63RD IFLA Council and General Conference

Workshop: Audiovisual and Multimedia joint with Preservation and Conservation, Information Technology, Library Buildings and Equipment, and the PAC Core Programme, September 4, 1997

A Digital Dark Ages?
Challenges in the Preservation of Electronic Information

Terry Kuny [1]
XIST Inc. / UDT Core Programme
email: terry.kuny@xist.com

Last revised: August 27, 1997

Who controls the past controls the future. Who controls the present controls the past.
George Orwell, Nineteen Eighty-Four, 1949

Monks and monasteries played a vital role in the Middle Ages in preserving and distributing books. It was their work which provided much of our present knowledge of the ancient past and of the rich heritage of Greek, Roman and Arabic traditions. With the advent of the printing press, this monastic tradition disappeared. However, the reverence for the historical record of text has been carried by librarians and archivists within private and public libraries to this very day.

The tenor of our time appears to regard history as having ended, with pronouncements from many techno-pundits claiming that the Internet is revolutionary and changes everything. We seem at times, to be living in what Umberto Eco has called an "epoch of forgetting." Within this hyperbolic environment of technology euphoria, there is a constant, albeit weaker, call among information professionals for a more sustained thinking about the impacts of the new technologies on society. One of these impacts is how we are to preserve the historic record in an electronic era where change and speed is valued more highly that conservation and longevity.

As we moves into the electronic era of digital objects[2], it is important to know that there are new barbarians at the gate and that we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever. We are, to my mind, living in the midst of digital Dark Ages; consequently, much as monks of times past, it falls to librarians and archivists to hold to the tradition which reveres history and the published heritage of our times.

THE DIGITAL DARK AGES

The following observations about the present environment provide the basis for the assertion that we are living in the midst of a digital dark age:

- Enormous amounts of digital information are already lost forever. Digital history cannot be recreated by individuals and organizations cannot recreate a digital history because it was not archived or managed properly or it resides in formats that cannot be accessed because the information is on out-dated word-processor files, old database formats, or saved on readable media. Many large data-sets in governments and universities world-wide have been made obsolete by changing technologies (think punch cards and 12" floppy disks) and will either be lost or subject to expensive "rescue" operations to save the information. Unsurprisingly, the Report of the Task Force On Archiving of Digital Information has identified in its recommendations, the development of "effective fail-safe mechanisms to support the aggressive rescue of endangered digital information."[3]

- There will be a demographic bulge of electronic materials coming into libraries and archives as the Baby Boom generation of authors and academics begin to wind down their careers and begin off-loading their materials to various libraries and archives. These materials will come to libraries on a wide-variety of storage devices, perhaps even in entire computer systems, and will probably have equally significant paper collections associated with them. To assist the archivist of 2015, we need to find methods for helping organize this information today.

- Information technologies are essentially obsolete every 18 months. This dynamic creates an unstable and unpredictable environment for the continuance of hardware and software over a long period of time and represents a greater challenge than the deterioration of the physical medium. Many technologies and devices disappear as the companies that provide them move on to new product lines, often without backwards compatibility and ability to handle older technologies, or the companies themselves disappear.

- There is a proliferation of document and media formats, each one potentially carrying their own hardware and software dependencies. Copying these formats from one storage device to another is simple. However, merely copying bits is not sufficient for preservation purposes: if the software for making sense of the bits (that is for retrieving, displaying, or printing) is not available, then the information will be, for all practical purposes, lost. Libraries will have to contend with this wide variety of digital formats. Many digital library collections will not have originated in digital form but come from materials that were digitized for particular purposes. Those digital resources which come to libraries from creators or other content providers will be wildly heterogeneous in their storage media, retrieval technologies and data formats. Libraries which seek out materials on the Internet will quickly discover the complexity of maintaining the integrity of links and dealing with dynamic documents that have multimedia contents, back-end script support, and embedded objects and programming.

- Financial resources available for libraries and archives continue to decrease and will likely do so for the near future. The argument for preserving digital information has not effectively made it into public policy. There is little enthusiasm for spending resources on preservation at the best of times and without a concerted effort to bring the issues into the public eye, the preservation of digital information will remain a cloistered issue. The importance of libraries has been diminished in the popular press as the pressures from industry encourage consumers to see libraries as anachronistic while the Internet and electronic products such as Microsoft Encarta are promoted as inevitable replacements. Until this situation changes, libraries and archives will continue to be asked to do more with less—both in terms of providing traditional library services, as well as new digital library services: preservation will have to encompass both kinds of collections.

