Preserving Digital Information

Report of the Task Force on Archiving of Digital Information

commissioned
by
The Commission on Preservation and Access
and
The Research Libraries Group

May 1, 1996

Executive Summary

In December 1994, the Commission on Preservation and Access and the Research Libraries Group created the Task Force on Digital Archiving. The purpose of the Task Force was to investigate the means of ensuring "continued access indefinitely into the future of records stored in digital electronic form." Composed of individuals drawn from industry, museums, archives and libraries, publishers, scholarly societies and government, the Task Force was charged specifically to:

- -- "Frame the key problems (organizational, technological, legal, economic etc.) that need to be resolved for technology refreshing to be considered an acceptable approach to ensuring continuing access to electronic digital records indefinitely into the future.
- -- "Define the critical issues that inhibit resolution of each identified problem.
- -- "For each issue, recommend actions to remove the issue from the list.
- -- "Consider alternatives to technology refreshing.
- -- "Make other generic recommendations as appropriate" (see Appendix A for the full charge).

The document before you is the final report of the Task Force. Following its initial deliberations, the Task Force issued a draft report in August, 1995. An extended comment period followed, during which a wide variety of interested parties located both here in the United States and abroad contributed numerous helpful and thoughtful suggestions for improving the draft report. The Task Force benefited greatly from the comments it received and incorporated many of them in this final report.

In taking up its charge, the Task Force on Archiving of Digital Information focused on materials already in digital form and recognized the need to protect against both media deterioration and technological obsolescence. It started from the premise that migration is a broader and richer concept than "refreshing" for identifying the range of options for digital preservation. Migration is a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. The Task Force regards migration as an essential function of digital archives.

The Task Force envisions the development of a national system of digital archives, which it defines as repositories of digital information that are collectively responsible for the long-term accessibility of the nation's social, economic, cultural and intellectual heritage instantiated in digital form. Digital archives are distinct from digital libraries in the sense that digital libraries are repositories that collect and provide access to digital information, but may or may not provide for the long-term storage and access of that information. The Task Force has deliberately taken a functional approach in these critical definitions and in its general treatment of digital preservation so as to prejudge neither the question of institutional structure nor the specific content that actual digital archives will select to preserve.

The Task Force sees repositories of digital information as held together in a national archival system primarily through the operation of two essential mechanisms. First, repositories claiming to serve an archival function must be able to prove that they are who they say they are by meeting or exceeding the standards and criteria of an independently-administered program for archival certification. Second, certified digital archives will have available to them a critical fail-safe mechanism. Such a mechanism, supported by organizational will, economic means and legal right, would enable a certified archival repository to exercise an aggressive rescue function to save culturally significant digital information. Without the operation of a formal certification program and a fail-safe mechanism, preservation of the nation's cultural heritage in digital form will likely be overly dependent on marketplace forces, which may value information for too short a period and without applying broader, public interest criteria.

In order to lay out the framework for digital preservation that it has envisioned, the Task Force provides an analysis of the digital landscape, focusing on features, including stakeholder interests, that affect the integrity of digital information objects and which determine the ability of digital archives to preserve such objects over the long term. The Task Force then introduces the principle that responsibility for archiving rests initially with the creator or owner of the information and that digital archives may invoke a fail-safe mechanism to protect culturally valuable information. The report explores in detail the roles and responsibilities associated with the critical functions of managing the operating environment of digital archives, strategies for migration of digital information, and costs and financial matters.

The report concludes with a set of recommendations on which the Commission on Preservation and Access and the Research Libraries Group need to act, either separately or together and in concert with other individuals or organizations as appropriate. The Commission and the Research Libraries Group should:

1. Solicit proposals from existing and potential digital archives around the country and provide coordinating services for selected participants in a cooperative project designed to place information objects from the early digital age into trust for use by future generations.
2. Secure funding and sponsor an open competition for proposals to advance digital archives, particularly with respect to removing legal and economic barriers to preservation.
3. Foster practical experiments or demonstration projects in the archival application of technologies and services, such as hardware and software emulation algorithms, transaction systems for property rights and authentication mechanisms, which promise to facilitate the preservation of the cultural record in digital form.
4. Engage actively in national policy efforts to design and develop the national information infrastructure to ensure that longevity of information is an explicit goal.
5. Sponsor the preparation of a white paper on the legal and institutional foundations needed for the development of effective fail-safe mechanisms to support the aggressive rescue of endangered digital information.
6. Organize representatives of professional societies from a variety of disciplines in a series of forums designed to elicit creative thinking about the means of creating and financing digital archives of specific bodies of information.
7. Institute a dialogue among the appropriate organizations and individuals on the standards, criteria and mechanisms needed to certify repositories of digital information as archives.
8. Identify an administrative point of contact for coordinating digital preservation initiatives in the United States with similar efforts abroad.
9. Commission follow-on case studies of digital archiving to identify current best practices and to benchmark costs in the following areas: (a) design of systems that facilitate archiving at the creation stage; (b) storage of massive quantities of culturally valuable digital information; (c) requirements and standards for describing and managing digital information; and (d) migration paths for digital preservation of culturally valuable digital information.

Given the analysis in this report and its findings, we expect the Commission and the Research Libraries Group to pursue these recommendations on a national and, where appropriate, an international front, and to generate dialogue, interaction and products that will advance the development of trusted systems for digital preservation. There are numerous challenges before us, but also enormous opportunities to contribute to the development of a national infrastructure that positively supports the long-term preservation of culturally significant digital information .

John Garrett (co-chair)
CyberVillages Corporation
jgarrett@cmgi.com

Donald Waters (co-chair)
Yale University
donald.waters@yale.edu

Task Force on Archiving of Digital Information
Members

Pamela Q.C. Andre
Director
National Agricultural Library

Howard Besser
Visiting Associate Professor
School of Information
University of Michigan

Nancy Elkington
Assistant Director for Preservation Services
Research Libraries Group

John Garrett (co-chair)
Chief Executive Officer
CyberVillages Corporation

Henry Gladney
Research Staff Member
IBM Almaden Research Center

Margaret Hedstrom
Associate Professor
School of Information
University of Michigan

Peter B. Hirtle
Policy and IRM Services
National Archives at College Park

Karen Hunter
Vice President and Assistant to the Chairman
Elsevier Science

Robert Kelly
Director, Journal Information Systems
American Physical Society

Diane Kresh
Director for Preservation
Library of Congress

Michael E. Lesk
Manager, Computer Science Research Division
Bell Communications Research

Mary Berghaus Levering
Associate Register for National Copyright Programs
U.S. Copyright Office
Library of Congress

Wendy Lougee
Assistant Director, Digital Library Initiatives
University of Michigan Library

Clifford Lynch
Director, Library Automation
Office of the President
University of California

Carol Mandel
Deputy University Librarian
Columbia University

Stephen P. Mooney
Copyright Clearance Center, Inc.

Ann Okerson
Associate University Librarian
Yale University

James G. Neal
Sheridan Director of Libraries
Johns Hopkins University

Susan Rosenblatt
Deputy University Librarian
University of California, Berkeley

Donald Waters (co-chair)
Associate University Librarian
Yale University

Stuart Weibel
Senior Research Scientist
OCLC, Inc.

Table of Contents

PRESERVING DIGITAL INFORMATION

Introduction

Today we can only imagine the content of and audience reaction to the lost plays of Aeschylus. We do not know how Mozart sounded when performing his own music. We can have no direct experience of David Garrick on stage. Nor can we fully appreciate the power of Patrick Henry's oratory. Will future generations be able to encounter a Mikhail Baryshnikov ballet, a Barbara Jordan speech, a Walter Cronkite newscast, or an Ella Fitzgerald scat on an Ellington tune?

We may think that libraries and archives have stemmed the tide of cultural memory loss. We rely on them to track our genealogies, to understand what science has discovered, to appreciate the stories people told a hundred years ago, and to know how we educated our children during the Depression. Even seemingly trivial, ephemeral, or innocuous information that libraries maintain has unanticipated uses. For example, early in this century railways provided the primary means of transporting oil. The competition for this lucrative business led to rebates, kickbacks and other dubious business practices. The United States countered such practices by enacting severe antitrust laws. Germany, however, prohibited secret rates by requiring all oil carriage firms to publish their tariffs in the railway trade press. During World War II, nobody in Germany thought to repeal the law. Every week, an American agent went to a Swiss library, read the relevant newspaper, and worked out how much oil the Nazis were transporting and where.

Society, of course, has a vital interest in preserving materials that document issues, concerns, ideas, discourse and events. We may never know with certitude how many children Thomas Jefferson fathered or exactly how Hitler died. However, to understand the evils of slavery and counter assertions that the Holocaust never happened, we need to ensure that documents and other raw materials, as well as accumulated works about our history survive so that future generations can reflect on and learn from them. The Soviet Union stands as an example of a society in which history was routinely rewritten and pages of encyclopedias were cut out and replaced according to current political whim. The ability of a culture to survive into the future depends on the richness and acuity of its members' sense of history.

The Fragility of Cultural Memory in a Digital Age

But our ability and commitment as a society to preserve our cultural memory are far from secure. Custodians of the cultural record have always had to manage the inherent conflict between letting people use manuscripts, books, recordings or videos, and being sure that they are preserved for future use. For works printed on acidic wood pulp paper, as most books have been since 1850, we measure the remaining lifetime of those materials in decades, not centuries.

And what of the information we are now creating and storing using digital technology? In 1993, forty-five percent of U.S. workers were using a computer (United States Census Bureau 1993); the number is surely larger now. Virtually all printing and a rapidly increasing amount of writing is accomplished with computers. Professional sound recording is digital, and digital video is on the verge of moving from experimental to practical applications.

The Limits of Digital Technology

As a means of recording and providing access to our cultural memory, digital technology has numerous advantages and may help relieve the traditional conflict between preservation and access. For materials stored digitally, users operate on exact images of the original works stored in their local computers. Separating usage from the original in this way, digital technology affords multiple, simultaneous uses from a single original in ways that are simply not possible for materials stored in any other form. Digital technology also yields additional, effective means of access. In full text documents, a reader can retrieve needed information by searching for words, combinations of words, phrases or ideas. Readers can also manipulate the display of digital materials by choosing whether to view digital materials on a screen, store them on their computer or external media, or to print them.

Digital technology, however, poses new threats and problems as well as new opportunities. Its functionality comes with complexity. Anyone with a compass (or a clear night to view the position of the stars in relation to true north) could theoretically set up or repair a sundial. A digital watch is more useful and accurate for telling time than a sundial, but few people can repair it or even understand how it works. Reading and understanding information in digital form requires equipment and software, which is changing constantly and may not be available within a decade of its introduction. Who today has a punched card reader, a Dectape drive, or a working copy of FORTRAN II? Even newer technology such as 9-track tape is rapidly becoming obsolete. We cannot save the machines if there are no spare parts available, and we cannot save the software if no one is left who knows how to use it.

Rapid changes in the means of recording information, in the formats for storage, and in the technologies for use threaten to render the life of information in the digital age as, to borrow a phrase from Hobbes, "nasty, brutish and short." Some information no doubt deserves such a fate, but numerous examples illustrate the danger of losing valuable cultural memories that may appear in digital form. Consider, for example, the now famous, but often misrepresented, case of the 1960 Census.

As it compiled the decennial census in the early sixties, the Census Bureau retained records for its own use in what it regarded as “permanent” storage. In 1976, the National Archives identified seven series of aggregated data from the 1960 Census files as having long-term historical value. A large portion of the selected records, however, resided on tapes that the Bureau could read only with a UNIVAC type II-A tape drive. By the mid-seventies, that particular tape drive was long obsolete, and the Census Bureau faced a significant engineering challenge in preserving the data from the UNIVAC type II-A tapes. By 1979, the Bureau had successfully copied onto industry-standard tapes nearly all the data judged then to have long-term value.

