Werk

Titel: Enduring Access to Digital Information: Understanding the Challenge

Autor: Mandel, Carol A.

Ort: Graz

Jahr: 1996

PURL: https://resolver.sub.uni-goettingen.de/purl?514854804_0006|log70

Enduring Access to Digital Information:
Understanding the Challenge

Carol A. Mandel
*Columbia University Library**

Introduction: Preservation in the new realm of digital information

The role of libraries as repositories--places for safekeeping--is well established in society. Libraries, together with museums and archives, are charged with society's mission to preserve its accumulated knowledge and culture. In the context of the material world, this mission and its importance are well understood by librarians and the communities they serve: libraries must make accessible books, manuscripts, and other artifacts for use long into the indefinite future. However, information in its virtual form, digital information, presents a new context in which to undertake the library's traditional mission of safekeeping. In this new world, many old assumptions no longer hold, known techniques do not apply and unanswered questions abound. Preservation of digital information is a challenge we do not yet know how to meet.

Digital information differs fundamentally and conceptually from print on paper in three ways that are relevant to developing preservation strategies. First, in preserving print on paper, we are seeking to create a more permanent medium from the fragile medium of paper. We do this either by conserving the paper object, or by reformatting information on to a more permanent medium such as microfilm. When we confront digital information, medium is of much less concern. Instead the challenge is impermanent technology. Digital media are, of course, not long-lasting, and require a continual process of reformatting. But more important is the short life of the software and hardware that are necessary to translate digital information into usable form. The techniques for preserving technology are more complex and less well understood than those for preserving media; they present a new challenge.

A second new challenge is also related to the difference between preserving a physical object owned by a library and preserving virtual content that a library has licensed for use. When a library preserves a book or a manuscript it owns, it has the legal right to preserve that object. Copyright of the content of the book may limit the multiple reproduction or distribution of its preserved version, but there are no further restrictions on use of the preserved object than there were on the original. In the digital world, libraries increasingly are licensing online content from publishers. In some cases, a physical magnetic tape or optical disk is acquired; increasingly, the licensed information is made available by the provider online via a network and no material object is accessioned by the library. In both cases, the library does not have the legal right to maintain the database in perpetuity; in the latter case, it also does not have the operational means to do so. The role of the library as repository is no longer clear in the world of licensed networked information.

The third new challenge libraries face in preserving digital information is that the lifecycle of this information is, as yet,

unknown. Digital publication has only a short history and libraries have limited experience in handling it. Preservation requires a commitment to provide access into an indefinite future, and we cannot yet predict the cycle of activity and expense that such preservation will require.

The Task Force on Archiving of Digital Information

Recognizing that preservation of digital information presents such fundamentally new challenges, two American research library organizations formed a task force and charged it to develop a report that would frame the key problems engendered by digital preservation, define the critical issues that inhibit resolution of these problems, and develop strategies for addressing them. In December of 1994, The Commission on Preservation and Access (CPA) and The Research Libraries Group (RLG) formed the Task Force on Archiving of Digital Information. The charge to the Task Force is shown below. (Brief descriptions of RLG and CPA are shown in the Appendix.) Membership included representation from research libraries, library and information organizations, archives, technology corporations and the publishing community. A draft report issued in August 1995 elicited much thoughtful, proactive comment, especially from the international community. The draft report stimulated new work in the UK and in Australia, and related closely to work concerning deposit of electronic publications undertaken by the European Union. The extent of activity and comment related to the draft has demonstrated the need for ongoing international communication and coordination in addressing this enormous and complex new challenge. The final report of the Task Force was issued in May 1996, and is available on the Research Libraries Group website, *http://www.rlg.org.*

The report, and hence the summary presented in this paper, focuses on three essential questions:

1) What does digital preservation entail?
2) How do we organize ourselves to do it?
3) What steps should we take to move forward?

What does digital preservation entail?

Preservation, in both the paper and digital realms, includes: preserving the content, or substance, of information; preserving the context required to understand and interpret the information well into the future; and providing the means to access and use the information. In each of these functions, preservation of digital information is analogous to preserving traditional media, but presents new sets of problems.

Preserving content

The first act of preservation in any environment is identifying or selecting material that merits preservation activity. The principles and judgements required to predict the future need for information in various disciplines are as complex and varied in their application to digital information as they are to more familiar media. The fluidity and dynamic nature of digital data, however, add some new dimensions. The choice of version to retain is more difficult since many more versions of a resource are likely to exist. Interactive and dynamic databases, which change from moment to moment, can only be "preserved" through samples and snapshots; no commonly accepted body of practice yet governs such choices and little experience exists demonstrating the future usefulness of different sampling techniques. Hyperlinked files add further layers of decision-making, as the extent to which linked resources should also be selected and included for preservation must be determined.