- Increasingly restrictive intellectual property and licensing regimes will ensure that many materials never make it into library collections for preservation. These will be corporate assets and will not be deposited into public collections without substantive financial and licensing arrangements that few libraries will be able to afford. From a positive perspective, this fact will allow libraries to essential ignore the preservation question for many kinds of key information resources (examples will include newspapers, electronic serials, directories) as these may be preserved by their corporate owners. The flip-side of this argument is whether corporate owners will develop a public-spirited interest in providing this archival role for future generations and whether the resources will be accessible to the public.

- The archiving and preservation functions within a digital environment will become increasingly privatized as information continues to be commodified. Companies will be the place where the most valuable information is retained and preserved, and this will be done only insofar as there is a corporate recognition of the information as an asset. But companies have no binding commitment to making information available over a long-term. Those librarians that suggest legal deposit is the means for addressing this issue are not likely to be successful. As a full discussion of this topic is beyond the scope of this paper, let it suffice to state that libraries would have a very limited ability to cope with the volume and variety of digital resources that publishers could potentially dump on them. Still more problematic are the rights management and access control issues that content providers will require—demands which strongly argue that legal deposit in a digital era will have limited effectiveness. Libraries will be the archive of last resort and will be repositories of ephemera and "public domain" information—those materials considered as largely without commercial value.

- The Commission on Preservation and Access suggests that the first line of defense against the loss of valuable digital information rests with the creators, providers and owners of digital information.[4] This fact is a critical one for preservation purposes as it strongly suggests that the role that librarians and archivists must play will be an increasingly public one. Preservation is a desktop issue, not merely an institutional one.

The role of preservationists must be to interact with users and to address preservation and information management issues on their desktops, not the archivists desktop.

- Standards will not emerge to solve fundamental issues with respect to digital information. The challenge in preserving electronic information is not primarily a technological one, it is a sociological one. The dynamism of the market for information technologies and products ensures the fundamental instability of hardware and software primarily because product obsolescence is often key to corporate survival in a competitive capitalist democracy. Product differentiation manifests itself at the very level of the document standard. Proprietary systems provide commercial enterprises with profitable products whereas static (i.e. preservable) formats do not create a continuing need for upgrading which software and hardware companies depend upon. This situation conspire against standards that create a stable nexus of hardware, software, and administration.

- Libraries and archives will be required to continue their existing archival and preservation practices as the current paper publishing boom continues. Clearly, digital collections are not going to be a substitute for existing and future library collections and plans must be made to accommodate both. A significant concern of libraries and archives is that the financial resources necessary to address expensive IT upgrades, embark on data rescue operations, and undertake digital preservation will have detrimental impacts on other aspects of library and archival operations such as building collections and providing services for the public.

THE PRESERVATION NEXUS

Let us be absolutely clear from the outset: no one understands how to archive digital documents. Microfilm remains the long-term medium of choice for projects seeking to preserve large numbers of documents.

Sustainable solutions to digital preservation problems are not available. The research program for digital preservation is still being established. For example, the Preservation of Electronic Materials: a Programme of Studies funded by the U.K. Joint Information Systems Committee of the Higher Education Funding Councils has recently put forward a research agenda which illustrates the situation. This programme includes:

- Developing a topology of major data types and formats and identifying issues affecting preservation of each category of material.
- Investigating the attitudes of originators and rights owners to the responsibilities of digital preservation.
- Examine costing models for long term preservation of digital materials.
- Examining the three main methods of digital preservation: technology preservation; technology emulation; information migration.

- Investigating the digital preservation needs of universities and research funders.
- Investigating progress towards permissive guidelines for digital preservation.
- Report on sampling methods and techniques for collecting materials, on the nature and extent of institutional electronic archives, and or the relevance of current archival practice to digital preservation.
- An investigation of post hoc rescue, or data archaeology, of high value digital material which cannot be accessed because the required IT environment is no longer available.[5]

Other organizations such as the National Digital Library Federation and Research Libraries Group are at a similar point in their conclusions.