Though largely successful in the end, the data rescue effort was a signal event that helped move the Committee on the Records of Government six years later to proclaim that “the United States is in danger of losing its memory.” The Committee did not bother to describe the actual details of the migration of the 1960 census records. Nor did it analyze the effects on the integrity of the

constitutionally-mandated census of the nearly 10,000 (of approximately 1.5 million) records of aggregated data that the rescue effort did not successfully recover. Instead, it chose to register its warning on the dangers of machine obsolescence in apocryphal terms. With more than a little hyperbole, it wrote that "when the computer tapes containing the raw data from the 1960 federal census came to the attention of NARS [the National Archives and Records Service], there were only two machines in the world capable of reading those tapes: one in Japan and the other already deposited in the Smithsonian as a relic" (1985:9, 86-87).[1]

Other examples lack the memorable but false details accompanying the 1960 Census story, but they do equally illustrate how readily we can lose our heritage in electronic form when the custodian makes no plans for long-term retention in a changing technical environment. In 1964, the first electronic mail message was sent from either the Massachusetts Institute of Technology, the Carnegie Institute of Technology or Cambridge University. The message does not survive, however, and so there is no documentary record to determine which group sent the pathbreaking message. Satellite observations of Brazil in the 1970s, critical for establishing a time-line of changes in the Amazon basin, are also lost on the now obsolete tapes to which they were written (cf. National Research Council 1995a: 31; Eisenbeis 1995: 136, 173-74).

Similarly, in the late 1960s, the New York State Department of Commerce and Cornell University undertook the Land Use and Natural Resources Inventory Project (LUNR). The LUNR project produced a computerized map of New York State depicting patterns of land usage and identifying natural resources. It created a primitive geographic information system by superimposing a matrix over aerial photographs of the entire state and coding each cell according to its predominant features. The data were used for several comprehensive studies of land use patterns that informed urban planning, economic development, and environmental policy. In the mid-1980s, the New York State Archives obtained copies of the tapes containing the data from the LUNR inventory along with the original aerial photographs and several thousand mylar transparencies. Staff at the State Archives attempted to preserve the LUNR tapes, but the problems proved insurmountable. The LUNR project had depended on customized software programs to represent and analyze the data, and these programs were not saved with the data. Even if the software had been retained, the hardware and operating system needed to run the software were no longer available. As a consequence, researchers wishing to study changes in land use patterns now have to re-digitize the data from the hard-copy base maps and transparencies at the State Archives or rekey data from printouts at Cornell's Department of Manuscripts & University Archives.[2]

Today, information technologies that are increasingly powerful and easy to use, especially like those that support the World Wide Web, have unleashed the production and distribution of digital information. Such information is penetrating and transforming nearly every aspect of our culture. If we are effectively to preserve for future generations the portion of this rapidly expanding corpus of information in digital form that represents our cultural record, we need to understand the costs of doing so and we need to commit ourselves technically, legally, economically and organizationally to the full

dimensions of the task. Failure to look for trusted means and methods of digital preservation will certainly exact a stiff, long-term cultural penalty. The Task Force on Archiving of Digital Information here reports on its search for some of those means and methods.

The Challenge of Archiving Digital Information

The question of preserving or archiving digital information is not a new one and has been explored at a variety of levels over the last five decades. Archivists responsible for governmental and corporate records have been acutely aware of the difficulties entailed in trying to ensure that digital information survives for future generations. Far more than their library colleagues, who have continued to collect and organize published materials primarily in paper form, archivists have observed the materials for which they are responsible shift rapidly from paper objects produced on typewriters and other analog devices to include files created in word processor, spreadsheet and many other digital forms (see, e.g. Hedstrom 1991: 343-44; National Academy of Public Administration 1989).

Technological Obsolescence

Early attention to the difficulties in preserving digital information focused on the longevity of the physical media on which the information is stored. Even under the best storage conditions, however, digital media can be fragile and have limited shelf life. Moreover, new devices, processes and software are replacing the products and methods used to record, store, and retrieve digital information on breathtaking cycles of 2- to 5- years. Given such rates of technological change, even the most fragile media may well outlive the continued availability of readers for those media. Efforts to preserve physical media thus provide only a short-term, partial solution to the general problem of preserving digital information. Indeed, technological obsolescence represents a far greater threat to information in digital form than the inherent physical fragility of many digital media (Mallinson 1986; Gavrel 1986).

In the face of rapid technological obsolescence and to overcome the problem of media fragility, archivists have adopted the technique of "refreshing" digital information by copying it onto new media (see Bearman 1989:21-22; The University of the State of New York, et al. 1988; Lesk 1990: 5, 1992).[3] Copying from medium to medium, however, also suffers limitations as a means of digital preservation. Refreshing digital information by copying will work as an effective preservation technique only as long as the information is encoded in a format that is independent of the particular hardware and software needed to use it and as long as there exists software to manipulate the format in current use. Otherwise, copying depends either on the compatibility of present and past versions of software and generations of hardware or the ability of competing hardware and software product lines to interoperate. In respect of these factors -- backward compatibility and interoperability -- the rate of technological change exacts a serious toll on efforts to ensure the longevity of digital information.

Digital information today is produced in highly varying degrees of dependence on particular hardware and software. Moreover, it is costly and difficult for vendors to assure that their products are either "backwardly compatible" with previous versions or that they can interoperate with competing products. Refreshing thus cannot serve as a general solution for preserving digital

information and this conclusion has prompted discussion of other kinds of solutions. Jeff Rothenberg, for example, has recently suggested that there may be sufficient demand for entrepreneurs to create and archive emulators of software and operating systems that would allow the contents of digital information to be carried forward and used in its original format (Rothenberg 1995; see also Creque 1995).

Migration of Digital Information

Refreshing digital information by copying it from medium to medium and the possibility of maintaining a complex set of emulators describe two distinct points on a continuum of approaches to preserving digital information. However, neither refreshing nor emulation sufficiently describes the full range of options needed and available for digital preservation. Instead, a better and more general concept to describe these options is migration.

Migration is the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. Migration includes refreshing as a means of digital preservation but differs from it in the sense that it is not always possible to make an exact digital copy or replica of a database or other information object as hardware and software change and still maintain the compatibility of the object with the new generation of technology. Even for information that is encoded in a contemporary standard form (e.g., a bibliographic database in USMARC or a corporate financial database in SQL relational tables), forward migration of the information to a new standard or application program is, as anyone knows who has witnessed or participated in such a process, time-consuming, costly and much more complex than simple refreshing (see Michelson and Rothenberg 1992).

Legal and Organizational Issues

Compounding the technical challenges of migrating digital information is the problem of managing the process in a legal and organizational environment that is in flux as it moves to accommodate rapidly changing digital technologies. Consider, for example, the barriers to decisive preservation action caused by widespread uncertainty about legal and organizational requirements for managing the intellectual property that digital information represents. Addressing and resolving the legal and practical questions of migrating intellectual property in digital form necessarily involves a complex set of interested parties, including the creators and owners of intellectual property, managers of digital archives, representatives of the public interest, and actual and potential users of intellectual property. In addition, the parties who represent, for example, owners and users of different kinds of intellectual property (e.g., text and other document-like objects, photographs, film, software, multimedia objects) each operate under very different organizational regimes, with different experiences and expectations.

Adding to these complexities is the deep uncertainty that these groups face in confronting an increasingly digital world. Owners of intellectual property are unsure about how to price access to their information and worry, given the ease of copying, that the first digital copy of a work they sell may be the only copy they sell. Users of intellectual property are unsure about what rights are

conveyed with the use of a particular digital information object and about how much such use is worth. Government representatives seek in the public interest to manage the powerful changes that accompany digital technologies, but are unsure about what levels of influece are available to them and how to generate appropriate public policy. The state of the copyright law, which was generated and developed in an analog world but is applied in an increasingly digital universe, is itself confusing and uncertain (see, for example, Barlow 1994 and Garrett, et al. 1993). And all of these concerns are exacerbated by the fact that bits know no borders.

The costs and the technical, legal and organizational complexities of moving digital information forward into the future raise our greatest fear about the life of information in the digital future: namely, that owners or custodians who can no longer bear the expense and difficulty will deliberately or inadvertently, through a simple failure to act, destroy the objects without regard for future use. Repeated anecdotes about the loss of land use information, satellite imagery or census data, even when false or misleading, feed our general anxiety about the future of the cultural record we are accumulating in digital form. And uncertainty and lack of confidence about our will and ability to carry digital information forward into the future exert a major inhibiting force in our disposition to fully exploit the digital medium to generate, publish and disseminate information. But how well does the evidence really support our fears, anxieties and inhibitions regarding digital information?

The Need for Deep Infrastructure

Even after more than forty years of growth, the digital world of information technology and communication is still relatively young and immature in relation to the larger information universe, parts of which have been under development for centuries. Our experiences of expense, intricacy and error in digital preservation surely reflect, at least in part, our inexperience with this emerging world as we operate in its early stages. Viewed developmentally, the problem of preserving digital information for the future is not only, or even primarily, a problem of fine tuning a narrow set of technical variables. It is not a clearly defined problem like preserving the embrittled books that are self-destructing from the acid in the paper on which they were printed. Rather, it is a grander problem of organizing ourselves over time and as a society to maneuver effectively in a digital landscape. It is a problem of building -- almost from scratch -- the various systematic supports, or deep infrastructure, that will enable us to tame anxieties and move our cultural records naturally and confidently into the future.[4]

For digital preservation, the organizational effort -- the process of building deep infrastructure -- necessarily involves multiple, interrelated factors, many of which are either unknown or poorly defined. One of the biggest unknowns is the full impact on traditional information handling functions of distributed computing over electronic networks. The effort to meet the cultural imperative of digital preservation thus requires a complex iteration and reiteration of exploration, development and solution as the relevant factors and their interrelationships emerge and become clearer and more tractable. And the first task in the effort is not to posit answers, but to frame questions and issues in such a way as to engage the many parties already working in various ways with digital information so that they can help us understand the relevant issues and,

within the context of their work, help us identify, define and incorporate solutions that contribute to the larger, common goal of preserving our cultural heritage.

Conceptual Framework

The Commission on Preservation and Access and the Research Libraries Group (RLG) have joined together in charging the Task Force on Archiving of Digital Information to take this first essential step toward a national system of digital preservation. They have asked the Task Force to "consult broadly among librarians, archivists, curators, technologists, relevant government and private sector organizations, and other interested parties" in an effort to:

- -- "frame the key problems (organizational, technological, legal, economic etc.)" associated with digital preservation,
- -- "define the critical issues that inhibit resolution of each identified problem," and
- -- "recommend actions to remove the issue from the list."

The Task Force charge, however, itself frames the anticipated solutions in terms of the concept of "refreshing" (see Appendix 1).

In taking up this charge, the Task Force on Archiving of Digital Information starts from the premise, argued above, that migration is a broader and richer concept than "refreshing" for identifying the range of options for digital preservation. For purposes of this report, the Task Force casts migration, in the sense defined in the argument above, as an essential function of digital archives. Moreover, it envisions the development of a national system of digital archives.

The Task Force defines digital archives strictly in functional terms as repositories of digital information that are collectively responsible for ensuring, through the exercise of various migration strategies, the integrity and long-term accessibility of the nation's social, economic, cultural and intellectual heritage instantiated in digital form. Digital archives are distinct from digital libraries in the sense that digital libraries are repositories that collect and provide access to digital information, but may or may not provide for the long-term storage and access of that information. Digital libraries thus may or may not be, in functional terms, digital archives and, in fact, much of the recent work on digital libraries is notably silent on the archival issues of ensuring long-term storage and access (for some exceptions, see Ackerman and Fielding 1995; Conway 1994; Graham 1995a). Conversely, digital archives necessarily embrace digital library functions to the extent that they must select, obtain, store, and provide access to digital information. Many of the functional requirements for digital archives defined in this report thus overlap those for digital libraries.[5]

The Task Force has deliberately taken a functional approach in these critical definitions and in its general treatment of digital preservation so as not to prejudge the question of institutional structure. Many traditional libraries, archives and museums, as institutions, have taken and may well continue to assume digital library and archival functions. However, the Task Force recognizes that, as the digital environment emerges and its requirements become clearer, traditional institutions may need to change in various structurally significant ways and new kinds of institutions and institutional structures may

emerge to perform all or parts of key archival functions for digital information. For this report, the Task Force has thus ruled as out of scope answers to such structural questions as: What existing institutions should assume archival responsibilities for various kinds of digital information? What specific digital materials should an institution select for archival storage? What should the organizational hierarchy of digital archives look like?

