Archivists, Librarians, and Documentalists:

Our Common Ground in the Information Age

Charles Dollar
National Archives and Records Administration
Washington, DC

Introduction

I am delighted to have the opportunity to attend this Congress and to participate in your discussions. In North America, we generally do not have meetings that draw together the communities and disciplines represented in this Congress, largely because we have allowed our differences in disciplines and practice to over shadow how much we have in common. It is for this reason that I look forward to exchanging views with you and discussing common problems.

The convergence of information technologies in the last five years or so is on the verge of causing enormous changes in the way we work, play, socialize, organize, teach, and even make war. By the end of this century electronic information systems that support routine activities - at home, at school, at work, and in business, military, science, medicine, education, and goverment , among others. These information technologies, which can be characterized as information capture, information processing, information storage, and information sharing, will have a dramatic impact on the work of archivists, librarians, and documentalists in the years ahead. I will forego a discussion of these information technologies in order to get to the heart of my paper this afternoon.

The increasing use of information technologies by the scholarly community, the publishing community, by government agencies to carry out business, and the public at large is breaking down barriers that traditionally have separated information professionals and is expanding the number of players who facilitate

information sharing. By the latter, I mean telecommunication specialists, managers of computer networks, and the designers of information systems, among others. As archivists, librians, documentalists, we must be participants in this expanded information handling community if our disciplines are to survive in the information age.

I believe there is a common ground for archivists, librarians, and documentalists in the information age. This common ground involves a function that goes to the heart of what we do. It is to preserve and provide access to information in a variety of forms. Thus, we come to the theme of my paper that the convergence of information technologies requires archivists, librarians, and documentalists to redefine preservation and access, and in the light of this redefinition to adapt our practices and methodologies in order to accommodate new and emerging information technologies.

Before developing this theme in some detail, I wish to discuss three information technology based imperatives that are part of our common environment. The first is the increasing use of new and emerging information technologies in the creation and use of electronic or digital documents. The second imperative is the use of new and emerging information technologies to convert traditional page image hard copy documents to digital images for preservation and access purposes. The third imperative is the increasing use scholars in the social sciences and humanities are making of computer processable text and documents in their research.

1. Increased Use of Information Technologies

Few people dispute the contention that information technologies are increasingly displacing manual and paper-intensive information handling activities. Numerous publications and conferences over the last year or so attest to this displacement, to say nothing of increased expenditures for powerful local area networks, high-speed personal computers, and new storage media and a variety of related hardware and software development. I have developed this theme in another study and will say only that these technologies are manifested in E-Mail, EDI transmissions, local arera networks, and high speed digital telecommunications, among others.

2. Digital Conversion for Preservation and Access

A major problem with much of the paper-based information in our libraries, archives, and documentation centers is that they are brittle and decaying, accompanied by incrasing legibility problems. This problem is further exacerbated each time researchers use the material. Library, archives, and documentation center staffs understand the problem. Traditionally, the solution has been to photocopy or microfilm the material for researchers to use and curtail access to the original material. Photocopy replicates the portability of original material in either page or book form but it is a short-term solution because eventually the photocopyied material will

deteriorate from use. Microfilm, of course, is a very powerful and well-established low-risk conversion technology. For example, if processed and stored in accordance with strict national standards, microfilm has an extremely long life , at least 100 years or more. Furthermore, copies of the master negative can be used to make duplicates almost indefinitely with virtually no loss of information. Despite these strengths, microfilm is awkward for researchers, lacks the portability of page images, and does not lend itself easily to hard copy production.

In the United States and Europe there are several projects either already completed or are now underway in which digital imaging technologies are being used in lieu of microfilm converstion and storage. Digital imaging essentially involves making an electronic image of original paper or microfilm based material. Because electronic page images are not searchable like ASCII text, it is ncessary to create a detailed index for retrieval purposes. The advantages of digital imaging include the capacity to improve significantly the legibility or readaility of material with only a modest effort in most cases. In addition, digital imges can be read and repoduced over and over again with no loss in image quality.

