P Method is the traditional method of archival retrieval, based on principles of archives administration and reference practices of archivists. Subject retrieval in the P Method proceeds by linking subject queries with provenance information contained in administrative histories or biographies, thereby producing leads to files which are searched by using their internal structures. Information in the pure or theoretically defined P Method derives only from what is known about the file — the activities of the creating person or organization and the structure or organizing principles of the file itself.[4]

The process of provenance-based retrieval requires expansion beyond previously published explanations. It is familiar to most archivists who perform reference service for the archives of large organizations. A user poses a subject question which the archivist (assuming no previous knowledge of relevant records) retrieves by relating the subject to the activities of the organization. That is, the archivist translates a user's subject query into the terms of organizational activity. Then either the records or their inventories are searched for information pertinent to the subject query, using the file classification structures created by the originating office and recorded by the archivist in container lists

2 Richard H. Lytle, "Subject Retrieval in Archives: A Comparison of the Provenance and Content Indexing Methods," (Ph.D. thesis, University of Maryland, 1979); "Intellectual Access to Archives: I. Provenance and Content Indexing Methods of Subject Retrieval," *American Archivist* 43 (Winter 1980), pp. 64-75; "Intellectual Access to Archives: II. Report of an Experiment Comparing Provenance and Content Indexing Methods of Subject Retrieval," *American Archivist* 43 (Spring 1980), pp. 191-207.

3 NISTF Working Group on Data Elements and Formats for Archival Information Exchange, "Data Elements for Archives and Manuscript Repositories: A Dictionary, Thesaurus and Format for Information Interchange," Society of American Archivists, National Information Systems Task Force, Washington, D.C. (February 1982); "Data Elements Used in Archives, Manuscript and Records Information Systems: A Dictionary of Standard Terminology," SAA, NISTF (October 1982).

4 Lytle, "Intellectual Access to Archives: I," *American Archivist*, pp. 64-65. The definition of the content-indexing method, the second method in Lytle's 1979 study, has been omitted here, since the emphasis of this paper is on the provenance method.

or the like. An example reference question at the Smithsonian Archives might be the following: "What do you have on the design and construction of large telescopes?" Based on research into the administrative history and functional mandates of the organization, the archivist knows that only the Astrophysical Observatory has such projects in its charge, but the records of the Observatory are voluminous; the scope of records to be searched can be reduced, however, by further careful selection of organizational subunits which might have created the desired information. But the archivist also knows that the Assistant Secretary for Science has staff assistants who sometimes get deeply involved in such projects, and so those records are identified as well for searching. Conversely, the records of top Smithsonian officials would not be good candidates for review unless the searcher was interested in the internal or external politics of scientific construction projects. As inventories and records are examined, further information is found about Observatory activities as well as information directly pertinent to the query.

Lytle has argued that the transformation of subject queries into organizational activity statements is an *inferential* process, in that the archivist infers from provenance information which organizational units might have undertaken relevant activities and therefore might have produced documentation pertinent to the subject query at hand.[5]

2. Provenance Beyond Hierarchy: Expanding the Archival View of Organizations

The provenance method of retrieval obviously rests on a detailed understanding of both the structure and processes of the organizations which created the records in question. The most important question, then, is how adequate is the North American archival view of organizations?

Archival theory has been strongly influenced by the nineteenth-century view of organizations.[6] Classical organizational theory assumes that the typical organization is autonomous and sovereign. At the highest levels, the organization's actions and the structures it produces are assumed to be the result solely of internally formulated policy. Even if this view were valid for a simpler time, it is far too simplistic for modern organizations operating in a world of multi-national corporations, inter-governmental units, regulatory organizations, and federal programmes administered by state, provincial, and local governments. Other assumptions of classical bureaucracies are also violated as the internal workings of modern organizations are explored. In the ideal organization, decisions were supposedly made at one level, implemented at the next. In the modern world of task forces and committees, staff roles and sub-contracting, this seemingly simple structural relationship is in reality immensely complex. On organization charts this complexity is indicated by dotted lines, influence arrows and circles, two-way authority links, and other shorthands which represent a host of non-hierarchical relationships. Management by consensus, collegial relationships, professional boundaries and rights, job responsibilities limited by union contracts, independent ombudsmen, or central agency arbiters further complicate these relationships.