The Task Force sees repositories of digital information as held together in a national archival system primarily through the operation of two essential mechanisms. First, to ensure that no valued digital information is lost to future generations, repositories claiming to serve an archival function must be able to prove that they are who they say they are by meeting or exceeding the standards and criteria of an independently-administered program for archival certification. Second, certified digital archives will have available to them a critical fail-safe mechanism. Such a mechanism, supported by organizational will, economic means and legal right, would enable a certified archival repository to exercise an aggressive rescue function to save digital information that it judges to be culturally significant and which is endangered in its current repository. The current repository may be a digital library, another digital archives, or some other individual, organizational, public or private source of digital information. Without the operation of a formal certification program and a fail-safe mechanism, preservation of the nation's cultural heritage in digital form will likely be overly dependent on marketplace forces, which may value information for too short a period and without applying broader, public interest criteria.

Plan of Work

In order to lay out the framework for digital preservation that it has envisioned, the Task Force begins with an analysis of the digital landscape, focusing on features, including stakeholder interest, that affect the integrity of digital information objects and which determine the ability of digital archives to preserve such objects over the long term. The Task Force then introduces the principles that responsibility for archiving rests initially with the creator or owner of the information and that digital archives may invoke a fail-safe mechanism to protect culturally valuable information. The report explores in detail the roles and responsibilities associated with the critical functions of managing:

-- the operating environment of digital archives;
-- strategies for migration of digital information; and
-- costs and financial matters.

The report concludes with a set of recommendations that provide a strong and urgent forward agenda for the Task Force sponsors, the Commission on Preservation and Access and the Research Libraries Group. The Task Force expects the Commission and RLG to pursue these recommendations on a national and, where appropriate, an international front, and to generate dialogue, interaction and products that will advance the development of trusted systems for the preservation of digital information. Such development is necessary for and will contribute powerfully to the overall growth of an information-based society and economy. We can afford to continue and increase economic and social investments in digital information objects and in the repositories for them on the information superhighway if, and only if, we also create the archival

means for the knowledge the objects and repositories contain to endure and redound to the benefit of future generations.

Information Objects in the Digital Landscape

Information objects in digital form, like those in other forms, move through life cycles. They are created, edited, described and indexed, disseminated, acquired, used, annotated, revised, re-created, modified and retained for future use or destroyed by a complex, interwoven community of creators and other owners, disseminators, value-added services, and institutional and individual users. The digital world is still too new for us to describe fully the life cycle of the information objects that do now or will in the future reside there, but what surely unites the community of actors in their various information-based activities is their common purpose in support of the pursuit of knowledge.

The pursuit of knowledge is a process in which the emergence of new knowledge builds on and reconstructs the old. Knowledge cannot advance without consistent and reliable access to information sources, past and present. It is the archival function in the system of knowledge creation and use that serves to identify and retain important sources of information and to ensure continuing access to them. How reliable the archival process proves to be in the emerging digital environment hinges on the trustworthy operation of digital archives and on their ability to maintain the integrity of the objects they are charged to preserve (Bearman 1995: 8-9, Duranti 1995, Hedstrom 1995, Lynch 1994a: 737-38; see also Pelikan 1992: 120, Waters 1996a,b).

For digital objects, no less than for objects of other kinds, knowing *how* operationally to preserve them depends, at least in part, on being able to discriminate the essential features of *what* needs to be preserved. In this section, the Task Force focuses on the integrity of information objects in the digital environment -- on the kinds of attributes they acquire during the course of their lives to give them a distinct identity -- and on the claims of various interested parties, or stakeholders, in the different kinds and attributes of digital information. In the next section, the Task Force explores the operational roles and responsibilities of digital archives.

The Integrity of Digital Information

Digital technologies increasingly serve to integrate information resources. Text, numeric data, images, voice, and video have heretofore resided in print or other analog media for storage and transmission. When they are encoded digitally, either by conversion or at the point of creation, these various kinds of resources share layers of technology -- a common means of storage and transmission -- that allows them to be brought together and used in both old and new ways.

Multimedia and hyperlinked objects on the World Wide Web represent some of the new kinds of information and new ways of knowing in the digital realm that bring together the traditional forms of information and transform their use. The application of computer hardware and software has also generated other new kinds of information objects, including the products of simulation, remote sensing, computer-aided design (CAD) and geographic information (GIS)

systems. These objects come into being and exist as creatures of the digital environment; if nurtured well, digital technologies will certainly beget still other kinds of information objects, which we can now only anticipate.

The processes of preserving digital information will vary significantly with the different kinds of objects -- textual, numeric, image, video, sound, multimedia, simulation and so on -- being preserved. Whatever preservation method is applied, however, the central goal must be to preserve information integrity; that is, to define and preserve those features of an information object that distinguish it as a whole and singular work. In the digital environment, the features that determine information integrity and deserve special attention for archival purposes include the following: content, fixity, reference, provenance, and context.

-- Content

Questions of preserving information integrity turn at their core on questions of content. What digital archives are trying to preserve after all is the intellectual substance contained in information objects. The notion of content, however, is itself a complex idea that operates at several different levels of abstraction. To preserve the integrity of objects in their charge, digital archives must decide at which level, or levels, of abstraction they are defining information content.

At the lowest level of abstraction, all digital information objects consist of simple bitstreams of 0s and 1s. One can distinguish objects from one another merely by distinguishing the configuration of bits. Preserving the integrity of an information object in this sense means preserving the bit configuration that uniquely defines the object. There are various well-established techniques, such as checksums and digests, for tracking the bit-level equivalence of digital objects and ensuring that a preserved object is identical to the original (Lynch 1994a: 739, 1996: 138; see also Graham 1994).

Defining content as a collection of bits, however, is often too limited and simplistic to be useful. In the digital environment, as we have seen, ideas are typically embedded in particular formats and structures that are dependent on hardware and software technologies subject to rapid change. Conceived at this higher level of abstraction -- that is, in terms of format and structure -- the definition of content poses considerable difficulties for managing information integrity in an archival context, and we are just beginning to come to terms with the expression of these problems in the digital environment.[6]

Consider textual objects. Text today is generally covered by a formal, international ASCII standard for representing character formats. Standard extensions exist for encoding diacritic characters in romance languages other than English and a new standard is slowly emerging to incorporate scripted languages under a new common encoding scheme (UNICODE). Alternative encoding methods, however, abound. IBM maintains its own EBCDIC character encoding scheme and Apple and Intel-based personal computers differ in the ways that they support extended ASCII character sets. For documents in which any of these character codes can adequately represent the contents, the differences in encoding schemes may matter little and digital archives can manage object integrity by mapping character sets from one to the other. However, for works involving multiple languages or complex equations and

formula, where character mapping is imperfect or not possible, character set format takes on considerably more significance over the long-term as a matter of content integrity.

In addition to character set issues, digital archives must also grapple with the means of representing and preserving textual content embedded in layout and structure. Markup systems, such as implementations of TeX and the Standard Generalized Markup Language (SGML), do exist as platform-independent mechanisms for identifying and tagging for subsequent layout and retrieval detailed structural elements of documents. The use of TeX and its variants, for example, is relatively common among scholars in some scientific disciplines such as mathematics and computer science, and the federal government and scholarly publishers are increasingly employing the SGML standard in documents that they produce and distribute electronically (see, for example, Cole and Kazmer 1995).[7] Beyond these relatively specialized segments, however, word processing and desktop publishing systems still dominate the market for the creation of documents with complex structure and layout, and the software for such use typically models and stores document structure and layout in proprietary terms. Although software may provide mechanisms for converting documents to common interchange formats, use of such mechanisms often results in the loss or inadequate rendering of content such as page structures and the layout of headers, footers and section headings.

The preservation of images comprises another example of the ways in which content, defined in terms of structure and format, poses integrity problems for digital archives. Image resolution, accuracy of color representation and compression for storage all require attention, but the interaction of these structural factors tends to pit judgments about the quality of content against the need for its efficient archival storage and use. In general, one can express the tradeoffs as follows: the higher the resolution and the richer the color register, the larger the file size and more costly the storage. The use of compression technologies further complicates the equation because the application of some algorithms to reduce space needs may involve an irreversible loss of data. The popular JPEG algorithm, for example, supports “lossy” compression, whereby the greatest degree of compression requires the maximum settings for allowable loss (Lynch 1994a: 743).

At the highest level of abstraction, digital archives define content in a way that transcends the limits of the hardware and software systems needed for reading and interpreting the bits of an information object and for rendering it for use in a specific format and structural representation; that is, they define content in terms of the knowledge or ideas the object contains. Expressed in this way, the preservation challenge for digital archives is to migrate (or to enable the user to migrate) intellectual content using standard interchange algorithms and other appropriate migration strategies so that the ideas available in the end are identical to those contained in the original object. The measure of integrity in the preservation process thus turns, at least in part, on informed and skillful judgments about the appropriate definition of the content of an digital information object -- about the extent to which content depends on its configuration of bits, on the structure and format of its representation, and on the ideas it contains -- and for what purposes.

-- Fixity

The process of identifying and preserving a digital information object as a whole and singular work goes well beyond considerations of content. It also depends, for instance, on the way that the content is fixed as a discrete object. If an object is not fixed, and the content is subject to change or withdrawal without notice, then its integrity may be compromised and its value as a cultural record would be severely diminished.

Outside the digital landscape, business practices, editorial policies, and legal constraints, have created disciplines for fixing information in discrete objects. For example, business records, such as an annual report, a letter of hire or an engineering drawing for the manufacture of a new product, are fixed in the transactions for which they constitute evidence. The acts of production and broadcast establish radio and television programs as discrete objects. And the act of publication marks a specific version or edition of a literary work. In each of these cases, it is virtually impossible to change or withdraw the cultural record that the release of an information object establishes.

On the digital landscape, by contrast, it is still relatively easy for a creator to alter or retract previously released information. Such actions can eliminate or overlay significant content and thereby corrupt the record. It is also relatively easy in an on-line environment -- and equally confusing for purposes of preserving information integrity -- for an author (or user) to make available concurrently multiple representations of what he or she considers to be the same work, providing no definition of a canonical version and viewing all representations as having equal quality and validity (Lynch 1994a: 743).[8]

To address these problems, a wide range of cryptographic techniques, such as watermarking, already exist for various kinds of digital information objects. These could serve well to mark and identify specific, canonical versions and editions of textual, audio and visual works, and to establish trusted, protected channels of distribution for those objects. However, the standards and infrastructure and the policies and practices for applying such techniques to the creation of fixed versions of digital information objects still need considerable development (Lynch 1994a: 740-741, 1993: 69-72). Absent such development and the fixity of information that would result, digital archives face considerable challenges in trying to preserve the integrity of digital objects (Lynch 1996: 138-142).[9]

Some kinds of digital objects present the problem of fixity of information in yet another way. An increasing number of networked information resources are better modeled, not in terms of versions or editions of works, but as continuously updated databases. The Human Genome Project supports such a database, which serves as a vehicle and record for a worldwide collaborative effort. The financial communities also depend on large, continuously updated databases. There is no natural way to fix these resources at the database level, no natural set of publication points for the objects as a whole. Their integrity, or singularity, as databases resides in the coherence of the database as a whole and in the continuousness of the updates.

To preserve the integrity of such information resources, the complete record of changes has to be built into the design of the database and fixed at the record level. Technology exists for constructing databases that maintain all previous versions of the records in the database as well as details about when new information is added or logically deleted. A database designed to a so-called time-travel specification, however, is expensive to construct and operate. In the absence of such a preservation-oriented design, digital archives may have no choice but to fix the database artificially in time and capture a series of snapshots of its state at periodic intervals as a way of preserving its integrity (Lynch 1994a: 741-742, 1993: 37).