The fluidity of digital information and the ease with which it can be changed introduce the need for authentication of the

version that is selected for preservation. A repository of digital information must protect information from tampering and must employ techniques for ensuring users that the information held is actually what it is represented as being. A variety of techniques for protecting and authenticating data exist, but their employment is not yet widespread and easily put into production.

In the short term, any provider of digital information must store the data and provide the systems management and engineering to maintain and deliver it, such as back-up, maintenance of redundant files, etc. Longer term retention of files requires additional techniques. Initially, media must be "refreshed" to maintain usability and conform to new formats. Because digital files are dependent on software and hardware to use them, these too must be kept active. Maintaining a museum of unsupported hardware and software platforms is not practical; data must be migrated to work on new platforms. If data are stored in "flat files", i.e., files in which the data content is easily separable from software to use it, then this migration is not necessarily difficult. However, increasingly, content and functionality are inseparable, and both must be preserved. Migration then becomes an increasingly complex systems engineering task. One potentially important technique for migration is emulation, that is, development of software that can emulate the environment in which the original software operated. New systems designs and technologies are needed to facilitate this work. To the extent that major software vendors are encouraged to maintain backward compatibility as they develop new versions, the task of migration will be greatly facilitated.

The challenges presented by migration illustrate the fundamentally different conceptual bases that distinguish preservation as it applies to digital rather than analog media. Archival preservation via migration requires a commitment to unknown future activities with unpredictable future costs. This essentially requires libraries to change their definition of archival,

and their understanding of the commitments and resources necessary to function as a repository.

Preserving context

The issues involved in preserving context are conceptually the same for digital and traditional formats, but it is important to be aware of the new options introduced by the characteristics of digital information. The concept of fixity in selecting and maintaining a version of a digital resource is related to understanding its context. How and why was a version selected, and how does the preserved version relate to those that no longer remain? Provenance is important contextual information; the provenance of published books can be adequately documented by publication information, but digital resources have no such conventions. The history of reproduction and migration is essential for digital objects that document phenomena, such as reproductions of artistic images, scientific data, etc., as the layers of translation and transformation for digital resources can be enormously complex. Maintaining the original "look and feel" of a digital resource, or at least documenting it for the future, will be increasingly challenging as dynamic documents must be sampled for preservation and as objects migrate to new platforms.

Preserving access

There is little use to preserving a resource unless those who need it can easily discover its existence and make use of it. Network discovery and retrieval tools may make cataloging and indexing of digital resources easier, but it is the responsibility of a digital repository to ensure that access is adequate. Once a resource is discovered by a potential user, an infrastructure for delivery and use is required. Thus a repository must maintain adequate network connectivity; must provide software for

retrieving, viewing and, if appropriate, manipulating digital information; must maintain a reliable system for referencing and locating the resources it holds; and must maintain systems for protecting the security and integrity of the data it holds and for implementing any restrictions placed on access by the owners of the intellectual property. Since these access requirements are not specific to long term preservation, but are necessary as well for short term use of digital information, there is a large community of publishers and information providers interested in developing economical techniques for all of these functions. Libraries and other repositories will not be alone in their efforts to maintain access.

How do we organize ourselves to preserve digital information?

Because there are so many kinds of individuals and organizations interested in providing access to digital information, the roles of information providers versus those of libraries and other repositories are not clearly defined in this new environment. In the world of material information, the distinction between preservation and use is quite clear. The more an object is used, the more it is subjected to wear and tear; use and preservation are not only different activities, they may be antithetical to each other. In this world, only libraries and archives have become concerned with long term preservation.

The boundaries between roles in the digital environment are, at present, less clear. Use keeps digital objects alive. If they are in-demand, the activities that enable their use--e.g., access, systems engineering, refreshing--are the same activities that will ensure their continued functionality. Digital objects that sit unused will become unusable. These blurred boundaries between immediate and long-term use have led many information providers to view themselves as the archival keepers of their data. They see no need for intermediaries, such as libraries, to provide archival functions.

Their interest in maintaining archival control is related to their interest in maintaining intellectual property control over their information. As noted earlier, the focus of digital preservation is on content, not on objects, and content owners do not wish to lose control over their property.

In an ideal environment, content developers, publishers and libraries would work closely together in developing interdependent roles. Those who create digital information would design digital resources with access and long term use in mind, e.g., by providing needed metadata, using standard formats, documenting software, etc. Those who publish digital information would deposit it with appropriate repositories and develop agreements for long-term preservation and access. We need to strive for this ideal environment. However, there will be cases where digital information has not been created and deposited with such foresight, and libraries will need the legal basis to acquire and aggressively rescue files that have essentially been abandoned by those who created or published them.