In short, the classical view of organizations emphasizes the importance of hierarchy, in a theoretical world where a given bureaucratic unit is directly subordinate to no more than one higher unit. That kind of hierarchy is called a *mono-hierarchical* structure by information scientists, and its application to organizations emphasizes the chain-of-command dimension of organizations. Mono-hierarchy is thus a poor model for understanding modern organizations.

The inability of mono-hierarchical systems to capture the complexity or large organizations can be illustrated by organization tables in *The United States Government Manual*.[7] The example reproduced here (see chart) is one of the federal government's most traditional organizations, the Department of the Army; the reader should note that it requires six footnotes to clarify distinctions between reporting authority, supervision, advisory roles, and policy roles. Two other example organizations, for which charts are not provided, also defy representation in traditional organization charts. The Smithsonian Institution identifies no fewer than eighteen advisory boards and commissions and three separate boards of trustees with undefined relationships to the Smithsonian Board of Regents and the Secretary of the Institution. The National Foundation on the Arts and Humanities exists only as a theoretical construct surrounding three independent Councils with imprecise relationships to each other and undefined responsibility to the President of the United States.

One striking characteristic of these examples is the virtual absence of hierarchy at higher organizational levels. For example, the Army chart contains two levels with nine offices reporting to the Office of the Secretary of the Army and thirty-one offices reporting to the Chief of Staff. The richness of provenance information has little to do with the hierarchical structure of these organizations.

Hierarchical schemata might be useful if they had meaning across organizations, but unfortunately they do not. Levels within one organization have a totally different function from the same levels in another organization, and absolute positions are meaningless. Hierarchical schemata cannot be used, for example, to identify where the personnel hiring function or the new product testing function resides within one organization or to locate this function across organizations. Indeed, in one organization, these functions might be a small section; in other organizations they are larger divisions or major branches.

Having used these examples to demonstrate the weakness of mono-hierarchical schemas for explaining structure, processes, and activities in modern organizations, two important points are apparent. First, there is a richness of information about organizations which has been captured at least in part by those who create and administer them. This information may be found in a variety of organization charts, mission statements, annual reports, parliamentary or congressional submissions, and the like. Secondly, a better theoretical model that captures the complexity of modern, living organizations begins to emerge. Insofar as hierarchical structure is concerned, the model is *poly-hierarchical* because it captures traditional hierarchical relationships *across time* as organizations

5 The inferential process is implied in Lytle's publications cited above, and Lytle has been explicit with colleagues about the notion of inference (correspondence available). Despite its provocative presentation and potential importance for archival practice, the inferential provenance method process has never been studied. Some information scientists have shown interest in it, but archivists have not.

6 Max Weber, "Essentials of Bureaucratic Organization: An Ideal Type Construction," in R.K. Merton, ed., *Reader in Bureaucracy* (Glencoe, Illinois, 1952); Michael A. Lutzker, "Max Weber and the Analysis of Bureaucratic Organization: Notes Toward a Theory of Appraisal," *American Archivist* 45 (Spring 1982), pp. 119-30.

7 *The United States Government Manual 1984-85* (Washington, 1984).

[1]The Chief of Legislative Liaison and the Chief of Public Affairs report directly to the Secretary of the Army and are responsible to the Chief of Staff.
[2]The Inspector General serves as the confidential representative of, and reports directly to, the Secretary of the Army and to the Chief of Staff upon the morale, discipline, efficiency, and economy of the Army.
[3]The Auditor General reports directly to the Chief of Staff with concurrent responsibility to the Secretary of the Army.
[4]The Comptroller of the Army is under the direction and supervision of, and is directly responsible to, the Assistant Secretary of the Army (Installations, Logistics and Financial Management).
[5]With concurrent responsibility to the Chief of Staff.
[6]The Chief of Engineers reports through the Assistant Secretary of the Army (Civil Works) to the Secretary of the Army on civil works matters and commands the U.S. Army Corps of Engineers.
[6]Commander, U.S. Army Criminal Investigation Command reports directly and concurrently to the Secretary of the Army and the Chief of Staff on criminal investigation matters.