-- Reference

A third aspect of information integrity that bears crucially on the preservation process is that information objects must have a consistent means of reference. Information objects come into being and acquire their distinctiveness in relation to various other objects in an information space. For an object to maintain its integrity, its wholeness and singularity, one must be able to locate it definitively and reliably over time among other objects (Dollar 1992:62-65; Michelson and Rothenberg 1992).

Systems of citation, description and classification provide the necessary means of reference so that one can consistently discover, identify and retrieve relevant objects.[10] Such systems traditionally include bibliographies, catalogs, indices, data dictionaries, directory systems, finding aids and the like, all of which collect, assemble and organize references to bodies of work. Creators of these resources are already extending them with such innovations, for example, as the use of the 856 field in bibliographic records, to incorporate reference to digital objects. Moreover, in the electronic environment, there are also emerging sophisticated tools for automatically gathering and presenting in new kinds of products referential information from textual, image, sound, video and other kinds of digital objects.

To meet the test of consistent reference, however, citations to digital objects from these various manually and automatically generated sources must over time reliably identify the works to which they refer. One way of establishing confidence in a reference system is to generate citations in the first place -- either manually or automatically -- from the self-referential information that a work itself carries with it, perhaps in a header, a digital envelope or some other distinct part of the digital object. Unfortunately, except for documents marked up to the standard of the Text Encoding Initiative, few digital objects today contain self-referential information that meets conventional citation quality. Matching citation and digital work to meet the test of consistent reference thus is difficult. Moreover, reliably identifying multiple versions and editions of works in electronic form remains a persistent reference problem (Lynch 1993: 73-74, 1996: 138-142; see also Davis 1995, European Commission 1996, Levy and Marshall 1995).

Part of the general problem of identifying and consistently referring to variant digital works is the specific problem of resolving names and locations for them. Coordinated by such groups as the Internet Engineering Task Force (IETF), active research and development has led to an emerging consensus that several factors contribute to the unique identification of digital information objects.

Two of the most important factors are the Uniform Resource Name (URN) and the Uniform Resource Locator (URL).[11]

The URL refers to the specific place where a digital object resides and is currently the dominant method of object location on the World Wide Web. The weakness of the URL, however, is that it may frequently change, especially as an object migrates from one machine to the next. By contrast, a name, the URN, is supposed to apply uniquely and permanently to a distinct object and to designate it independently of its particular location at any point in time. URNs today exist more in concept than in practice. Thus, resolving the name and location for variant digital works and thereby providing consistent reference to them means moving from a conceptual design of the relation between names and locations to an operational reality through the implementation of naming authorities, which assign URNs, and and the development of digital services that translate names into currently valid URLs (Lynch 1994b).

In order to provide a consistent means of reference for digital objects, systems of citation, description and classification will need to dispense more than name and location information. The IETF framework provides a general category for additional reference information in what it calls Uniform Resource Characteristics (URCs). Among the kinds of information that a reference system might provide under this general category is information about the intellectual property rights governing the use of a particular object and its cost. For archival purposes, however, there are two other qualities of digital information objects that a reference system must take into account. These are the features of provenance and context.

-- Provenance

Provenance has become one of the central organizing concepts of modern archival science (Dollar 1992: 48-51; see also Bearman 1989, 1995; Hedstrom 1995). The assumption underlying the principle of provenance is that the integrity of an information object is partly embodied in tracing from where it came. To preserve the integrity of an information object, digital archives must preserve a record of its origin and chain of custody.

For some information objects, the formal process of publication creates a trusted channel of distribution and serves to establish a sophisticated record of provenance, at least from the creator through the point of release. In the digital environment, the problems of documenting provenance through a record of publication are bound up with the problems of fixity, and particularly with issues involved in tracing multiple versions and editions, as described above (see Lynch 1996: 143). Outside the path of publication, however, there are several other channels of information creation and distribution, including tracing the path of migration within their own organization, that digital archives need to document in order to preserve the integrity of digital objects in their custody.

First, there is a chain of provenance from individuals, a trail which has comprised a traditional concern of archivists. In the digital environment, as in other domains, individuals produce and accrete much information that relates to their private lives and to their public roles and responsibilities, including their means of livelihood. Among the kinds of information objects that may count as personal records in digital form are electronic mail, notes, manuscripts, journals,

photographs or videos, and databases of personal finances. For objects of these kinds, which make their way into digital archives and which the custodian deems worth keeping as part of the historical record, the archival responsibility in establishing provenance is to establish, as unambiguously as possible, the identity of the individuals who are the source of the objects and to trace the chain of custody to arrival in the archives.

A second kind of digital information flow that requires special attention with respect to provenance are those mediated or governed by corporate information systems. The digital products of such systems may include databases of employment and financial records, contracts and other legal records, planning documents, FDA filings, technical reports and so on. The distinctive feature of such objects is that they are all best viewed not as the result of individual works but as proceeding from the business, governmental or other social practices that the information system organizes and instantiates. The archival challenge for establishing the provenance of such objects is to find ways to preserve an understanding of the corporate policies and processes and roles and responsibilities thus represented in the information system and its products (see Hedstrom 1991:349; Dollar 1992:62)

Third, there is the chain of provenance that traces information, particularly scientific data, to a source in electronic instrumentation. In some cases, the instrumentation produces data in the service of individual experiments or of clinical practice; in other cases, remote sensors gather streams of observational data about physical systems in space or on earth (see, for example, National Research Council 1995a: 2-29). Of course, one ultimately accounts for the origin of the resulting data in the design and conduct of the experiment or clinical service, or in the character of the remote sensing program. However, the use of the instrumentation creates a special requirement to trace the provenance of the data to the design characteristics of the instrument itself. The data are a direct product of and derive their meaning and usefulness from the calibration, units of measure, sampling rate, conditions of recording and other relevant features embedded in the technology of the instrument.

Finally, a concern for provenance in preserving the integrity of digital information means that digital archives must document what happens to the information within their own organizations. They must keep a record of migration activity, and particularly of the transformations they make to keep information objects current with new technologies for use. Only by tracing such migratory activity, can a digital archives establish its own chain of custody back to the original object.

In the end, the investment that digital archives make in establishing the provenance of objects in their care serves to preserve the integrity of digital information in two distinct ways. First, a tracing of chain of custody from the point of creation helps to create the presumption that an object is authentic, that it is what it purports to be and that its content, however defined, has not been manipulated, altered or falsified (Duranti 1995: 7-8). The second effect of establishing provenance through a chain of custody is to document, at least in part, the particular uses of the object by the custodians. In thus creating a record

of use, the archival concern with provenance is intimately related to the notion of context as a matter of information integrity.

-- Context

The fifth attribute of information integrity that bears on the preservation of digital information objects is their context, the ways in which they interact with elements in the wider digital environment (Dollar 1992: 48; Bearman 1995: 4; Hedstrom 1995). The context of digital information includes a technical dimension, a dimension of linkage to other objects, a communication dimension, and a wider social dimension. Digital archives must be attentive to the ways in which each of these contextual dimensions affects the integrity of the objects in their care.

First, to specify the technical context of digital information is to specify its hardware and software dependencies (Bearman and Sochats 1995). Digital objects, by nature, require the use of computer hardware and software to create and use them. In some cases, an object may be closely dependent on a configuration of technology. For example, a particular digital document may require a particular word processing program running on a machine with a special kind of computer processing chip and operating system. Other objects, such as an image file, may be readable using a variety of software programs, but may be stored on disk in a format that is readable only with a special computer. For still other objects, like World Wide Web documents marked-up in HTML, the hardware and software dependencies may be specified in very general terms because they are readable with a software browser that is available on almost any hardware platform. The archival challenges for preserving the integrity of these various kinds of objects are, on the one hand, to represent faithfully the context of the objects in terms of their hardware and software dependencies and, on the other hand, to overcome, through appropriate migration efforts, those dependencies that threaten to hinder future use.

A second dimension of the context of digital information objects is the linkages that may exist among them. On the World Wide Web, for example, the HyperText Markup Language (HTML) provides a way to place in one object references to various other objects. A click of a mouse key then enables one to move quickly from one linked object to another. If the integrity of these objects is seen as residing in the network of linkages among them, rather than in the individual objects, or nodes, on the network, then the archival challenge would be to preserve both the objects and the linkages, a task that would today be exceedingly complex. At present, there appears to be no good archiving solution; a possible stop-gap measure would be to treat the network in terms of its component parts and to take periodic snapshots of the individual WWW objects.

The communications medium comprises a third dimension of the context of digital information. The character and integrity of digital information objects depends, in part, on their mode of distribution. Material distributed on CD-ROM, for example, has a set of file format and other characteristics that set it apart from material distributed in, say, networked form. As digital materials increasingly come into being and subsist in an electronically networked environment, contextual features of the network, such as bandwidth and security, will account for characteristics of the digital objects. Increasing

bandwidth, for example, will stimulate the production and dissemination of high-bandwidth digital materials such as full-motion video. Similarly, improved security on the network will encourage the conduct of a wide variety of confidential transactions. As network-based objects, such as video and records of confidential transactions, make their way into digital repositories, an archival account of their integrity must include an account of the features of the network context that supports their existence.

Finally, the wider social environment plays a significant contextual role that contributes to the integrity of digital information objects. Networked information, for example, depends on specific policy and implementation decisions that address the bandwidth, security and other qualities of the network and related technical infrastructure. Electronic mail is subject to a variety of social distinctions when, in some contexts, it is used to convey highly personal messages among friends and, in other contexts, is a vehicle for formal communication among academic or business colleagues (Hedstrom 1995). One might account for digital business records, too, in the social context of the political and organizational regime in which they were generated. Indeed, the social context of digital information objects is an expression of what one might characterize in other terms as the interests of stakeholders in those objects.

Stakeholder Interests

Digital information objects acquire the qualities of content, fixity, reference, provenance and context -- and thereby their integrity -- as they move through a life cycle in a series of relationships with parties, or stakeholders, who have specific interests in their creation, management, dissemination, use or retention. The initial stakeholder in a digital information object is, of course, the creator of its content. Following creation, a digital information object may pass through a series of gateways of increasingly public release and access. Some digital information objects (like videos of family picnics or private journals) may never be released beyond the initial creator; others will be limited to an immediate circle, which may or may not be physically co-located with the creator. Still other objects will be very widely disseminated.

Once a creator fixes a digital information object by releasing it, for either personal or wider use, the object invites various other kinds of stakeholder interests. Depending on the extent of the release and the nature of the object, users and other stakeholders may refer to it, index it or cite it. Still others may acquire, license or otherwise take custody of it to edit and publish it, to invoke it as evidence of a transaction or perhaps, as in the case of scientific data, to study, analyze and report it. They will all use it in the context of specific hardware and software applications and perhaps link it to related objects. And eventually during the course of its life, a digital information object may attract the interest of stakeholders, such as a digital libraries, which serve as collectors/disseminators of information objects.

What unites all these various stakeholders, of course, is their particular interest in adding to or making use of the value of digital information objects. However, the integrity of information objects in the digital environment is today so fragile that as stakeholders disseminate, use, reuse, recreate and re-disseminate various kinds of digital information, they can easily, even inadvertently, destroy valuable information, corrupt the cultural record, and ultimately thwart the

pursuit of knowledge that is their common end. Against such a danger, a safety-net is needed to ensure that digital information objects with long-term cultural and intellectual value survive the expressions of stakeholder interest with their integrity intact.

Among the stakeholders serving as collector/disseminators, those who perform archival functions provide just such a safety-net. Digital archives build and maintain reliable collections of well-defined digital information objects and they preserve the features -- content, fixity, reference, provenance and context -- that give those objects their integrity and enduring value. In the next section of this report, we discuss the organizational principles and requirements for creating a safety net of distributed digital archives. Such an archival network will establish a critical part of the foundation needed to support a burgeoning, but still fragile, national and international knowledge economy.