In an abstract sense, the preservation of digital materials is not complex. As long as the relationship between hardware, software and humanware (organizations and people) is maintained, a kind of "preservation nexus" exists and a digital object can be preserved forever. The problem is the centrifugal forces such as time and money that pull each of these elements away from each other—software and hardware becomes outdated, migrating information may require expensive recoding, and organizations lack resources to address the problems. This creates an environment where the object is basically left in a digital limbo—trapped in an obsolete format or captured on an unreadable medium or lacking the administrative capacity, resources, or willingness to refresh the data.

The archiving of digital information is not a conservation problem. To quote a conclusion of the Technology Assessment Advisory Committee to the Commission on Preservation and Access, "Preservation means copying."[6] It is the "contents" that must be preserved not conserved. Unlike conservation practices where an item can often be treated, stored and essentially forgotten for some period of time, digital objects will require frequent refreshing and recopying to new storage media. Keeping the "original" digital artifact is not important.

Further, refreshing or "copying" of digital information will not be confined to merely moving from one storage medium to another but will also entail translation into new formats or structures. It is also likely that this translation will be an imperfect copy, as well as a costly and ongoing expense that must be budgeted for accordingly. Some types of digital objects will not be transferable to either new media or to new formats, for example: software executable files; self-extracting archives coded for particular operating systems; and formats unique to a software implementation. Even maintaining materials on a particular storage media like CD-ROM will not be easy. Multimedia materials that require particular hardware and software platforms to access the contents will disappear and will not be easily migrated to any new system. It is quite possible that a significant portion of multimedia CD-ROM titles presently in-print today, will not be accessible to the next generation operating systems.

The handling of the physical storage of the digital object is the least of our worries. However, as a profession, librarians and archivists know how to do this well—how to have climate controlled environments, how to conserve and store materials of various

types, how to ensure that disaster recovery practices are in place, and so on.[7] What is lacking in our knowledge base, and in that of technologists, is how to preserve over time, and on a cost-effective basis, the relationship between the storage, retrieval and display hardware and software—a relationship which is required to make "being digital" also being understandable.

Preserving digital objects over time and in a cost effective manner, requires technologies and formats to be stabilized to a greater degree than we have presently. But market forces drive the introduction of newer, faster, bigger, "better" technologies for the production, distribution and storage of electronic information. No standards process can keep up with the dynamic changes which have occurred in the past 20 years. Documents are becoming complex, dynamic creations made of multiple objects, embedded programming and hypertext links. This is a significant departure from the solitary book or artifact with which preservationists traditionally work. Organizations are being asked to make fiscal commitments to creating complex technical infrastructures that change every 3-5 years and which require increasingly expensive technical expertise to keep functioning.

WHAT IS TO BE DONE?

What must libraries do to address the challenges of digital preservation? There are a number of key areas in which concerned institutions and professionals can contribute:

1. Knowledge Creation: Contribute to Research and Development

There is an urgent need to augment research in the area of digital preservation. Projects which further our knowledge in the challenges of preserving various types of materials—maps, archival materials, color documents, bound volumes, data-sets, music, and electronic formats like SGML, PDF, ASCII, HTML—must be undertaken. The research needs to include a careful accounting of the actual costs of preserving these materials. If projects do not provide cost effective preservation solutions or have only marginal benefits, we need to be informed of this. If some techniques or strategies work better than others, we need this information clearly stated.[8]

2. Digital Triage: Developing Guidelines for What Can and Should Be Saved

There should be informed skepticism about the claims of organizations that say they will archive the Internet.[9] The library and archival communities already know that not everything can and should be saved. What is key is selecting which digital resources to preserve and which ones not to preserve.

Librarians and archivists must develop digital collection development and evaluation guidelines to assist in deciding what can be saved and what should be saved, and what can't be, on a case-by-case basis. The Research Libraries Group Preservation Working Group on Digital Archiving, as well as that of the JISC in the UK, have identified the

development of guidelines for appraisal, selection, and priority setting for preserving digital information as being a key task for future work.[10]

3. Rescue Operations: Ensure Vital Electronic Documents are Preserved Now

Digitally produced images of documents are not a substitute for microfilm preservation. Digital copying will not necessarily ensure the preservation of a digital document. The fact is that digital information may be outputted to microfilm for preservation purposes and it may even be appropriate to print an electronic document on acid-free paper and handle it according to established archival practices. These hybrid methods may be an effective transition step until stable guidelines and technologies evolve for long-term digital preservation.