Source: *The United States Government Manual 1984-85* (Washington, 1984), Appendix C, p. 827.

re-form themselves, and because it captures relationships which are not within the scope of superior/subordinate relationships. Some of the most important relationships are not hierarchical at all. All of these relationships can be encompassed by the concept of networking — capturing significant formal and informal relationships in an organization which together explain its mission, structure, and activities.

Unfortunately, North American archival theory has accepted the nineteenth-century view of organizations, probably more than most archivists realize. Intellectual limitations imposed by these assumptions, and consequent defects in archival practice, must be overcome before the power of provenance can be realized.

3. Identifying Defects in Current Application of Provenance

3.1 *Consequences of the Received View of Organizations.* Current archival practice overemphasizes the importance of hierarchy. The placement of records-generating activities in organizational hierarchies has become almost an archival obsession. Provenance of archival records is indicated in archival information systems by terms which identify offices of origin and subsequent custodians (including offices responsible for the records); these terms are then linked in archival retrieval systems in a hierarchical schema which serves as a proxy for relationships between the actual offices of origin.

This distortion translates directly into records-keeping practices. In classical bureaucratic organizations, records were kept by and for the offices which used them. In the modern age of central management information systems, records retention and reporting regulations, and vigilant corporate legal staffs, these assumptions are rapidly becoming unwarranted. Offices which generate records frequently do not maintain them. When entire agencies are created or abolished, the connection with a hierarchical schema is difficult enough, but when changes take place within an agency, creating and abolishing divisions, departments, bureaus, and offices, or transferring them to other agencies, merging or dividing existing units, then a mono-hierarchical representation of these changes is impossible. Even more difficult to cope with are subtle shifts and changes over time in missions, functions, responsibilities, and reporting relationships. Yet all these changes take place within real organizations. The problem is that the current archival model does not fit such living organizations.

3.2 *Dysfunctions Resulting from the Record Group Concept.* The record group concept, developed and applied at the National Archives of the United States and adopted widely by Canadian and American archives, is the central implementation of the principle of provenance in the arrangement and description of archives. In fact, the principle of provenance and the record group have become so closely identified that, when drafts of this paper were circulated, archivists often equated criticism of the record group with a rejection of the principle of provenance. Given the identification of "record group" and "principle of provenance" in the minds of many North American archivists, it is important to emphasize that this critique aims at strengthening the application of the principle of provenance by pointing out serious limitations of the record group concept. Far from a rejection of the principle of provenance, the purpose is to strengthen its application to archival theory and practice.

Limitations of the record group concept are severe. It has debilitated archival theory and has limited effective development of provenance-based retrieval in North American repositories. There are two causes of its defects. One is the assumed importance of linking

documentation with the hierarchical placement of the creating unit, a ramification of the classical view of superior/subordinate mono-hierarchical organizations which has been discussed. A second is that the record group, like traditional library classification schemes, is essentially a shelf-order system; since a record unit can go only one place on a shelf, a mono-hierarchical structure results.[8] Supporters of the record group concept point out that, in its implementation at the National Archives and other large repositories, it is not treated as a shelf-order classification. That rebuttal is beside the point. The record group imposes the *intellectual* constraints of a *physical* shelf-order classification; and in fact, many archivists act as though its purpose is to guide shelf-order arrangement. Despite its limited validity as a snapshot of one aspect of an organization, the record group concept has become an albatross.

3.3 *Consequent Confusion Caused by the Record Group Concept.* In case the reader doubts the reality of the record group albatross, some examples can be given of consequent practical and intellectual problems, quite beyond those already noted concerning its immediate implementation in repositories.

Early in its deliberations, the National Information Systems Task Force (NISTF) set out to identify descriptive elements (date, title, etc.) which had been applied to archival materials. This effort met with the insistence that the data elements be keyed to "hierarchical level," presumably to hierarchical level of record units. A thorough empirical study of the descriptive practices of archival repositories[9] demonstrated no connection between "hierarchical level" and descriptive elements, but the conviction continued, even within the NISTF, that such a relationship existed.