Archival Roles and Responsibilities

As the networked environment for digital information expands and matures, intense interactions among the parties with stakes in digital information are providing the opportunity and stimulus for new stakeholders to emerge and add value, and for the relationships and division of labor among existing stakeholders to assume new forms. For example, naming authorities for digital information objects are being organized, creators are assuming the role of publishers, publishers are contemplating the construction of digital archives, and collector/disseminators are facing the need to resolve the names and locations of digital objects. Moreover, all of the stakeholders, including the new ones who are emerging, are exercising their various functions simultaneously in relation to particular digital information objects. That is, for example, the creator of an object may retain a financial or other stake in its dissemination and use for an extended period, just as a publisher/disseminator of the object exercises the same or similar interest. Similarly, stakeholders acting as collectors/disseminators may retain an interest in the continuing use of the object, at the same time as they or other stakeholders acting as digital archives take steps to preserve it for future use (Wiederhold 1995).

In a time of such sustained flux and change, during which these various divisions of labor are taking shape, the most effective and affordable strategy for developing a system of digital archives is to assume a distributed, rather than centralized, structure for collecting digital information objects, protecting their integrity over the long term, and retaining them for future use. A distributed structure, built on a foundation of electronic networks, places archival responsibility with those who presumably care most about and have the greatest understanding of the value of particular digital information objects. Moreover, such a structure locates the economic and cultural incentives where they are most likely to prompt those preserving digital information to respond with the greatest agility to the changing digital landscape and to the shifting tides of technology.[12]

Effective structures for digital archives in a distributed network will surely take various forms and will include corporations, federations and consortia, each of which may specialize in the archiving of digital information and range over regional and national boundaries. Both informal collaborations (associations and alliances) and formal partnerships among contractors and subcontractors, will also surely arise, in which responsibilities for archiving are allocated among various other interests in digital information. Moreover, shared interests in, for example, intellectual discipline, in type of information, in function, such as storage or cataloging, and even interests in the output of information within national boundaries will all form a varied and rich basis for the kinds of formal and informal interactions that lead to the design of particular archival organizations. In the end, existing archival organizations may successfully adapt their current roles and responsibilities to changing needs, existing archival roles and responsibilities may be redistributed among other kinds of organizations with stakes in digital information, and focused attention on the

General Principles

division of labor in the digital information economy will lead to the emergence and growth of new kinds of archival structures.

Whatever the particular structural outcomes may be as stakeholders, new and old, interact and give shape to the distributed network of digital information, distributed responsibility for preserving that information requires commitment at least to the following set of organizing principles:

1. Information creators/providers/owners have initial responsibility for archiving their digital information objects and thereby ensuring the long-term preservation of those objects.

    -- The creator/provider/owner may engage other parties, such as certified digital archives, to take over some or all of the archival responsibility.

    -- Libraries and archival organizations may interact with creators/providers as subcontractors for maintaining digital archives during and after the active life of their information objects.

2. Certified digital archives have the right and duty to exercise an aggressive rescue function as a fail-safe mechanism to preserve information objects that become endangered because the creator/provider/owner does not accept responsibility for the preservation function and does not take steps formally to convey responsibility, or because there is no natural institutional home for the objects.

The conditions of creating digital information and giving it a useful life are essentially the same as those required for the information to persist over time. That is, it must be stored and maintained in an accessible form. It is not unreasonable, therefore, to assert, as the first principle does, that initial responsibility for preservation begins with creation of the information and rests with the creator, owner or provider of it. Individuals and public and private agencies already regard such a responsibility as a natural one for their critical internal records. Creators of knowledge in other spheres depend on past knowledge and regularly acknowledge their responsibility to add to the enduring record when they publish their own ideas and findings. As the properties and usefulness of other kinds of information objects become more widely known, the ability in the digital environment to reuse and repackage these objects may generate revenue or other benefits. Creators, owners or providers of these objects, including some publishers, who may not have otherwise counted long-term preservation among their key responsibilities, may thus be more inclined to do so and may either assume the responsibility themselves or find a qualified partner to do so. Among potential partners, they will likely find libraries, archives or similar agencies, which have specific collection agendas and will seek to take or share responsibility for preserving organized collections of digital materials.[13]

The second organizing principle calls, in effect, for a fail-safe mechanism. A variety of factors -- budgetary constraints, reorganization of priorities or focus, change of business, the need to go out of existence, or expiration of copyright -- might prompt custodians to neglect, abandon or destroy their collections of

digital information. No distributed system of digital archives will afford effective protection of electronic information unless it provides for a powerful rescue function allowing one agency, acting in the long-term public interest of protecting the cultural record, to override another's neglect of or active interest in abandoning or destroying parts of that record.[14]

Digital archives operating normally and those operating in fail-safe mode differ mainly in the rights and obligations they have with respect to rescuing materials that have fallen in jeopardy because a custodian no longer accepts preservation responsibility. Section 108 of the 1976 Copyright Act of the United States defines the rights and obligations for libraries and archives operating normally to preserve copies of printed materials. The proposed revision of the copyright law reserves to libraries and archives similar rights and duties for digital materials. However, neither in its present form nor in its proposed revision does the Copyright Act provide a legal foundation defining the rights and duties that must come into play to facilitate and encourage an aggressive rescue function.

The Copyright Act, in fact, may not provide the best legal and institutional framework for establishing fail-safe mechanisms for digital archives. However, such a framework is urgently needed because we know that one of the greatest dangers to the long life of digital information is the ease with which it can be abandoned and then deliberately or inadvertently destroyed. One appropriate foundation for an archival fail-safe mechanism against this danger might rest in articulating for the various domains of intellectual property -- on analogy from the domain of real estate -- a definition of the concept of abandonment and of the rights and duties that follow for digital archives operating in fail-safe mode when they can show that present custodians have abandoned, or are in the process of abandoning, culturally valuable objects. Other means of providing a fail-safe mechanism for digital archives also exist, however. For example, a depository system might well serve the purpose. Under such a system, publishers could be legally bound to place with a certified digital archives a copy of their published digital works in a standard archival format (cf. European Commission 1996 and Consortium of University Research Libraries 1996).[15]

Given the principles articulated here, it follows that a commitment to preservation -- collecting digital information objects, protecting their integrity over the long term, and retaining them in an accessible form for future use -- is a defining feature of a digital archives, whether it is operating in normal or fail-safe mode. This commitment is fulfilled in practice in any digital archives by the exercise of three crucial functions: managing the operating environment of the archives, the migration of the archives as the operating environment changes, and the costs and finances of the operating environment and of periodic migrations.

The operating environment of digital archives

For assuring the longevity of information, perhaps the most important role in the operation of a digital archives is managing the identity, integrity and quality of the archives itself as a trusted source of the cultural record. Users of archived information in electronic form and of archival services relating to that information need to have assurance that a digital archives is what it says that it is and that the information stored there is safe for the long term. In the view of the Task Force, a formal process of certification, in which digital archives meet or

exceed the standards and criteria of an independent certifying agency, would serve to establish an overall climate of value and trust about the prospects of preserving digital information.[16] Appropriate organizations and individuals need to begin now developing such standards and criteria, including standard methods for a repository to declare its existence as a digital archives and therefore its intentions to preserve the contents over which it has custody, and for describing what the archives contains and what services it provides. The certification process must also address the standards and best practices for other dimensions of archival operation, including selection and accessioning of material, its storage and access and the engineering of the systems environment.

-- Appraisal and Selection

Archives cannot save all information objects; they must appraise and select for retention the most valuable items. Selection processes for archives of all kinds -- paper and digital -- are matters of intellectual judgment about what to include and save and what to exclude. Criteria for such judgments are largely tied to the intrinsic qualities of the material and many of the criteria that have proven useful in the paper world will no doubt translate to and prove equally effective in the digital environment. In general, selection criteria include an appraisal of the content of the object -- its subject and discipline -- in relation to the collection goals of the digital archives, the quality and uniqueness of the object, its accessibility in terms of available hardware and software, its present value and its likely future value (see National Research Council 1995a: 33-41, Owen and van de Walle 1995: 13; see also Atkinson 1986).

Selection is also dependent on a number of extrinsic factors. For example, it is acutely dependent on search and retrieval mechanisms to help place and evaluate candidate objects in a larger universe of related materials. Search and retrieval in large universes of diverse digital materials is the subject of very active computer science research, but search and retrieval today against a rapidly expanding universe of materials in all forms -- digital, paper, microform -- remains difficult, and the relatively primitive systems currently in operation may, in the short term at least, inhibit the effective selection of materials for digital archives.

Redundancy of information is also important for effective selection. Selectors need to know before accepting an object where a copy of record is stored and if and where additional copies are distributed either as backup or for more efficient retrieval. In addition, selectors ultimately need to have a rich understanding of the software and hardware dependencies of candidate digital information objects so that they can factor the carrying costs for an object into their overall assessment of its value. Such understandings remain in relatively short supply in some measure because the educational processes for wedding selection and technical skills still need to be devised and perfected. In larger measure, however, technical understandings are deficient because the digital environment is so immature that the dependencies of information objects on underlying hardware and software are still in many respects unknown.

Finally, selection for digital archives must be a continuing process. Given the need to migrate digital information regularly from its hardware and software environment, the stimulus and occasion will recur to reappraise the value of the material being migrated. Materials may lose value over time and may need to

be withdrawn and discarded (deaccessioned). Hard decisions may also be necessary about how and whether digital objects migrate if changes in a technical environment force alters what is preserved so that it differs in fundamental ways from what existed in the original object. Digital archives thus need to formulate and implement policies and practices for ongoing appraisal (Conway 1996a).

-- Accession

Once an information object is identified for inclusion in a digital archives, it needs to be accessioned, that is, prepared for the archives. The accession process involves both describing and cataloging selected objects, including their provenance and context, and securing them for storage and access. Standards for description are well developed for certain kinds of materials that are likely to appear in digital archives, such as monographs and serials. Because of the high degree of standardization, one might reasonably expect such descriptions eventually to accompany the digital object, thus simplifying the accessioning process. For other kinds of materials, such as WWW pages, motion video and multimedia, relevant attributes are less well understood. Standards of description are less well practiced and one cannot reasonably expect satisfactory descriptive material to accompany the digital object. In all cases, special attention is needed in the accession process to creating, describing and tracking the versions of the object in the digital archives to satisfy the various requirements for display and other forms of access and for long-term storage. If material is to be deaccessioned then public declarations need to be made to that effect, particularly if it is the last known copy, so that rescue efforts by others can proceed if appropriate.

In the accession process, selected digital objects also need to be made secure for indefinite future use. A digital archives may need to establish access controls for its information objects and the means for authenticating them to future users. Establishing access controls involves setting terms and conditions for authorized use by specific users or classes of users. For certain extreme cases, such as rescued objects still under intellectual property protection, there may be no authorized uses for certain periods of time except for internal use within the digital archives to ensure that the object will be accessible in the future. Authentication, which may make use of cryptography, provides verification that a digital object is what it purports to be and contains the contents that the author/creator or publisher originally intended. One of the characteristics of materials in digital archives may need to be that they contain a digital digest or signature that users can independently verify to assure themselves that the object is unchanged from its archival state.[17]

-- Storage

The storage operation in digital archives attends primarily to the media level formatting of information objects. Primary considerations include levels of hierarchy and redundancy. In any digital archives, there may be multiple levels of storage graded to levels of expected use and needed performance in retrieval. Little used material may be stored most efficiently off-line, usually in tape format. For objects in high demand, where retrieval time is at a premium, on-line storage in magnetic media may serve best. In a distributed network, they may need to be stored on-line in multiple locations. An intermediate solution is near-line storage, where information objects may be stored on optical or tape media and loaded in a jukebox. Retrieval time in near-line storage systems

suffers by comparison to on-line storage, but is considerably more responsive to user demand than off-line storage. Digital archives may use any or all of these methods. The most sophisticated systems combine the resources so that objects in use or recent use are stored on-line and, as they age from the time of most recent use, they move to near-line storage and then eventually off-line.

Another important storage consideration is redundancy. In a system that is completely dependent on the interaction of various kinds and levels of hardware and software, failure in any one of the subsystems could mean the loss or corruption of the information object. Effective storage management thus means providing for redundant copies of the archived objects as an insurance against loss. Depending on the copyright status of the objects, digital archives may choose to make backup copies on their own or to make arrangements for other sites, which hold the same object, in the network of digital archives to serve as backup, or they may choose to do both.