Librarians and archivists need to work with industry to develop simple and cost effective print-to-microfilm systems; this will enable archives to preserve documentary collections that are provided in proprietary formats such as word-processors in a cost-effective fashion to be effectively preserved.

By transferring electronic information into non-electronic form there will be a loss of functionality for some kinds of information. Paper or microfilm documents may no longer have active hypertext links or be searchable by keyword and there will be some cases where it is not reasonable to migrate the information to non-electronic forms as it would render the information useless, e.g. software or large data-sets. Re-coding information of this type will be sharply constrained by the resources and in many cases will not be feasible.

4. Document Formats: So Many to Chose From

Mixed media and multiple document formats will continue to remain the fly-in-the-ointment of digital collections. Multiple formats may require maintaining multiple hardware/software platforms and will confound simple migration to new storage media.

Whether possible or even preferred, requiring that data be stored in a common format is unlikely for the foreseeable future. Similarly, existing translation software available for the migration and translation of document formats illustrates that the problems are significant and the results are often less than satisfactory. The simple case of conversion between the most popular document formats, MS-Word and WordPerfect, provides ample illustration of the challenges that are faced and argues for skepticism about claims for future systems which will make this task easier.

There will probably be no effective solution to this problem. If a library will be receiving electronic objects from content providers such as authors or publishers, they may want to specify a limited range of acceptable document formats. Working with creators to bring these problems to their attention and providing guidance on how to organize files and which formats to use for purposes of archiving will help. Once again, more research and

effective communications with content and technology providers are required to address the issue of multiple formats.

5. Being Legal: Rights Management and Access Control

The management of rights and access controls for electronic objects is an increasingly complex area of concern for libraries and archives. A library may have the rights to access and use electronic materials, but the right to preserve the materials may not be the same thing. Restrictions on access placed by rights-holders or by licensing arrangements for particular resources needs to be addressed when questioning whether the information can be preserved. A simple example is the case of whether a library can retain old versions of a CD-ROM database to which it may subscribe or whether the CD-ROM must be discarded after the subscription is finished or destroyed when a new version is issued. More complex legal issues arise with the automated collection of Internet information for preservation purposes in efforts such as the Internet Archive, where it seems that intellectual property rights are being ignored. Similar intellectual property concerns about the legality of unauthorized and automated indexing of Internet WWW sites are also being raised.

Licensing will be one of the most important things that a library will be required to do in the electronic realm. The management of diverse licensing arrangements promises to be a significant administrative and technical challenge for preservation purposes. For example, the University of California Digital Library framework is quite explicit about the need to ensure that where digital materials are printable, that the licenses and contracts associated with them allow you to print a copy on acid-free paper for preservation purposes.[11]

Librarians need to work on contractual issues. There must be a concerted effort on the part of all libraries to work together to get the best contractual arrangements possible and to be more aware of the contractual issues associated with the licensing of electronic resources. The Council on Library Resources, Commission on Preservation and Access and Yale University's initiative with the LibLicense Project is an important starting point for this work.[12]

Technical advances such as those in the EC COPEARMS (Coordinating Project for Electronic Authors' Right Management Systems)[13] and IMPRIMATUR (Intellectual Multimedia Property RIghts Model And Terminology for Universal Reference)[14] projects will assist significantly in rights management. However, these technologies remain in the prototype stage and formidable challenges exist in the formalization and intellectual description of rights and in developing common contractual languages. The efforts to provide these descriptions is only beginning and is key to effective rights management. As a result, rights management systems are probably 5-10 years from commercial deployment.

6. Wave the Flag: Promoting the Importance of Preservation

Librarians and archivists must engage in a concerted effort to raise the profile of preservation.[15] The Commission on Preservation and Access[16] and Research Libraries Group (RLG)[17] in the United States, the Joint Information Systems Committee (JISC)[18] in the UK and PADI Working Group[19] and National Library[20] in Australia, among others, have all been active in framing the problems of digital preservation within their constituencies. This work needs to be supported, expanded and brought to greater public attention.