The importance of hierarchy has even been carried over to the design of computerized databases about archives. It was widely held that the SPINDEX software system was ideally suited to archives because its database structure reflected the "hierarchical" structure of archives. Thus the influence of the hierarchy assumption is apparent: a machine-readable database is structured to mimic hierarchical characteristics presumed to inhere in archival units. Some readers of this paper commented that the dependence on hierarchy was overemphasized in the SPINDEX example, or that it was certainly no longer considered crucial to the design of automated archival information systems. To that comment, a quotation from a recent *Computerworld* article on the American Presidential Libraries automation project follows: "[the project is now at the point of] designing a nine-level hierarchical data base that will index the vast collections, many of which are quite detailed."[10]

These examples illustrate that numerous archivists have come to operate as though archives (i.e., the records themselves) exhibit hierarchical relationships to each other.[11]

8 For a devastating criticism of the record group, see Peter J. Scott, "The Record Group Concept: A Case for Abandonment," *American Archivist* 29 (October 1966), pp. 493-504, and in a subsequent exchange of letters between Scott and Meyer Fishbein. Unfortunately, Mr. Scott's advice has not been heeded.

9 Elaine Engst, *Report on Archival Practices for the National Information Systems Task Force* (unpublished paper, Society of American Archivists, October 1980).

10 *Computerworld* 18 (24 December 1984).

11 Of course, records do have some hierarchical characteristics, such as item/folder/box; these are not relevant here because they pertain to whole/part physical relationships rather than record-to-organization relationships.

Whether many archivists would agree with the proposition "archives are hierarchical" is not known, but many operate as though they believed it. The intellectual lineage of this misconception is clearly tied to the implementation of the record group concept in archival institutions, specifically the attempt to arrange records on shelves to reflect the hierarchical structure of organizations.

3.4 *Absence of an Authority Record Approach.* Arrangement and description practices add to the confusion caused by the record group because information about organizations and information about records are all mixed up in administrative histories, prefaces to inventories, and the like. These problems should be addressed by a disciplined approach to authorities, the primary authority being information about creating organizations.

An authority record is a formal means by which the creators and users of an information system maintain a common "language" between them to support information retrieval. An example is a geographical authority record maintained by scientists to support uniform description of where natural history specimens were collected. Good authority records maintain communications and support retrieval, and therefore must change over time. Archivists should not reject the use of authority records because they can point to problems with rigid library authority records such as the Library of Congress Subject Headings. A proposal for implementation of authority records is included in the following section.

4. Expanding the Power of Provenance-based Retrieval for Traditional Archival Environments

If dysfunctional aspects of record group systems are discarded, the path is cleared for a substantial improvement in provenance-based retrieval effectiveness for traditional archival repositories. The major areas of improvement discussed here indicate how progress can be made, without attempting a comprehensive description of a provenance-retrieval environment.

4.1 *View Provenance Information as Retrieval Access Points.* Archivists have traditionally viewed the principle of provenance — along with respect for the original order of records — as the basis for the arrangement and description of records. Although the presumed purpose of arrangement and description is enhanced retrieval, the practical result more closely approximates mere preservation. Significant strides can be made by taking a more aggressive approach to the application of the principle of provenance to retrieval.

Provenance information should be thought of as a means for providing access points to records in archival custody. In that respect, provenance information access points are the same in function as other kinds of access points such as chronological or geographical or subject information. To retrieve anything, a handle is required. The handle, or access point, is a characteristic which can be used in conjunction with other characteristics to identify a set of objects for examination. This applies equally whether the objects of retrieval are items in a grocery store, books in a library, or records in an archives. What differs is the appropriate characteristics — or, more precisely, which characteristics will prove most discriminating and most useful to searchers.

Lytle has made the largely unexplored assertion that provenance-related information will be a discriminating selector in the retrieval of archives. If provenance-related information will provide a powerful tool for retrieval, it must provide useful access points.

4.2 *Specific Recommendations for Provenance-Related Access Points: Emphasize Form and Function.* What elements of information concerning the provenance of records might serve as access points for a retrieval system oriented toward exploiting the power of provenance? Bearman has argued strongly for the discriminating power of function and form of material.[12]

Whether the responsibility resides with the local magistrate, sheriff, county medical officer, parish priest, or census bureau, the governmental function "to record" births will generate predictable records, as will the function "to license" professions or "to authorize" expenditures. Functions are independent of organizational structures, more closely related to the significance of documentation than organizational structures, and both finite in number and linguistically simple.[13] Because archival records are the consequences of activities defined by organizational functions, such a vocabulary can be a powerful indexing language to point to the content of archival holdings, without need for actual examination of the materials themselves or for detailed subject indexing.