-- Access

Providing access to digital information in a distributed network environment means above all that digital archives are connected to networks using appropriate protocols and with bandwidth suitable for delivering the information their control. Digital archives have an obligation to maintain the information in a form so that users over the network can find it with appropriate retrieval engines and view, print, listen to or otherwise use it with appropriate output devices. In the descriptions of the resources they hold, responsible digital archives must provide to their users what they know about the provenance and context of their digital objects so that users can make informed decisions about the reliability and quality of the evidence before them. With respect to access, digital archives also have the responsibility to manage intellectual property rights by facilitating transactions between rights-holders in the information and users and by taking every reasonable precaution to prevent unauthorized use of the material.

-- Systems engineering

Because so many of the operational responsibilities of the digital archives -- selection, accessioning, storage and access -- are functionally identical to those of more traditional permanent repositories, they may successfully extend their scope to include digital materials. Many traditional archives have already embraced digital materials, and libraries and museums are not far behind. Wherever digital archives may reside organizationally, their operation is highly distinctive in one crucial respect. That is, they need, at least now and for the foreseeable future, a high level of systems engineering skill to manage the interlocking requirements of media, data formats, and hardware and software on which the operation of the digital archives essentially depends.

As the digital environment matures, the role of systems engineer will serve to integrate new technical developments that promise to streamline and strengthen the operation of digital archives. Commercially available systems for user authorization, and document authentication, systems for using resource descriptions to automate the maintenance and delivery of archived information objects, and networked-based services that help manage the conventions for naming digital resources and provide means for conducting intellectual property transactions all will need the attention of systems engineers to ensure that they are effectively incorporated into the normal operation of digital archives. In

addition, the systems engineering function will serve an essential role in helping to determine when objects in digital archives should migrate to new hardware and software.

Migration Strategies

As the operating environments of digital archives change, it becomes necessary to migrate their contents. There are a variety of migration strategies for transferring digital information from obsolete systems to current hardware and software systems so that the information remains accessible and usable. No single strategy applies to all formats of digital information and none of the current preservation methods is entirely satisfactory. Migration strategies and their associated costs vary in different application environments, for different formats of digital materials, and for preserving different degrees of computation, display, and retrieval capabilities.

Methods for migrating digital information in relatively simple files of data are quite well established, but the preservation community is only beginning to address migration of more complex digital objects. Additional research on migration is needed to test the technical feasibility of various approaches to migration, determine the costs associated with these approaches, and establish benchmarks and best practices. Although migration should become more effective as the digital preservation community gains practical experience and learns how to select appropriate and effective methods, migration remains largely experimental and provides fertile ground for research and development efforts.

Stewards of digital material have a range of options when faced with the need to preserve digital information. One might preserve an exact replica of a digital object with complete display, retrieval, and computational functionality, or a representation of it with only partial computation capabilities, or a surrogate such as an abstract, summary, or aggregation. Detail or background noise might be dropped out intentionally through successive generations of migration, and custodians might change the form, format or media of the information. Enhancements are technologically possible through clean-up, mark-up, and linkage, or by adding indexing and other features. These technological possibilities in turn impose serious new responsibilities for presenting digital materials to users in a way that allows them to determine the authenticity of the information and its relationship to the original object.

-- Change Media

One migration strategy is to transfer digital materials from less stable to more stable media. The most prevalent version of this strategy involves printing digital information on paper or recording it on microfilm. Paper and microfilm are more stable than most digital media, and no special hardware or software are needed to retrieve information from them. Retaining the information in digital form by copying it onto new digital storage media may be appropriate when the information exists in a "software-independent" format as ASCII text files or as flat files with simple, uniform structures. Several data archives hold large collections of numerical data that were captured on punch cards in the 1950s or 1960s, migrated to two or three different magnetic tape formats, and now reside on optical media. As new media and storage formats were introduced, the data were migrated without any significant change in their logical structure.

Copying from one medium to another has the distinct advantage of being universally available and easy to implement. It is a cost-effective strategy for preserving digital information in those cases where retaining the content is paramount, but display, indexing, and computational characteristics are not critical. As long as the preservation community lacks more robust and cost-effective migration strategies, printing to paper or film and preserving flat files will remain the preferred method of storage for many institutions and for certain formats of digital information.

Yet the simplicity and universality of copying as a migration strategy may come at the expense of great losses in the form or structure of digital information. When the access method for some non-standard data changes, one must, in order to migrate them, often eliminate, or "flatten," the structure of documents, the data relationships embedded in databases, and the means of authentication which are managed and interpreted through software. Computation capabilities, graphic display, indexing, and other features may also be lost, leaving behind the skeletal remnant of the original object. This strategy is not feasible, however, for preserving complex data objects from complex systems. It is not possible, for example to microfilm the equations embedded in a spreadsheet, to print out an interactive full motion video, or to preserve a multimedia document as a flat file.

-- Change Format

Another migration strategy for digital archives with large, complex, and diverse collections of digital materials is to migrate digital objects from the great multiplicity of formats used to create digital materials to a smaller, more manageable number of standard formats that can still encode the complexity of structure and form of the original. A digital archives might accept textual documents in several commonly available commercial word processing formats or require that documents conform to standards like SGML (ISO 8879). Databases might be stored in one of several common relational database management systems, while images would conform to a tagged image file format and standard compression algorithms (e.g., JFIF/JPEG).

Changing format as a migration strategy has the advantage of preserving more of the display, dissemination, and computational characteristics of the original object, while reducing the large variety of customized transformations that would otherwise be necessary to migrate material to future generations of technology. This strategy rests on the assumption that software products, which are either compliant with widely adopted standards or are widely dispersed in the marketplace, are less volatile than the software market as a whole. Also, most common commercial products provide utilities for upward migration and for swapping documents, databases, and more complex objects between software systems. Nevertheless, software and standards continue to evolve so this strategy simplifies but does not eliminate the need for periodic migration or the need for analysis of the potential effects of such migration on the integrity of the digital object.

-- Incorporate Standards

Digital archives will benefit from the widespread adoption of data and communication standards that facilitate reference to digital information objects and enable their interchange among systems. Business needs in many

institutions are driving the development and adoption of data standards. Organizations that create, use and maintain Geographic Information Systems (GIS), for example, are trying to reduce data conversion and maintenance costs by creating data that conform to widely accepted standards so that they can be exchanged, reused, or sold. Rapid implementation of electronic commerce depends on widespread development and adoption of standards for EDI (electronic data interchange) transaction sets under auspices of the ANSI X.12 committee. Standards initiatives that address business needs for the secure and reliable exchange of digital information among the current generation of systems will impose standardization and normalization of data that ultimately will facilitate migrations to new generations of technology. Digital archives must keep abreast of standards developments and make sure that their own technological infrastructure conforms to widely adopted standards.

-- Build Migration Paths

Planning for long-term preservation is a critical element of digital preservation. To the extent that creators/providers/owners of digital information accept initial responsibility for archiving their objects, they may begin to see the wisdom of incorporating migration paths or other provisions for preservation as an integral part of the process or system that generates digital information. To assist in this educational process, digital archives might work one-on-one with potential donors to develop agreements early in their careers and establish arrangements for regular, on-line deposit of digital materials in a format acceptable to the repository. In government, institutional, and corporate settings, archivists and librarians can issue guidelines and advice for digital preservation and encourage their parent institutions to adopt common usage rules, comply with data standards, and select applications software that supports migration. The preservation community as a whole needs to work with industry to create information systems and standards that build in archival considerations, such as backward compatibility from the point of initial design. Backward compatibility or migration paths would enable new generations of software to "read" data from older systems without substantial reformatting. Although backward compatibility is increasingly common within software product lines, migration paths are not commonly provided between competing software products or for products that fail in the marketplace.[18]

-- Use Processing Centers

Although standards and migration paths may become commonplace at some future date, a large body of digital material exists today in non-standard formats, and organizations and individuals continue to produce digital materials in formats that will require migration. Developing "processing centers" that specialize in migration and reformatting of obsolete materials may provide a cost-effective method of digital preservation. Processing centers might provide reformatting services for particular types of materials, such as text, certain database structures, geographic information systems (GIS), or multimedia products. Such centers might maintain older versions of hardware and software to support migration. They might provide a platform for reading and viewing digital information with the same "look and feel" as the original version by developing "software emulators" as suggested by Rothenberg (1995). Processing centers would take advantage of economies of scale and maximize the use of uncommon technical expertise. Migration/preservation services centers might resemble commercial firms that reformat old home movies and obsolete video formats or consortia of libraries and archives with distributed

preservation programs. A national laboratory for digital preservation, modeled after the National Media Laboratory, is another alternative. Feasibility studies and cost/benefit analyses would be necessary to determine the technological, economic, and commercial viability of such processing centers.

Managing Costs and Finances

In addition to managing their operating environment and the migration of information through hardware and software platforms, a third function by which digital archives fulfill their commitment to preserve electronic information is in managing the costs of these activities. The principal cost factors of the operating environment are those associated with selection, accession, storage, use, migration, property rights transactions, and the systems engineering needed to manage migration and to maintain the digital archives within a distributed network infrastructure. Some costs, like those for hardware and software as well as those for intellectual property if rights are purchased rather than leased, will appear as capital costs and will need to be amortized.

Operating costs will vary by the form of the information, by usage and over time. Digital information in full text form will be relatively cheap to store and use compared to other forms, such as image, sound, video and multimedia information. Full-text takes up less space in storage and less bandwidth in transmission than other forms. The modes of usage for full-text are relatively well, though by no means completely, understood, and so the delivery and access software is relatively more stable and less subject to costly turnover than that for information in other forms.

Usage will also substantially affect operating costs. Healthy demand for particular information objects will push the digital archives that provide them to shift delivery to more costly on-line storage or to sophisticated and expensive systems of hierarchical storage. The costs of software and intellectual property may also be pegged to demand. Another usage factor affecting costs is the high variability in the kinds of user access devices. For many types of information objects, modes of access are still closely tied to the type of workstation available to the user. The tradeoff for the digital archives is either to limit access only to "approved" devices or to support multiple platforms and incur the added associated costs in hardware, software, intellectual property and systems engineering.

Operating costs will also vary over time. Storage costs will likely continue to decline both absolutely and relative to other cost factors. The costs of access and of managing property rights transactions are relatively high today because the supporting systems are virtually non-existent; these systems are developing very rapidly and their relative costs will likely also fall. The costs of the property rights transactions themselves, however, cannot be reasonably predicted at this stage and the danger is that costs will rise to the point that they become a barrier to access. In the long run, the cost factor that will most likely determine the success or failure of digital archives is the investment in the systems engineering and infrastructure needed to support highly distributed network-based functions.

Migration costs will vary depending on the complexity of the original data objects, the frequency of migration, the extent to which the functionality for computation, display, indexing, and authentication must be maintained, and the need to compensate for acquisition of intellectual property rights. Migration costs are much greater for complex objects, such as geographic information systems, where it is necessary to retain multiple formats of data, color display, and complex relations between the "layers" of a geographic information system than for flat files of data or ASCII characters. Computer models that drive artificial intelligence systems are of little long-term value if their computation capabilities are not retained; but migrating the models from one generation of software to the next involves complex and expensive transformations. Some types of digital information may require frequent migration -- as often as every three to five years -- if they are stored in formats that are subject to frequent change. Digital archives may have to compensate for intellectual property rights and may be required to purchase software or site licenses in order to migrate digital information stored in proprietary formats.

Planning for migration is difficult because there is limited experience with the types of migrations needed to maintain access to complex digital objects. When a custodian assumes responsibility for preserving a digital object it may be difficult to predict when migration will be necessary, how much reformatting is needed, and how much migration will cost. There are no reliable or comprehensive data on costs associated with migrations, either for specific technologies and formats or for particular collections.

-- Cost Modeling

Answering questions about the costs and affordability of digital information and of preserving it in a system of digital archives is essential but exceedingly complex. There is a large array of cost factors to understand for a panoply of differing kinds of digital information objects in which numerous parties have a variety of different kinds of interests. A multidimensional matrix -- with present and future stakeholders mapped along one axis, kinds of digital information along a second, and cost factors against a third -- might serve well as a framework for systematically assessing the value of digital information and the affordability of preserving it.