As a profession, librarians and archivists need to encourage critical thinking and be highly pragmatic about the nature of the new medium and the challenges of digital preservation. Although there are positive benefits to digitization, particularly in providing remote and enhanced access to information, as Klaus-Dieter Lehmann, Director General of Die Deutsche Bibliothek, warns there is a danger here as well. The problem is that "digitization may come to be regarded as a panacea for all of the real and imagined problems libraries now face in connection with the preservation of physical collections: the growing need for storage space, the deterioration of books due to acid paper, and the rising costs of library operation."[21] Only by increasing public support and understanding of the issues of preservation, both digital *and* analog (i.e. physical collections) can we hope to address the shortfall in fiscal and human resources that continue to hinder preservation efforts and impact upon library services.

7. All for One, One for All: Working Together

Archiving decisions for materials which are common to many libraries will be made in consultation with other libraries to determine the appropriate forms and sharing-mechanisms. Few libraries will be positioned to effectively archive large quantities of electronic information. Any given library will necessarily be required to select resources that they can archive and preserve according to their particular mandates and user requirements. In many cases, it will not make sense to duplicate efforts. Digital preservation efforts will need to be coordinated. In other situations, it may make perfect sense to duplicate archival efforts, particularly if the information is too valuable for historic purposes to be entrusted to only one institution. Efforts such as the U.S. National Digital Library Federation and the Canadian Initiative on Digital Libraries are examples of important first steps.

8. Digital Preservation as a Public Good

Librarians and archivists protect the public interest by making information available to the community and by asserting the importance of maintaining a record of our collective intellectual heritage. This task will be a continuing challenge because libraries and archives are too often considered to be competitors to publishers, document delivery services, and other private sector content providers. It is unlikely, unless a substantive ground-swell of public support is generated, that libraries will win in the battle against commodification of

information resources or be able to generate the public support (and fiscal resources) necessary to meet the challenges.

Despite the present lack of public interest in digital preservation, it is necessary to believe, perhaps as an article of faith, that the efforts of librarians and archivists will be appreciated in the future. The traces of information that we are able to save from our digital vellum will be valuable sources of information to the future. Even if the task of digital preservation remains thankless, it is a vital one and must still be undertaken. The objective is a noble and necessary one even as the problems many appear insurmountable.

CONCLUSION

Digital collections facilitate access, but do not facilitate preservation. Being digital means being ephemeral. Digital places greater emphasis on the here-and-now rather than the long-term, just-in-time information rather than just-in-case. The research program for digital preservation has only recently been initiated to develop strategies, guidelines, and standards. Although tremendous work has been undertaken in defining the problems and challenges, much more remains to be done, and the tough task of actually doing digital preservation (and digital rescue) remains ahead.

A critical appraisal of where we are vis-a-vis our digital culture, and what we want for the future—something which may not be defined in technical terms at all—is required both inside and outside of the library and archival professions. If history and cultural heritage are to be important, then it will likely fall to librarians and archivists, the monastic orders of the future, to ensure that something of the heady days of our "digital revolution" remains for future generations. The challenges to digital preservation are considerable and will require a concerted effort on the part of librarians and archivists to rise up to these challenges and assert in public forums the importance of protecting a fragile digital heritage.

Those who cannot remember the past are condemned to repeat it.
George Santayana, The Life of Reason, 1906

ENDNOTES

[1] The author would like to thank his colleagues, Michael Williamson and Fay Turner, for their comments and sharp pencils, and accepts all responsibility and is unrepentant for hyperbolic outbursts.

[2] Making distinctions between "electronic records" or "electronic texts" or other semantic debates are of questionable utility. The emergence of multi-media forms of communication and the increased complexity of the "document" in the digital era argues that we must begin to consider that what we are really looking to manage and preserve is electronic objects.

[3] Task Force on Archiving of Digital Information. *Preserving Digital Information: Final Report and Recommendations*. Commission on Preservation and Access and the Research Libraries Group. May 1, 1996. <URL: http://www.rlg.org/ArchTF/index.html >

[4] *Preserving Digital Information: Final Report and Recommendations*.