The National Archives of the United States recently completed a pilot project to scan, enhance, and store on optical disks selected documents from the U.S. Civil War. The study, which was begun in 1987, demonstrated that digital imaging technology and the retrieval system used in the project could produce superior

images and speed retrieval time of staff and researchers. These benefits were off-set by the substantial cost of conversion and maintaining the system.[1]

As many of you undoubtedly know, your neighbor to the east has a major digital image and optical storage system project underway in Seville. The project, which is near completion as part of the Quincentenary Celebration of Christopher Columbus travels to the New World, involves some 9,000,000 pages of material from the late 15th century through the end of the 19th century. Those of you who have seen the system demonstrated would agree, I believe, that digital imaging and optical storage technolgoies can produce remarkable results in terms of improved readability and access. I might add in passing that I understand some consideration is being given to a similar project, although on a much smaller scale, at the National Archives of Portugal.

From a library and archives perspective, the greatest impetus to the use of digital imaging technologies for preservation and access purposes is the work of the U.S. National Commission on Preservation and Access. In 1990 the Commission published a study that recommended the exploration of the use of digital imaging technologies for preservation and access.[2] With the support of the Commission and the Xerox Corporation, Cornell University began a pilot study to digitally scan and store on

---

1. See Report of the Optical Digital Image Storage System (National Archives: Washington, 1991).

2. Michael Lask, Image Formats for Preservation and Access. A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access (Washington, 1990).

optical disks some 1,000 britle books. Unlike the project of the National Archives, the Cornell Project concludes that the cost of digital scanning equals that of microfilm. Another study at Yale University is looking at the digital conversion of microfilm, particularly microfilm with poor image resutling from improper filming, chemcial processing, or storage.

A number of other digital imaging projects underway at agencies of the U.S. federal governemnt and numerous state and logcal government agencies are engaged in similar efforts. It seems clear that digital imaging technologies are cost-competitive with microfilm and photocopies. As the cost of digital imaging declines even further, this technology will become a powerful preservation technology for archives, libraries, and documentation centers everywhere.

Scholarly Use of Machine-Readable Text and Data

A paradigm shift in scholarly research practices is clearly underway as more and more scholars are converting primary textual sources to machine-readable form in order to conduct computer-assisted analysis and intepretation. The National Center for Machine-Readable Text in the Humanities, headquartered at Rutgers University estimates that about 8,000 series of converted electronic text have been created.

Perhaps the first such effort began in 1957 when the French Government began a project to develop a new dictionary of the French language.[1] In preparing this dictionary, some 150 million words were taken from major French literature and philosophy and science and technical literature. In 1982 the French government deposited some 1500 machine-readable texts at the University of Chicago to support a project called American and French Research on teh Treasury of the French Language (ARTFL). These machine-readable texts, augmented by troubadour poetry, texts from the 1848 revolution, and a collection of 17th century French theater posters, are now available to a variety of scholars.

The earliest American conversion project, the Thesaraus Linguae Graecae (TGL), was begun in 1972. Today, the database consists of more than 8,000 works of classical Greek text drawn largely from the period of Homer (ca. 750 BC) through AD 600. This electronic collection is used primarily by researchers in literature, linguistics, ancient history, philosophy, and religion.

A third large file, the Medieval and Modern Databank (MEMDB), was established in 1982 at Rutgers University in order to make an electronic library available for medieval and early modern historians. Currently, the databank consists largely of some 13,000 medieval currency exchange quotations from about 1100 to

---

1. This section draws heavily upon the research of Avra Michaelson and Jeff Rothenberg in their forthcoming study Scholarly Communication, Information Technology and Archives (Rand Corporation, 1992).

1500 AD. Eventually, the databank will include taxation records, wills, and inventories, vital statistics, import/export records, household/estate accounts, glossaries of weights and measures, and calenders of dates, among others.

There are other major electronic compilation projects underway in the United States, Great Britain, Israel, and Spain, along with hundreds of smaller conversion projects throughout Europe.

One of the more interesting and instructive (for this audience) spin-offs of these text conversion projects is the Text Encoding Initiative, an effort to determine the elements and methods for encoding machine-readable text for electronic interchange. The instructive aspect is that the text encoding methods closely follow ISO 8879, the Standardized General Markup Language (SGML). This standard specifies a format encode (or mark up) text that can be processed in a machine- and software-independent form by scholars using incompatible computer systems.

The on-going conversion and encoding of electronic text involves an enormous investment in time and resources. A crucial question for these electronic databanks, and ultimately for archives, libraries, and documentation centers, is how to ensure the preservation of these digital collections over time, a topic which I now wish to examine in some detail.