There is much we need to know before we can fully elaborate an economic framework of this kind, and much to be learned from detailed studies of a wide variety of public and private institutional experiences with preserving digital materials. A number of organizations, including the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and the Inter-university Consortium for Political and Social Research (ICPSR), have relatively long and rich experiences in maintaining huge archives of certain kinds of digital information, and they are increasingly doing so in a distributed environment over electronic networks. Yet, despite these various experiences, little systematic understanding has yet emerged of the actual costs of digital archiving. As a result, there is almost no sense of the detailed interplay of cost factors that might promote the kinds of specialization, division of labor and competition needed, in turn, to drive digital archives not just to manage costs against a standard of information quality and integrity, but to strive vigorously to lower those costs while maintaining and improving the standard of quality.[19]

To advance our overall understanding of the long-term implications of digital preservation, and to help define a context for developing the economy of archival management, we need formal, detailed models of the actual costs now being encountered in digital archives. One such model, which projects the costs needed to preserve the digital products of Yale's Project Open Book, is presented below. Within the larger framework of archival activity, this model represents just one small, restricted domain -- namely, the preservation of images of textual documents produced by a research university library in a networked context. The analysis embedded in the Yale model illustrates how we can begin systematically to enrich our understandings of the value and usefulness of specific digital preservation efforts. It also suggests where we need to focus our energies as we undertake the fundamental work needed to advance the economy of digital archiving.

The Yale Cost Model

Assumptions about the value of digital information frequently turn on assertions that cheap storage and easy access are uniquely available in the digital environment. At the same time, many publishers, librarians and others contemplating the digital future express considerable fear that the digital world of information will not prove less expensive than the traditional paper environment, but in fact more expensive. Given a body of digital works, what resources are needed to store and provide access to them indefinitely into the future?

Through Project Open Book, the Yale University Library has accumulated an archives of over 2,000 digital texts. The texts are collections of black and white TIFF images at 600 dots per inch resolution under CCITT Group IV compression. Based on experience it has gained to date in handling these texts, the Yale Library has begun constructing a model that projects the costs of storage and access for a much larger digital archives, and has provided some of the details for presentation here (see also Waters 1994).

The fundamental assumption of the projection in the Yale model is that the digital archives is built primarily for the Yale community and is composed not of converted volumes, but of newly published digital texts distributed in bit-mapped form.[20] According to the model, the archives purchases or leases these new texts and they accumulate at an annual rate of 200,000 volumes per year. Only a fraction of the newly acquired material is used each year. The usage rate is based on actual circulation rates at the Yale Library of about 15% of volumes. Such a rate is not unusual for large research libraries and seems appropriate as a basis for modeling a digital archives intended to collect material for future use as well as present use. The usage rate in the digital archives, however, is assumed to be 20%, slightly more than 15% in the traditional library on the theory that access is easier and therefore demand is more for materials in digital form.

Costs of Digital Archives

Table 1 below contains a summary of the estimated digital archives storage and access costs per volume for Year 1 (See Appendix 2 for details). Components of storage costs in the digital archives include storage device costs -- a jukebox, in this model -- storage media costs, and the costs of operating and maintaining the storage equipment, and of periodically refreshing the media and migrating

the data. The costs of providing access to the digital archives include providing a document server and software for the server that will support client machines on users' desks, maintaining the server and access software and of operating the server, and printing on demand a selected portion of volumes used and of delivering those printed copies to the user. Systems engineering and management overhead costs are also included for both the storage and access services. Note, however, that many costs -- for example, those for acquisitions, cataloging and reference services -- are factored out of the model for purposes of this discussion.

In this model, all equipment and software are capitalized over a life of five years, whereupon it is assumed to be obsolete. The media are assumed to be refreshed and the data migrated as the technology changes on the same five year cycle. Hardware and software maintenance costs as well as equipment operations costs are estimated as a proportion of the original purchase price. The model assumes the cost of hardware and software, as well as migration and operational services all to be declining at a relatively rapid rate of 50% every 5 years. On the other hand, the management and systems engineering services costs are calculated as a proportion of the salary of a full-time, benefited employee whose salary rises each year at a 4% rate of inflation.

Cost Factors	Costs per Volume in Year 1
Storage	
Device Costs	$0.97
Device Maintenance	0.40
Operations Costs	0.40
Media Costs	0.25
Media Refreshment and Data Migration	0.49
Storage Systems Engineering	0.04
Storage Management Overhead	0.03
Total Storage Costs per Volume	**$2.58**
Access	
Document Server	$2.13
Server Maintenance	0.88
Server Operations	0.88
Access Software	1.22
Software Maintenance	0.50
Access Systems Engineering	0.04
Access Management Overhead	0.03
Printing	0.96
Delivery	0.09
Total Access Costs per Volume	**$6.72**

Table 1. Costs of the Digital Archives

One can reasonably assume that local printers are available to network users of the digital archives, and that many users would not avail themselves of the printing and delivery services posited here. This model is built on the assumption that only 10% of the use of the archives would result in use of the archives' own printing and delivery services. The unit costs represented here are thus 10% of the actual unit costs for on-demand printing and delivery. Moreover, the printing costs include an estimated charge for copyright clearance and are assumed to be stable into the future. Delivery charges are assumed to be labor intensive and to inflate by 4% per year over time.

Depository Library Costs

Are the storage and access costs for this first year high or low? How would the costs compare of over time? What is an appropriate standard of comparison?

To generate a comparison benchmark, Yale imagined that the same 200,000 volumes that it assumed it would acquire each year for the digital archives could just as readily be acquired in paper form. Note that this kind of benchmark is only possible for document-like objects and is simply not an option for many of the materials that originate in digital form and cannot exist in any other form. Yale further imagined, because its libraries are full or nearly so, that all new volumes would be stored, not in a new and expensive full-service library in the center of campus but in a low-cost depository facility from which a highly responsive service would deliver needed items directly to faculty offices and student rooms. Leaving all competitive rivalries aside, Yale granted for purposes of the model that it could not build and run a depository facility at unit costs lower than the published unit prices at the Harvard Depository Library. Yale thus projected its depository costs based on the depository prices charged by Harvard. For purposes of this presentation, other relevant costs -- for example, those for acquisitions, cataloging, and reference services -- are factored out and assumed to be equal to similar costs factored out of the digital archives model.

Table 2 contains a summary of the storage and access costs per volume in a depository library for Year 1. Given an average size of document and an average number of documents per linear foot, the projected library storage costs for paper documents are easily calculated from the published price list of the Harvard Depository Library, and presumably include in their base a means of recovering the costs of building construction and other capital costs as well as maintenance costs and management overhead. The costs of providing access to the paper-based depository library consist of four components: retrieval from the depository shelf, transfer to the campus service point, circulation of the volume and delivery to the user. Estimates of retrieval and courier service costs are also based on the published price list of the Harvard Depository Library.[21] The estimate of circulation cost is derived from actual circulation costs at Yale. And a cost is assigned for delivery service to the faculty office or student room as a substitute for reader use of a browsable stack in a full service library. All these costs are assumed to rise with inflation at an annual rate of 4%.

Cost Factors	Costs per Volume in Year 1
Total Storage Costs per Volume	**$0.21**
Access	
Retrieval	$2.63
Courier	0.27
Circulation	0.60
Delivery	0.90
Total Access Costs per Volume	**$4.40**

Table 2. Depository Library Costs

Today, in Year 1, the differences between the unit costs of the depository library compared to those of the digital archives are striking. The storage costs in this model are more than 12 times higher for a digital archives composed of texts in image form, and the access costs are 50% higher. Skepticism about the purported cost advantages of digital libraries over traditional libraries thus seems well-founded in this model, at least in Year 1. What would happen over the longer term?

A Ten-Year Scenario

Proponents of digital libraries rest their case, at least in part, on arguments about the rapidly declining rates of technology costs. This highly simplified model highlights that argument and sets up a stark contrast between the digital archives, the costs of which (except for management overheard, systems engineering, demand printing and delivery) decline at a steady rate, while the costs of the depository library rise by an inflationary rate each year.[22] If these assumptions are set in play over a ten year period, the changes in unit costs are remarkable (see Table 3).

	Year 1	Year 4	Year 7	Year 10
Depository Library				
Depository Storage Costs Per Volume	$0.21	$0.24	$0.27	$0.30
Depository Access Costs Per Volume Used	$4.40	$4.95	$5.57	$6.27
Digital Archives				
Digital Storage Costs Per Volume	$2.58	$1.73	$1.18	$0.82
Digital Access Costs Per Volume Used	$6.72	$4.84	$3.60	$2.79

Table 3. Projected Costs Per Volume Over 10 Years

The calculation for this table ignores the annual carrying costs a digital archives would incur to maintain the material it acquires each year. Instead, the table presents the costs as if each year was the first year of an archives' operations. By centering on the operational costs in this way, the table clearly reveals that real unit costs of storage for the digital archives fall by about 70% over the period. However, in Year 10 they still remain more than double the unit costs of storage in the depository library. By contrast, unit costs of access in the digital archives fall to less than half of the unit access costs in the depository library over the period, overtaking them in about Year 5.

So formulated, the model seems to confirm a widespread sense of the value of the digital world in providing easy access to digital information. However, if over the next decade storage in the digital archives is managed the same as storage in conventional paper-based libraries -- that is, if the number of volumes stored is the same as the number of volumes acquired -- then the overall cost advantage would still favor the depository library. In Year 10, the cost to store 200,000 new volumes in the digital archives is $164,000; the costs of access to 20%, or 40,000, of the volumes is $112,000; together these yield a total cost of $276,000. By contrast, the cost to store 200,000 new volumes in the depository library in Year 10 is $61,000; the costs of access to 15%, or 30,000, volumes would be $188,000; together these total $249,000.

Obstacles and Prospects for Digital Archives

With its stark assumptions that seem to favor the digital archives and its surprising results that favor the paper-based library, the cost analysis presented

here raises a critical question. If the costs of providing storage and access to texts in digital image form are truly greater than the costs of providing storage and access to the same texts in paper form, are the highly touted advantages of the digital environment merely a chimera?

There are at least two answers to this question. One is to challenge the high costs of the digital archives over time by asserting, for example, that the assumed rate of decline in technology-based costs should be steeper. This kind of challenge to the model and its results is risky for two reasons. First, although there may be evidence today that some technology costs are declining at a steeper rate than the overall rate posited in the model, it is difficult to argue from the evidence that the overall rate should be lowered or that such a lowered rate could be sustained over the period. It is difficult because, second, any expectation of declining costs has to be balanced against the equally persistent expectation of rising functionality, which tends to drive technology-based costs up -- or, at least, to slow their decline.

Another, and perhaps more fruitful, answer is to think of the organization of digital information storage and access in fundamentally different terms from those which govern the conventional paper library. As we have seen, one of the significant qualities of digital information is that it lives in a networked environment. Given sufficient capacity or bandwidth, adequate security and reliability and wide extension of the network, one can alter a fundamental assumption of the model, namely, that the digital archives, like the paper-based library, best serves its client community by taking physical possession of *all* the materials it acquires. Instead, one can imagine a distributed storage environment, supported by various kinds of consortia, partnership and other kinds of contractual arrangements with suppliers. Under these arrangements, a digital archives or other user agent could purchase or license the full 200,000 volumes, but then secure the right, either for a period of time or in perpetuity, to move a digital work from another archives on the network into local storage only when it is needed. And exercising a fail-safe prerogative might comprise

	Year 1	**Year 4**	**Year 7**	**Year 10**
New Volumes	*200,000*	*200,000*	*200,000*	*200,000*
Depository Library (volumes stored = new volumes)				
Estimated annual use (15% of volumes)	*30,000*	*30,000*	*30,000*	*30,000*
Depository Storage Costs for New Volumes	$42,769	$48,110	$54,117	$60,874
Depository Access Costs for Volumes Used	$132,090	$148,583	$167,136	$188,005
Total Depository Storage and Access Costs	**$174,859**	**$196,693**	**$221,253**	**$248,879**
Digital Archives (volumes stored = volumes used)				
Estimated annual use (20% of volumes)	*40,000*	*40,000*	*40,000*	*40,000*
Digital Storage Costs for Volumes Used	$103,208	$69,371	$47,207	$32,763
Digital Access Costs for Volumes Used	$268,870	$193,400	$143,977	$111,785
Total Digital Storage and Access Costs	**$372,078**	**$262,771**	**$191,184**	**$144,549**
Difference: Depository - Digital	**($197,219)**	**($66,078)**	**$30,069**	**$104,331**

one highly specialized definition of need.

