SECTION IV

Standards

CHAPTER SEVEN

Information Technology Standards and Archives

CHAPTER EIGHT

Documenting Documentation

CHAPTER SEVEN

Information Technology Standards and Archives*

Standards are designed to overcome boundaries. The boundaries presented by information technologies have been envisioned as a series of seven steps (the OSI model) each of which provides a platform for communications across systems. Archivists are most concerned with interchange standards on the application (seventh) level. These standards, if appropriately formulated, could convey the context in which information is created and managed and the structural relationships between components of the data content as well as the raw data in the system. This chapter discusses the archival requirements for interchange standards at this level. It suggests that archivists have employed tools in their discipline that give them a valuable insight into requirements for accountable management of these new forms of cultural communications. The implications of these for new technology standards will be elucidated.

* Originally published in *Janus* (1992.2): 161-166.

INTRODUCTION

Archivists are confronted by rapidly changing methods of work made possible by the use of electronic information technologies. Often they are told that the only way they can hope to preserve the information generated by these technologies is to employ and influence information systems standards. But when they look for ways to become involved in the definition of standards, they encounter a vast array of information systems standards and discover an inchoate universe of standards development activity. Which standards are important to implement or to influence?

Without criteria by which to evaluate what is most important, archivists will have little impact on information systems standards, even if they try to become involved. In this chapter, I present a framework for evaluating the potential significance of an information systems standard to archives based on an analysis of what makes a record archival. I suggest how this criterion can be employed to effect standards development and in what way that would have an impact on the documentation of the twenty-first century.

THE CHANGING CHARACTER OF WORK

In order to understand the potential significance of successful intervention in the definition of information systems standards, we must first appreciate the changes that are taking place in the conduct of work in modern bureaucracies. To illustrate these, I will present a hypothetical, but by no means unlikely, case study of how the Office of the Attorney General of one of the states in the United States would defend a newly legislated method of statewide school funding being challenged in a state Supreme Court. The Attorney General is the principal legal officer of the jurisdiction and, until recently, most states funded education through local (county level) taxation. Recently, however, the Supreme Courts of some of these

states have recently required state legislatures to adopt different methods of funding on the grounds that local tax bases differ and thereby result in discrimination based on property values, or on wealth.

The advantages of this case study is that archivists will have little difficulty agreeing a priori with the presumption that the evidence of such a high level and critical governmental activity should be archival and that in the United States we have witnessed dramatic changes in the way in which lawyers, in and out of government, conduct their work since the advent of electronic information systems.

In this hypothetical case, a team of several lawyers would probably be assigned to work together on researching and writing a brief representing the government's position. They would use a "groupware" writing tool which tracks versions and revisions and permits several individuals to write, comment on, and revise the same document. To prepare their arguments, they would first search online databases for references to prior case law and download these references into a local database making them available for citation in their legal brief. At some point in the preparation of the case, they would also search census data and other demographic databases maintained by the state in order to demonstrate how statewide funding would serve the larger social good of providing equality in educational services. In addition to retrieving statistical data, they would probably view their retrieval results through a Geographic Information System, in order to illustrate graphically the equity issues involved in statewide financing.

During the period devoted to drafting the brief, some members of the team would be out of town on other business or gathering evidence for this case. This would not impede their use of the groupware environment or online databases which they would access by telecommunications. In addition, they would communicate with the other members of the team by voice mail and fax. The voice messages (digitized analog

signals) and the fax communications (digitized raster data) would be received on computers where they would be stored in software-controlled voice and fax mailboxes and indexed for subsequent retrieval. Some might also conduct depositions taped on audio or video tapes to provide evidence of school disparities. These tapes would also be indexed and stored. Evidence would include still frames captured from video clips and digital sound from interviews. These sounds and images could be directly incorporated into multimedia documents. While it is rare now, in the near future the legal team would submit their briefs to the court electronically. Such briefs could consist of hypermedia rather than just linear multimedia segments.

INFORMATION SYSTEMS STANDARDS

The archival interest in this case is to assure the preservation not only of the raw data of the brief submitted to the court, but also evidence of the way in which the government conducted and built its case. This objective requires that archivists be able to rely on standards for interchange of data, of information about data structure, and of information about data context.

The data (text, image, and sound) actually created by the government legal team, or recorded by it as evidence, needs to be usable, understandable, and available to future researchers. This requirement would be satisfied by data representation standards such as ASCII for text, JPEG for images, FM or CD sampling rates for audio, CCITT Group III or Group IV protocols for fax, and VHS/NTSC, PAL, or SECAM standards for video. While data representation standards today almost universally accepted for data interchange do not absolutely assure that the data will be usable 100 years from now, we can assume that they define a sufficiently widespread usage that a migration path will be provided between them and whatever standards prevail in the future. Vigilant archivists will be able

to move their data along this migration path without loss of information content.