The three information technologies imperatives just reviewed require rethinking and modification of preservation as permanent retention. As James M. O'Toole observed in a recent essay, the notion of permanence, at least in the United States, is by no means an absolute.[1] Over time, the meaning of permanence has ranged between permanence of the informational content of documents to permanence of the physical objects themselves. Interestingly, both concepts are rooted in information technologies. In the 18th and mid-19th Centuries in the United States the printing of multiple copies was seen as a way of perpetuating information. However, beginning in the early 20th Century, technological developments aroused hope that the usable lifetime of documents could be extended indefinitely. Hence, permanent retention came to mean the physical extension of the usable life of originals for an unlimited period of time.

Efforts to apply this idea of permanent retention to a mushrooming volume of paper records soon ran head on into financial reality. In the United States, and no doubt this is true in other countries, the costs of preserving the useful life of original material for an unlimited period of time are so great that the goal is unachievable. Consequently, many archivists are abandoning the idea of permanent retention as meaning that original documents will be preserved forever.

1."On the Idea of Permanence," The American Archivist 52, (Winter 1989): 10 - 25.

The inherent technological obsolescence of electronic media and the devices required to read them further undermines the notion of permanent retention. Periodic recopying of electronic material to ensure migration from old technologies to new technologies, of course, mitigates the effects of technology obsolescence. However, periodic recopying is not without substantial costs. It is unlikely that any national archives will have the financial resources necessary to continue periodic recopying of all "permanent electronic records" into the foreseeable future.

The high costs of minimizing the effects of technology obsolescence are not likely to decline in the foreseeable future. In order to deal with this reality, archivists,librarians, and documentationalists must bring a new perspective to bear upon "permanence" and "permanent retention." "Continuing value,"[1] which implies that information may lose value because of a declining need for it over time, is a useful concept. However, the notion of "continuing value" must be linked to a systematic effort to reevaluate the cost and benefits of the retention of digital information. This reassessment must take into account the costs of migrating electronic material from an old system to a new system, which are likely to be greater than storage costs between migration periods.[2] At the same time, realistic

---

1."Permanent value" is not used in the statutory authority of the National Archives of the United States. Instead, the terms "continuing value" and "appropriate for preservation" are used. In the context of this paper "continuing value" conveys the sense that use - actual or potential - is crucial in continued retention.

2.Management of Electronic Records: 45.

assessments of the benefits of retention of the records also must be undertaken. This realistic reassessment, as David Bearman has noted, should identify the risks involved in concluding that the costs of retaining electronic information exceed the benefits of retention.

In North America contemporary archival and library preservation involves at least three different actions. The first action is the prevention of further damage to documents, and this generally requires establishing a controlled environment and using specific techniques that stabilize the deterioration of paper documents. The second preservation action is the conversion of documents to other formats - typically paper to microform - when the original version itself possesses little or no intrinsic value. The third action is the restoration of usability of the original carrier of documentary information - in so far as this is possible. With a badly damaged document, it may be possible only to stabilize its condition so that no further deterioration occurs. Typically, restoration is confined to documents of high intrinsic value.[1]

The common idea running through these three preservation activities is that, because the physical carrier itself bears information, ensuring the preservation of the carrier - paper or microfilm - ensures that the information itself is preserved. Of course, this is meaningful when the information and its carrier are physically interdependent. However, an emphasis on

---

1.For a discussion of the concept of intrinsic value, see Intrinsic Value in Archival Material: Staff Information Paper 21, National Archives and Records Service (Washington, 1982).

the carrier of information offers little useful guidance for dealing with digital material in the 1990s, especially with those produced by integrated office systems or integrated database management systems in which records do not exist as physical entities but rather as virtual documents, smart documents, and database views. The preservation of digital material requires shifting the emphasis from preservation of the information carrier or physical storage media to the preservation of access to information electronically captured and stored. Shifting the emphasis from the carrier of information to the intellectual elements of information results in a fundamental reorientation to preservation activities.