Our greatest challenges are organizational rather than technical. As the report of the Task Force on Archiving of Digital Information notes, our most important objective is "Organizing ourselves over time and as a society to maneuver effectively in a digital landscape... Building ... the deep infrastructure that will enable us to ... move our cultural records naturally and confidently into the future." We currently lack the infrastructure of practices, standards and organizations that is needed to support preservation of digital information. Elements of the infrastructure that we need to begin to build include the following:

Legal bases for deposit and rescue. In individual countries and internationally, legislation and agreements are needed: to encourage legal deposit of electronic resources with archival repositories, to enable rescue of abandoned resources, and to facilitate access and use of archival files.

Standards for description. Current library cataloging standards are not sufficient to describe access and contextual information about digital resources.

Standards for design and formats. Migration on a broad scale is only feasible if standard formats and platforms are widely used.

Backward compatibility. Software manufacturers need to be educated and encouraged in the importance of maintaining the usability of older versions of their products.

Accepted best practices for systems engineering and migration. Information about these techniques is not widely shared across professional communities.

Enabling technologies. Functions such as migration, emulation and access can be facilitated by new systems design and technological development focused on these issues.

Guidelines for archival principles and practices. The expected operational requirements to serve as a responsible repository for digital information are not well understood. Guidelines need to be articulated and promulgated, perhaps even through a certification process, so that organizations can develop themselves to meet the needs.

Processing centers. Many libraries and archives will not have the technical capabilities or specialized expertise to maintain and migrate digital files, even though they may have the appropriate mission and skills to identify and take responsibility for files to be preserved. In a manner similar to vendors that provide services for microfilm production and storage, processing centers could provide operational services for digital preservation.

Models for cooperative arrangements. Digital preservation is complex and expensive; it can only be practically undertaken on a distributed basis. Cooperative agreements will be essential to future use of digital files.

What steps does the Task Force recommend?

The Task Force developed a tripartite strategy for creating the support environment and gaining the knowledge we need to effect preservation of digital resources:

Identify existing best practices

Conduct pilot projects that test new models and techniques.

Develop the legal and organizational support infrastructure.

Best practices

Although digital resources are relatively new in the long history of communication media, they have been with us for some thirty years, and many have already been preserved. This work has been done in specialized sectors, such as social science survey data repositories and scientific observatories, and the practices used in each sector have not been shared across disciplines. As libraries begin to face the preservation of digital files created by mass communication and electronic publishing, they can learn from practices elsewhere. The Task Force recommended that the Commission on Preservation and Access commission cases studies of successful work in areas such as: design of digital files, mass storage, resource description and migration paths.

New models and techniques

The best way to gain experience and learn about new approaches is to encourage and fund pilot projects. The Task Force identified three areas in which funding programs should be developed--most likely through government agencies-- to stimulate projects.

Cooperative efforts to rescue files. There are digital files now in need of aggressive rescue. Several well defined efforts could be developed in which libraries share in the rescue and maintenance of selected files.

Sponsored digital archives. Repositories, information creators and publishers will need to collaborate to create digital repositories. Funding programs that require collaboration often stimulate new partnerships and approaches.

Research and development in new technologies. Technology developers have been interested in products with commercial applicability, such as rights management systems that serve commercial needs. Programs of sponsored research could target efforts at technologies that facilitate archival practices, such as migration, emulation and authentication.

Support structures

Developing the needed organizational and societal infrastructure for digital preservation will require strategic initiatives in several areas. The Task Force recommended that The Commission on Preservation and Access and The Research Libraries Group develop implementation plans to achieve the following objectives:
-    Make preservation a explicit goal of the US Federal Government National Information Infrastructure initiative.
-    Articulate and lobby for the legal principles that will facilitate digital preservation.
-    Develop criteria for organizations that wish to serve as digital repositories.
-    Educate scholarly societies about the importance of digital preservation and engage them in preserving the information they produce.
-    Maintain international coordination.

The issues and strategies identified by the Task Force are not specific to the United States, as the nature and extent of international comment on the Task Force report has demonstrated. The preservation panel at the May 1996 LIBER meeting continued the international discussion, and the process must be ongoing. The challenges of preserving digital resources are large and complex; meeting them will require many approaches and many perspectives. Strong, active national libraries across the world have a special role to play, as do research libraries, archives and a variety of specialized repositories. We will need to share information about best practices and new technologies across a wide spectrum of communities that have not worked together in the past. Digital preservation has emerged as a new, critically important field of interdisciplinary and international activity.

Appendix

The Research Libraries Group, Inc. (RLG) is a not-for-profit membership corporation of universities, archives, historical societies, national libraries, and other institutions devoted to improving access to information that supports research and learning. RLG owns and operates databases and software to serve the information access and management needs of both its members and non-member institutions and individuals worldwide. RLG's address is: 1200 Villa Street, Mountain View, CA 94041-1100.

The Commission on Preservation and Access (CPA) is a private, non-profit organization acting on behalf of the nation's libraries, archives, and universities to develop and encourage collaborative strategies for preserving and providing access to the accumulated human record. The Commission's address is: 1400 16th Street, NW, Suite 740, Washington, DC 20036-2217.