It is important for archivists to realize that they can depend on these standards for data interchange because system designers will implement this level of data interchange capability in response to widely expressed operational requirements of business and government. The need archivists have to transfer this data outside the systems in which it was created do not differ from, and have nothing to add to, the requirements of the business community. It follows that archivists can have little impact on the evolution of such data representation standards.

However, simply transferring the words and symbols created by the lawyers and witnesses in this case will not preserve the archival record as generally understood. Archivists must also capture and preserve structural and contextual information which gives the data created by the legal team its significance as evidence.

Contextual data is information about the creation and use of information. It is not resident in the texts, images, and sounds of the "documents," but it is acquired and/or created, by the information systems in which these reside and used in those systems to manage documents. A simple example of contextual data is the information maintained by an electronic mail system which records the sender, addressee, security assigned by sender, time/date sent, time/date opened, and reply, forwarding, and/or filing history. Within an operating e-mail system, this information must be created and managed by the computer, but it is largely recorded and stored in a proprietary way and will not be interchanged with other systems unless the interchange protocol employed by the two systems requires it.

Business requirements for contextual data interchange are quite limited. In the case of electronic mail, for example, they are reflected in the interchange standards defined for e-mail headers and directories (IEEE X.400 and X.500) which name

the addressee and the response requested. They do not reflect the full range of archival concerns, however, as they do not provide for interchange of information about the provenance and revision history of the records in question. Other contextual information in our legal case study relates to the legal and demographic databases searched for citations and evidence and the GIS systems used to represent the demographic information. In order to use information as archival records, we need to be able to represent what data in these databases was available to the searchers (e.g., their permissions and views), what questions they asked (e.g., their search strategies), and what algorithms were used for the geographic representations. This information includes that which is called "metadata" about these systems (covered by standards for Information Resource Directory Systems); "information retrieval commands" using languages such as Common Command Language (CCL) or Structured Query Language (SQL); and representations of user access rules, security, and database views. We also need to retain data about the actual state of the databases at the time they were searched.

Requirements for a complete archival record do not end with capturing contextual data, however, because information is also conveyed by the structure of the records which are retained. Structure has long been recognized by archivists as a conveyor of meaning[1] but the importance of standards for conveying structural relations of electronic archival records -- including internal documents markings which graphically convey meaning -- has not been alluded to in the archival literature. This structural information might include both the internal structures of documents and the structural relations between records in a database which are used to the software to construct the equivalent of the physical record in the paper file. In automated systems, the "logical" record (for instance the case file of an individual) may consist of a large number of discrete physical records stored in, and under the control of, different information systems. The relations between these

records determine the meaning of the logical record, and also its currency and authenticity. In emerging object-oriented and hypermedia environments, these links and their rules govern whether and how the data can be viewed and what can be done with it.

Standards for structural information are underutilized and underdeveloped from an archival point of view. Standards for representing the versioning of documents in the groupware environment, for representing the links between objects in hypermedia, and for representing the logical components of textual documents which give them their distinguishing "form" have not yet been seen as critical for business operations. As a consequence little attention has been given to preserving this information in a software-independent fashion. Archivists need to press the case for why it is critical for organizations to preserve structural information across systems. At the same time they could begin to use the limited structural data interchange standards which exist, such as Standard Generalized Markup Language (SGML) which was developed initially for the publishing industry.[2]

The relationships between elements of information and physical records and objects in a database determines the meaning of the information; for example, whether the data about a person included, or could have included, a link to the record with that person's current address is an issue that could have evidential significance. In addition, the views of a database that are permitted to specific users are controlled by permission tables which are themselves data to the database. This kind of data about the database and its use is called metadata, and can currently be documented following an international standard for Information Resource Directory Systems.

ARCHIVAL STRATEGY FOR INFORMATION SYSTEMS STANDARDS

Archivists are only likely to have an impact on information systems standard development when they play an active,

concerted role. They should concentrate their efforts where archival requirements depart from those of everyday business operational needs. In these areas, which are related to the archival concern for provenance,[3] they can articulate their functional requirements for standards by exploiting the potential of standardizing structural and contextual data capture in widely used information system applications. In part they must advance this agenda by critiquing existing standards based on their contextual and structural requirements. They need to illustrate to senior managers that the short-term operational requirements for transportability of standard representations of data content will not ensure retention of archival evidence essential to reconstructing the transaction or activity which is the object of the archival record.