Function is not, however, adequate by itself to point to the intellectual content of archival materials. "Form of material" works with "function." "Form of material" is the name given to a particular type of record by cultural contemporaries who generate such records. "Forms of material" are known by contemporaries not from a detailed reading of their contents nor from the physical medium upon which they are written, but from commonalities in their structure.[14] The terms archivists have in the past uselessly assigned in construction of series titles (correspondence, day books, applications, surveyor field books, central registry files, etc.) are "forms of material" and are available to provenance-retrieval systems as intellectual content descriptors. For, if correctly defined by the archivist, such record "type" designations capture in cultural shorthand a description of the informational content of records. The distinction between a diary, a journal, and a day book in the nineteenth century represents a distinction between the categories of information each will contain and the perspective represented by their creator. Archivists know the differences between these "forms" and what information each contains without having to read each example; again archivists can thus know from provenance rather than

12 David A. Bearman, "Implications of the Work of the NISTF for National Cooperation Among Folklife Archives," (unpublished paper, April 1984); "Towards National Information Systems for Archives and Manuscript Repositories: Problems, Policies, and Prospects," (paper presented at the Society of American Archivists' National Information Systems Task Force Conference on Archival Information Interchange, Hoover Institution, Palo Alto, California, 14-15 March 1983); "Toward National Archival Policies: Assessment and Planning on a National Scale," (paper presented at the Society of American Archivists Annual Meeting, September 1984); "Who About What or From Whence, Why and How: Intellectual Access Approaches to Archives and Their Implication for National Archival Information Systems," (paper presented at the Conference on Archives, Automation, and Access, University of Victoria, British Columbia, 1-2 March 1985).

13 In a recent think piece circulated to colleagues, Bearman has advanced a list of less than five hundred transitive verbs which he suggests incorporate all functions of governmental, religious, and commercial organizations. Together with a relatively small number of nouns which serve as objects to these verbs, organizational functions are fully defined. David Bearman, "A Proposal for a Functions Vocabulary: Terms, Formats and Rules for Cooperative Development of an Authority File," (5 November 1985).

14 A bank cheque, written on a watermelon, is nonetheless a cheque and even negotiable! Less extreme, but equally interesting to the authors, is the fact that memoranda are recognizably memoranda even when circulated as electronic mail. Electronic letters are also recognizable as letters. Who knows what new forms of material might arise because of electronic media, but today's electronic exchange systems are still transporting pre-electronic "forms" of mail.

from subject indexing certain elements of the intellectual contents of records. Across time, the logs, registers, and inmate case files of public institutions, for example, represent different contents and a careful designation of these historical changes is sufficient to denote that fact to researchers (a single form of material will also change across time, a fact which historical researchers know and use). In definitions of "forms of material," archivists will move to universal series descriptions which can become the basis for records schedules and for intrajurisdictional comparisons of holdings.[15]

4.3 *Establish Provenance Authority Records and Rigorously Separate Authorities from Description of Records.* Authority records can provide consistency in the use of name, subject, geographical, or other access points. Libraries have established authority records for subject headings and for names of authors, for example. Archival practice regarding authority files is highly variable, but attention to authority questions is much less common in archives than in many other information services.

The purpose of an authority record is to maintain a common "language" between the community of users of an information system, and in complex environments they are essential to effective system use. But authority records must not become an end in themselves, where maintenance of ossified practices becomes enshrined in the rules and procedures promulgated by self-appointed guardians. The fact that such bizarre systems exist in library practice should not lead archivists to discard the authority record altogether. Archivists should identify where they need authority records and then proceed to implement them by the best modern standards of the information services community.

The unique and most powerful archival authority is the provenance authority record. For each organizational function and significant organizational event embodied in a department, committee, task force, working group, planning meeting, or other activity, an authority record should be created. Authority records would include as data elements and access points, at least the name of the entity, the source of its authority, its mission, its functions, the entity to which it reported or was otherwise related, and its active dates. Provenance authorities would be expanded and enriched by the on-going results of provenance-based retrieval. To this peculiarly archival dimension should be added authority records from relevant information services; for example, an archives with natural history data might wish to adopt a standard scientific geographical authority.