Access to digital material, therefore, becomes a question of readability and intelligibility. Readability means that the information can be processed on a computer system or device other than the one that initially created them or on which it is currently stored. Typically, non-readability involves some aspect of either the storage device (a tape or disk) that is physically incompatible and cannot be read by a computer or the coding of the information is such that a computer cannot recognize it. In contrast, intelligibility means that the information is comprehensible to a human being. Intelligibility functions at two levels. The first level occurs when the display of digital information requires nothing more than human recognition for it to be intelligible. An electronic image (raster bit map) or an ASCII text file are two examples. The second level occurs when digital information does not carry sufficient information (i.e., it is not self-referential) for a human to comprehend its con-

tent. Usually, this problem is associated with both coded and numeric data, and the intelligibility of such information is assured by the use of documentation defining the values represented by the numbers and codes. Achieving intelligibility of electronic material is extremely difficult and expensive when the documentation is electronic and is embedded in a proprietary software dependent system. This is particularly true for digital image systems in which proprietry compression techniques are used to reduce the amount of storage required for each image or where a proprietary image file header is used. In both instances, this constitutes an electronic encryption that is not reversible without the original software.

Preservation of electronic informations, therefore, means ensuring their readability and intelligibility in order to facilitate data exchange over time. Hardware obsolescence, of course, is a major barrier to this exchange, as storage devices and media used today will be incompatible with those likely to be developed in the future. Equally as important is the prospect that electronic material that is software dependent will lose its intelligibility when the software becomes obsolescent.

There are at least two alternatives that we may pursue in dealing with hardware and software dependence. One alternative involves retaining either a paper or microfilm copy of the original materia. Microfilm in particular is very attractive because of its well-established longevity and the fact that high-speed OCR devices are available that can scan computer output microfilm and convert images to ASCII text. Under this alterna-

tive, electronic material could be converted to COM and then converted back to ASCII on demand. Microfilm, of course, is a low-risk technology that ensures both readability and intelligibility, and in this context clearly obviates technology obsolescence.[1] Unfortunately, this alternative would only be effective in dealing with electronic material that is in traditional page image format (e.g., letters, reports, memos, books, and the like). This COM alternative would not be an effective way of dealing with numeric data (which requires documentation to be intelligible), relational databases, Geographic Information Systems, hypermedia and multimedia systems.

The other alternative to ensure the readability of electronic information over time is periodic recopying. However, as the volume of information in electronic form increases, this is likely to become a major financial burden. Data exchange standards, which support upward migration paths that bridge computer generations, potentially can extend the time between recopying from, say, ten years to twenty years. Similarly, standards for interactive electronic documentation, such as the Information Resource Dictionary System, are intended to provide a bridge between otherwise incompatible software systems,[2] thereby extending the intelligibility of electronic records.

---

1. John Mallison, "on the Preservation of Human- and Machine-Readable Records," Information Technology and Libraries 7 (March 1988): 19 - 23.

2. Incompatible software systems also includes software that has become obsolescent.

Ensuring the readability and intelligibility of electronic records over time through adherence to information technology standards involves techniques and tools that are substantially different from those with which most librarians, archivists, and documentalists are familiar. This information technology environment is called an "open systems environment," one in which those parts of computer processing that need to be shared are standardized. In 1979 the International Standards Organization adopted the Open Systems Interconnection (OSI) Model, which proposed development of specific public (i.e., non-proprietary) standards. Since then the United States Federal Government has adopted a version of OSI called GOSIP - Government Open Systems Interconnection Profile - while the United Kingdom has adopted its own version called UKGOSIP. Canada has adopted COSAC - Canadian Open Systems Application Criteria. The United Nations has a Working Group studying the possibility of developing an UNOSI profile. Undoubtedly other countries and organizations will develop similar profiles.

The critical issue for libranis, archivists, and documentationalists is how to ensure that these profiles and accompanying standards address their information handling requirements. Consequently, we must understand how the process of standards development and implementation operates, identify and concentrate upon those standards of greatest relevance for our programs, and then we must become actively involved in the development and implementation process.

Conclusion

In concluding this paper I wish to summarize what I have discussed and urge librarians, archivists, and documentatonalists to focus upon our common ground.

The technology application imperatives I reviewed will not abate nor can we significantly impede their growth and expansion. We must adapt our work to these and other yet unknown technology applications. Our common ground of preservation and access, if redefined as I suggested, will help ensure that the services we perform will remain a viable part of the unfolding information handling community of the future.