The focus should be on those applications which are already widespread and likely to play a significant role in the changing character of work:

- office automation and electronic mail
- databases and data analysis/display
- electronic information dissemination/publication
- automatic transaction processing

In each of these areas the existing standards for data interchange do not adequately account for archival concerns but could be made to carry information required by archivists to represent structural and contextual information now associated with records in proprietary ways.

In the area of office automation, the primary concern is to capture and transmit in a non-proprietary way, the history of documents including their authorship, the source of each version, and the rules governing access and use. Presently such documents are interchanged using electronic mail facilities which have a "header" (X.500) that permits the representation of the name and electronic addresses of the sender and recipient, the time and date of transmission, and little else. Archivists should make an effort to get extensions to header

standards that would require the originating system to record both structural properties of the records being sent (form notations designating parts of the document that are its content and parts that serve, for example, as a distribution list) and contextual properties (such as versioning, permissions, and views data), identification of the work process of origin, and data regarding the actions that are requested (such as replies, acknowledgments, or follow-ups requested by the sender).

With respect to databases, neither the retrieval request (query) nor the means of reporting or representing the results have been made subject to interchange standards, yet we can hardly expect to make sense of a decision based on querying a database if we do not have the question or the answer in hand in an interpretable form. Archivists should examine the standard developed for database retrieval by the U.S. library community (ANSI Z39.50) for its appropriateness for software-independent query interchange and also study methods for interchange of user-defined software display rules such as those embodied in a graphic generated by a spreadsheet or a map generated by a geographic information system.[4]

Archivists also need to exploit the abstract data structural representation capabilities of the Information Resource Directory Systems (IRDS) and other "metadata" facilities. If fully implemented, metadata systems can carry software independent representations of structure and also of such contextually significant data as permissions, views, report definitions, and calculation capabilities which affect the results of database reports.

Information dissemination and publication conveys structural information through the form of documents that cannot presently be interchanged between systems without loss. Some efforts to provide methods to standardize representation of structural features of documents (such as SGML) partially satisfy archival needs, but they do not yet represent the revisions between versions, multimedia elements in texts, and the navigation of non-linear documents. Archivists can

now get involved in the definition of the Office Document Architecture and Office Document Interchange Format (ODA/ODIF) standards which are more extensive than previous efforts but will still require archival insights to ensure that they satisfy the need for software-independent representation of evidence of transactions.

Business transactions, such as the filing of a legal brief, the withdrawal of money from a bank, or ordering supplies and invoicing for them, are increasingly being conducted by electronic means. Archivists need to become involved in the representation of electronic transactions to transaction processing systems. The premium in such routine systems is on reducing the amount of information to what is the minimum that must be conveyed to successfully conclude the transaction. It may not satisfy evidential requirements. Each market niche and interchange community is developing such transactional standards on its own, often within the international frameworks for EDI (Electronic Document Interchange) but occasionally outside that framework as well.

Assertive tactics will require archivists to become involved in the standards under development for improved specifications for IRDS, X.500, SGML, and Office Document Architecture (ODA), while promoting an understanding of the need for contextual and structural standards within their institutions. Archivists in organizations using other data communication protocols (EDI, ANSI Z39.50, etc.) need to examine the extent to which these transactions convey in their standard forms the information required to locate their organizational and programmatic provenance and their place within a series of communications internal to a business process.

NOTES

[1] David Bearman and Peter Sigmund, "Explorations of Form of Material Authority Files by Dutch Archivists," *American Archivist* 50 (Spring 1987): 249-253.

[2] Although the requirement for conveying the logical structures internal to printed documents is only partially addressed by SGML, archivists could employ SGML to develop abstract, form-of-material, "fingerprints" for organizationally significant types of documents which could then be recognized by automatic parsers developed to screen for archivally significant communications. They could also develop strategies to audit, create, and break record links in database environments rather than trying to audit all modifying transactions.

[3] David Bearman, "Archival Principles and the Electronic Office" in *Information Handling in Offices and Archives*, Angelika Menne-Haritz ed. (New York: K.G. Saur, 1993): 177-193, reprinted in this volume as Chapter 5.

[4] Archivists often find it difficult to understand that the user-defined variables in software determine the output of the database. A simple example is the database report that is imported into a spreadsheet in which the cells of the spreadsheet have been predefined by the user with underlying algorithms so that a percentage increase or decrease is automatically added to the database report to represent projections for a future year. The user sees a graphic, generated out of the database but through the spreadsheet which is interpreted data. Currently the interchange capabilities for representing data content are able to send the graphic report, but will not carry the spreadsheet algorithms that produced it.