Provenance authority data (as described above) and data about actual records must be rigorously segregated. For the latter, an archival control record should be created for every archival series, or for any smaller unit of records to which administrators and other clients might wish access.[16] Provisions should be made to indicate relationships between archival control records in the control system. The relationships to be represented would

15 Bearman initiated discussions of how to implement such authority controls with the State Archives in Utah, Alabama, and Kentucky. Recent work by these archives and the National Archives suggests that agreement on these authorities can be reached and should have the theoretically projected value.

16 In a recent paper following up on some of these ideas, Max Evans has argued that the series should be accorded a privileged position in information systems because he feels it is a "natural" organizational unit of documentary materials. While we agree that nothing except administrative convenience in one period in the history of the National Archives recommends the record group, we see no need for such a privileged status for series. Max Evans, "Authority Control: An Alternative to the Record Group Concept," (paper presented at the SAA Annual Meeting, 1 November 1985).

be limited to those which were about the records themselves: horizontal (media transformation for example), vertical (whole/part), and chronological (versions or records of successor series).

Relationships between offices of origin should be captured in the intellectual content of provenance authority records and such linkages would therefore be exploited through searching the indexes. These authority files would be linked, however, to the archival control records.

The entire system should preserve a clear distinction between four categories of information which archivists and records managers employ in establishing control over their holdings: 1) information about actions performed by the archival repositories and maintained in archival control records; 2) physical description (record "form") and related access data; 3) content descriptions (agency "functions") and related access data; and 4) authority data, including the authority data unique to archival organizations: information about the history, structures, processes, and activities of records-creating organizations.[17]

4.4 *Integrate Archival Processes.* The present practice of most archival repositories segregates the information which the archives may accumulate about the current status of the records-creating organizations and their activities and records (for example, the National Archives' *Federal Register*, the Canadian *Access to Information Register*, *Privacy Index*, and the *Canada Gazette*); the information gathered in the course of the management of that organization's current records; and the information gathered as its archives are appraised, arranged, and described. The intellectual aspects of these segregated efforts have much in common. For example, the provenance authority record should be systematically constructed as the earliest information is available before it is lost, and then all subsequent steps should enrich and verify that record. The process works the other way too. Once a full provenance authority record is available at the archival repository, the information it contains should be very useful to information specialists in the originating organization. In short, by integrating these three sources or systems of information, records management and records appraisal will be better supported as well as improved archival retrieval.

5. Proposal for a Provenance-based Universal Information Access System

The invaluable insight of the principle of provenance is the relationship it reveals between creating activity and information created by organizations. If the archivist's use of provenance in arrangement and description — which establishes links backwards from records to creating activities — is reversed, a potential exists for a practical and powerful means of gaining access to and managing information.

The suggestions made in the last section for improving traditional archival information systems can be greatly expanded. The ingredients of such a system, and the means for automating it, can be stated, although the practicality of fully implementing such a system is unknown.

The system concept is quite simple. The objective is to capture the full richness of provenance information — the structures, processes, and activities of organizations — and to make routine the inferential process which permits one to locate information which has been or is being created by organizational activities. The power of the system will be its ability to retrieve present as well as past information created by organizations; in fact, extrapolations to information yet to be created could be made within certain constraints.

Much of the necessary provenance data is presently being collected. For example, the American National Archives is the publisher of the *United States Government Manual*, the authoritative guide to the structure and mission of the American government, and of the *Federal Register* in which official actions of government agencies and public documents are announced. It could be involved with the Federal Information Locator System which the Office of Management and Budget now maintains. It could also be tied into the General Accounting Office information system on agency/bureau missions. These systems already contain much of the content of the projected provenance authority record, albeit in very primitive form, as it relates to current organizations. (Where such information is not readily available for older or defunct institutions, archivists would have to research the usual historical sources to uncover it.)

This proposal anticipates greatly enhanced techniques for dealing with provenance authority data. For example, it would be necessary to create dynamic, interrelated databases from presently static systems which capture organization structures, functions, missions, activities, and relationships. That new system could serve as the principal access tool for all documentation/information created by organizations. Such enhanced authority control would incorporate poly-hierarchical structural relationships and non-hierarchical relationships in a complex networking model, which would permit tracking of particular functions or activities over time and across jurisdictions. In addition to locating information, the system would serve as a first order institutional memory with independent value for current policy-making processes as well as for location, use, and management of information. As organizations grow larger, more complex, and less hierarchical, their management officials have need of an analysis tool as powerful as the one envisioned for day-to-day administration. The personnel office has an organization table, the budget office has a picture of some on-going projects, the public relations office knows of other projects, the archives has inventories of records series generated by various activities, the planning office has statements of functions and missions — but no one has an overview of what the organization was, how it became what it is today, and where it is headed tomorrow. Archivists are ideally situated to provide such a view through the information systems which they ought to be creating to provide intellectual control for historical materials.