[5] <URL: http://lyra.rlg.org/preserv/jisc.html >. For background information on the workshop that led to this programme, see *Long Term Preservation of Electronic Materials. A JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib).* 27th and 28th November 1995: University of Warwick. <URL: http://ukoln.bath.ac.uk/fresko/warwick/intro.html >

[6] Michael Lesk, *Preservation of New Technology: A report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access.* October 1992. <URL: http://palimpsest.stanford.edu/byauth/lesk/lesk2.html >

[7] Those interested in finding out more information about the conservation of electronic materials are directed to the following resources:

- Commission on Preservation and Access <URL: http://clir.stanford.edu/cpa/ >,
- Library of Congress Preservation Directorate <URL: http://lcweb.loc.gov/preserv/preserve.html >
- Conservation OnLine <URL: http://palimpsest.stanford.edu/ >

A useful overview of imaging and storage technologies is the "Long-Term Access Strategies for Federal Agencies: Digital-Imaging and Optical Digital Data Disk Storage Systems" (Technical Paper No.12, July 1994) published by the U.S. National Archives and Records Administration.

[8] For example, in December 1996, researchers at the Yale University library engaged in a project to migrate data tapes of public opinion research from an older system to a newer one. Their literature search was unable to locate significant information on the subject of copying data files *and* changing the way they were coded. For further information about this project, see Ann Gerken Green and JoAnn Dionne, *Preserving the Whole: A Two-track Approach to Rescuing Data and Metadata.* Commission on Preservation and Access, December 20, 1996. <URL: http://clir.stanford.edu/pubs/ >

[9] The work of Brewster Kahle's Internet Archive is a notable effort here. An important caveat is that the "archive" is, at best, a fuzzy snapshot of the electronic materials which are most easily available. The vast majority of Internet resources are not amenable to the kind of "sweep" that Kahle's archive is attempting to undertake as the resources reside in databases which cannot be traversed by automated systems. There is also an interesting legal question about whether the Internet Archive has the right to copy entire websites into its database. This question will likely be the most significant one for the success of such operations in the future. For details, see <URL: www.archive.org >.

[10] RLG Preservation Working Group on Digital Archiving. Progress Report. May 6, 1997. <URL: http://lyra.rlg.org/preserv/digarch1.html >

[11] The University of California Digital Library: A Framework for Planning and Strategic Initiatives, October 1996. <URL: http://elib.cs.berkeley.edu/ >. Another interesting comment in this report is that "caution must be exercised by those who would assume that the UCDL obviates the need for new buildings and building renovations."

[12] See LibLicense Website at <URL: http://www.library.yale.edu/~llicense/index.shtml >. The LIBLICENSE-L Mailing List is a key place for discussions on library licensing issues. Contact: liblicense-l@pantheon.yale.edu

[13] COPEARMS. <URL: http://www.nlc-bnc.ca/ifla/VI/2/p5/proj5.htm >

[14] IMPRIMATUR. <URL: http://www.imprimatur.alcs.co.uk/ >

[15] In the past 5 years, I can recall only one article of significance on digital preservation in literature intended for the general public. Jeff Rothenberg, "Ensuring the Longevity of Digital Documents", *Scientific American*, January 1995.

[16] <URL: http://clir.stanford.edu/cpa/ >

[17] <URL: http://lyra.rlg.org/preserv/ >

[18] <URL: http://www.niss.ac.uk/education/jasper/intro.html >

[19] Preserving Access to Digital Information (PADI) Working Group. <URL: http://www.nla.gov.au/dnc/tf2001/padi/padi.html >

[20] <URL: http://www.nla.gov.au/3/npo/epres.html >

[21] Klaus-Dieter Lehmann, "Making the Transitory Permanent: The Intellectual Heritage in a Digitized World of Knowledge." *Daedulus*. Fall, 1996, p. 310.

Challenges in the Preservation of Electronic Information". 63rd International Federation of Library Associations and Institutions) Council and General Conference . . 4.

- Digital Scholarship relies on accurate, reliable and unchanged digital information, including digital objects inside digital libraries.Â system crash have taken their toll on DLIST, the University of Arizona's Digitial Library of Information Science and Technology. We are currently exploring choices and alternatives both to short term recovery and long term sustainability .â€)

.