The second component of the ideal information system is the inference process which supports two-way translation between subject questions and provenance/organizational

---

17 Bearman has written a design model for a Collections Information System for the Smithsonian Institution's holdings (which range from paintings to beetles, from buttons to minerals) based on these same four categories of data. Smithsonian Institution, *Request for Comment, Smithsonian Institution Collections Information System — A Plan for the Acquisition of an Integrated, Generalized Collections Management Information System* (Office of Information Resource Management, April 1984), 28 pp. and three appendices. Most recently, Bearman coauthored a generic model for any action-based systems using heuristic models drawn from artificial intelligence research which further illustrates the fact that it is the "collectedness" rather than the "objectiveness" of items in archives, museums, zoos, and botanical gardens which makes it possible to define fundamental structures for information systems designed to support the management of such holdings. Office of Information Resource Management, Smithsonian Institution, *CMASS: Statement of Problem* (September 1985), 48 pp.

activity terms. This inference process can be observed whenever a reference archivist translates a user inquiry for information first into a question about what kinds of records hold that information and then into a question about what functions generate such forms of material and from there into a question about institutional history so as to locate the record series which is most likely to answer the user query. More rigorous study of the precise manner in which this inference process functions would be required to develop rules by which the process could be automated.

A fully automated archival retrieval system for naive users would have two components. The first would be a computerized information system containing complete knowledge of an organization's administrative history — one full complement of provenance information — in a very flexible database environment which supports the complex relationships involved in modern organizations. The second would be an "inference engine," a software system which executes the provenance-inference process in place of the reference archivist.[18] The inference engine will have the ability to make inferences from user's questions to provenance information and hence to the desired documentation or information.

Clearly, the ultimate system is not in our immediate future. But the preceding discussion — and the notion of a vastly expanded and automated provenance system — has its applications for archivists today and will support enhanced automated applications tomorrow.

6. Conclusion

Immediate steps for archivists to take in order to improve archival information systems have been suggested. In summary, they are as follows: view provenance information as a provider of retrieval access points; emphasize form of material and function in retrieval systems; establish provenance authority records; rigorously separate authorities from description or control of records; and integrate archival processes from records creation through records appraisal to records description. Two immediate steps should be taken to enhance present retrieval systems and to take advantage of imminent new automated capabilities.

Improvements presently within our grasp could use "form of material" and "function," in addition to more traditional archival descriptive access points, to support detailed content inquiries and permit reference archivists to retrieve with greater precision and recall than they can using currently existing approaches. Most of these new capabilities could be supported by appropriately designed manual systems. Immediate work is needed, however, to elaborate and standardize a "form of material" and "function" vocabulary for archival retrieval systems, for they are the critical components of the data architecture of the automated system. While these immediately realizable results are being achieved, improvements can also be made which will capitalize on the inference engine and other vastly more powerful information technologies when they become commercially available.

---

18 Archival readers may be uncomfortable with the notion of an inference engine. For some years, information and computer scientists have been exploring the potential of automated systems to perform "human-like" reasoning. Very primitive prototypes are in existence. See *Computerworld* 19 (19 August 1985).

The second step pertains to the inferential process itself that archivists use to proceed from subject queries to location of pertinent documentation. Archivists should study the application of the reference archivist's inferential process discussed at some length in this paper. At best, this process will be the engine which drives vastly enhanced information retrieval systems. At a minimum, archivists could achieve considerable improvement simply by understanding the process more completely.

In these and other efforts, archivists should keep in mind that the need for more effective management of information and other resources in modern organizations can be exploited by archivists to gain additional staff and money for their work. Perhaps the best way to attract such new resources is to exploit the power of existing provenance information by better structuring current systems around authority controlled access points. Archivists hold the key to provenance information becoming a major tool for management of all resources in modern organizations.