Table 4. Costs for All Volumes Stored and Used

With the prospect of a viable system of distributed networked-based digital archives, one can thus cast the model of a digital archives in the following way: assume that it begins storing permanently the volumes it has acquired only as they are needed. Observe in Table 4, the effects of this changed assumption. Note that in Year 10, because the digital archives is now storing only the volumes used, its storage costs have dropped from $160,000 (as calculated above) to $32,763. Still, the depository library continues to hold an overall cost advantage until Year 7. Access costs shift to the benefit of the digital archives in Year 6 and storage costs follow suit in the next year. Beginning in Year 7 and continuing on through the rest of the period, the digital archives, conceived in a digital networked environment, begins to demonstrate its affordability compared to conventional paper-based modes of information storage and access.

A different construction of the organization of digital storage and access compared to paper-based storage and access thus leads to a compellingly different construction of the relative economies. Even under highly restrictive assumptions -- the very specialized case of bit mapped images of text and the arguably conservative expectations about the rate of decline in technology-based costs -- the digital archives embedded in a highly distributed network of information resources begins to look economically attractive in a relatively short time.[23]

Now, if one begins to relax these restrictive assumptions, then the model of distributed digital archives starts bearing even more economic fruit. Incorporate in the model a faster rate of decline in technology-based costs. Or, rather than bit-mapped, inject into the model a different format, such as compressed TeX, which is much less storage intensive than bit-mapped images. In all these cases, one can expect the costs to fall relative to both the digital and paper-based scenarios initially presented here.

Richer and more detailed cost models than the simple analytic model advanced here -- ones that include costs of acquisition, cataloging, reference services and so on -- are needed to accurately assess the value and affordability of the digital environment. The Yale model, however, has the distinct advantage of helping to reveal that the key that unlocks the path to the economies of the digital environment is not technological, but organizational. Developing suitable and effective modes of distribution in a networked environment that lead to cost effective digital archives for preserving digital information is an organizational task requiring much ingenuity and numerous creative partnerships and alliances of various kinds among stakeholders.[24] We can look forward to this organizational effort as the digital environment matures, but a key question still remains: who will pay?

-- Financing

A key question in the management of archival costs for operations and migration, of course, is how to balance them with income, either from a sponsoring organization or philanthropy that absorbs the costs or from direct or indirect charges for use. Uncertainty about the answer to this question, as much as any other factor, creates a significant barrier to the coherent, systematic preservation of digital information. Some general actions might help relieve the uncertainty about archival costs. For example, tax incentives and accounting rules that favor the preservation of digital information in archives as investment

in long-term capital stock might spur the growth of digital archives. Otherwise, solutions to cost questions are likely to be found in relation to specific bodies of digital materials and the communities that are interested in them.

For some kinds of digital information direct charging for use will be entirely acceptable to the relevant user communities. One can imagine making an actuarial calculation of the lifetime cost of preserving a digital information object, finding creators/providers/owners with an economic interest in paying to preserve their information, and constructing an archival service that functions much like a safety deposit system for digital information objects. As facilities are developed and refined to exact charges, conduct transactions for intellectual property and maintain confidentiality, and as experience with such mechanisms grows, some communities of interest that presently resist the notion of charging for information services, such as archiving, may grow less resistant. In any case, more imaginative solutions need to be found by asking hard questions about who benefits from the archived information, when do they benefit and do the answers suggest how the costs of preservation might be afforded. Some instructive examples are beginning to emerge from communities in which the members have asked and tried to answer these hard questions.

Consider physics, for example. The direct beneficiaries of archived physics information are physicists and related professionals, who have, as their professional organization, the American Physical Society (APS). The APS already publishes the central corpus of physics information. It keeps a copy of most of what it has published and owns the copyrights to its publications. It has stability and longevity (it was founded earlier than the New York Public Library and considerably earlier than most extant commercial organizations), and a membership keenly aware of the value of its publications and thereby able to help select the most valuable among them. The Society has the technical skill and organizational wherewithal to manage its archives in digital form. It has mechanisms which could finance the digital archives, including member dues and access charges, and it has recently adopted a digital form of publication for its key journals. Moreover, after careful study, the Society has recently decided to embark on a systematic program that would lead it to build complete digital archives of its publications, past and present, and to maintain them into the future.

Like the APS, other professional associations are actively reconsidering how economically to preserve their key information assets in digital form. The Association for Computing Machinery (ACM), for example, has also embarked on a ambitious program designed to place the entire ACM literature in an on-line digital library and is explicitly questioning how to afford it: "Should there be an archiving fee replacing the current practice of page charges? Should uncited items be deleted from the archive after a minimum holding time? Should highly cited items be guaranteed a permanent place in the archive?" (Denning and Rous 1995: 103). Other associations are more concerned with preserving critical data than with the published literature. The American Geological Institute, a federation of geoscientific and professional associations, is in the process of forming a National Geoscience Data Repository System to capture and preserve geoscientific data (see American Geological Institute 1994). As these various efforts mature, and others emerge, what kind of

imaginative solutions to the problems of managing the costs of digital preservation will they produce or, with incentives and stimuli from partners and competitors, could they be provoked to produce?

Summary and Recommendations

Buckminster Fuller used to tell a story about the Master of one of the Cambridge (England) Colleges, who noticed a deep crack in the massive beam supporting the college's dining hall. Not knowing to whom he should report the problem, the Master eventually notified the Royal Forester. The Forester replied that he had been expecting the call. The Forester's predecessor's predecessor, he said, had planted the tree for the new beam, and it was ready. This, Fuller noted, was how a society ought to work.

The Commission on Preservation and Access and the Research Libraries Group (RLG) together have asked this Task Force on the Archiving of Digital Information to report in effect on ways that society should work with respect to the cultural record it is now creating in digital forms. The digital environment is still relatively uncultivated at this stage, but the need is urgent, the time is opportune and the conditions are fertile for a strong, far-sighted set of cultivating actions to help ensure that the digital record ultimately matures and flourishes.

Major Findings

By analyzing the emerging digital environment, and the place of digital archives there, we have aimed in this report to identify the most demanding preservation issues and to frame them for appropriate action. Asked to focus on the notion of "technology refreshing," we have found "data migration" to be a richer and more fruitful concept for describing what is necessary to protect the integrity of the cultural record. Prompted to "envision possible end-states," we have reached several general conclusions that inform our view of viable options and next steps. In sum, we have concluded that:

- -- The first line of defense against loss of valuable digital information rests with the creators, providers and owners of digital information.
- -- Long-term preservation of digital information on a scale adequate for the demands of future research and scholarship will require a deep infrastructure capable of supporting a distributed system of digital archives.
- -- A critical component of the digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating and providing access to digital collections.
- -- A process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information.
- -- Certified digital archives must have the right and duty to exercise an aggressive rescue function as a fail-safe mechanism for preserving valuable digital information that is in jeopardy of destruction, neglect or abandonment by its current custodian.

On the basis of these conclusions, we now advance our recommendations for next steps. We formulate them under three general headings: pilot projects, needed support structures, and the development of best practice. We urge our sponsors to take the recommended actions, either separately or together and in

concert with other individuals or organizations as appropriate. In particular, the Commission on Preservation and Access and the Research Libraries Group should:

Pilot Projects

1. **Solicit proposals from existing and potential digital archives around the country and provide coordinating services for selected participants in a cooperative project designed to place information objects from the early digital age into trust for use by future generations.**

   Action is urgently needed to ensure that documents, software products and other digital information objects that document the early digital age from 1945 to 1990 are preserved before they slip irrevocably away. A project designed with this particular focus as a cooperative venture would have the added advantage of providing a testbed for developing a system of linked but distributed digital archives. Because the objects in this focal area are at such risk of loss, the project would also provide a useful means of exploring the operations of certification and fail-safe mechanisms for digital archives.

2. **Secure funding and sponsor an open competition for proposals to advance digital archives, particularly with respect to removing legal and economic barriers to preservation.**

   The recent competition sponsored by the National Science Foundation generated an enormous amount of creative thinking about and commitment to the development of digital libraries. A similar approach is needed for digital archives. The competition might best be focused on fostering creative alliances, especially with publishers, and practical, joint efforts designed to lower the legal and economic barriers to the effective operation of digital archives.

3. **Foster practical experiments or demonstration projects in the archival application of technologies and services, such as hardware and software emulation algorithms, transaction systems for property rights and authentication mechanisms, which promise to facilitate the preservation of the cultural record in digital form.**

   Only through early and active use will digital archives be able to influence the development of key new technologies and services and help to ensure that they support information longevity. The matrix of kinds of information, information attributes and stakeholders suggested in this report can provide a useful framework for crafting specific targeted efforts. There are many cells in that matrix which are presently void, but which could be filled and verified (or nullified) through a series of imaginative demonstration projects. Moreover, there is growing need for evidence that digital archives can practically and effectively incorporate in their daily operations automated systems for emulating obsolete hardware and software, transacting intellectual property and using cryptographic and other mechanisms for creating trusted distribution channels for digital information.

Support Structures

4. Engage actively in national policy efforts to design and develop the national information infrastructure to ensure that longevity of information is an explicit goal.

The Task Force envisions a highly distributed network of linked digital information archives as the environment in which digital information will flourish over the long-term. Communication and information network policy decisions regarding pricing, security and network extension will greatly affect the viability of these archives and their efforts to preserve digital information. These policy decisions need to be informed with an understanding of the importance and complexity of digital preservation. Development of an infrastructure conducive to preservation might also be aided by consideration of tax incentives and accounting practices that treat the creation of digital information archives favorably as an investment in the nation's long-term capital stock.

5. Sponsor the preparation of a white paper on the legal and institutional foundations needed for the development of effective fail-safe mechanisms to support the aggressive rescue of endangered digital information.

The Task Force has suggested that applying the concept of abandonment to the domain of intellectual property might serve as a one kind of legal foundation for a fail-safe mechanism. A legally mandated system of deposit for published works, in which publishers are required to place with a certified digital archives a copy of a work in a standard archival format, might serve as another kind of foundation. Limits in time and expertise have prevented the Task Force from giving systematic attention to the full array of options for instituting fail-safe mechanisms for digital information. Such attention is urgently needed.

6. Organize representatives of professional societies from a variety of disciplines in a series of forums designed to elicit creative thinking about the means of creating and financing digital archives of specific bodies of information.

The American Physical Society, the Association for Computing Machinery and the American Geological Institute have each embarked on a creative mission to build digital archives of the published material that it owns or of other digital information relevant to its membership. What can others learn from these ventures? What other possibilities exist for treating digital preservation as a problem to be solved by those with clear interests in the long-term availability and use of specific sets of related information objects?

7. Institute a dialogue among the appropriate organizations and individuals on the standards, criteria and mechanisms needed to certify repositories of digital information as archives.

If valued digital information is to be preserved for future generations, repositories claiming to serve an archival function must be able to prove

that are who they say they are and that they can deliver on their preservation promise. One of the ways for digital archives to provide such proof is to submit to an independently-administered program for archival certification. The appropriate individuals and organizations need now to begin systematically to identify and describe the standards, criteria and mechanisms for archival certification and thereby launch the process that would lead ultimately to a formal certification program.

8. Identify an administrative point of contact for coordinating digital preservation initiatives in the United States with similar efforts abroad.

There is considerable evidence of worldwide interest in the means of preserving digital information. Since the Task Force on Archiving of Digital Information issued its draft report in August 1995, the European Union, the Consortium of University Research Libraries in Great Britain and a national Working Party in Australia on the management of material in electronic format have all generated working papers on the topic of digital preservation, commented on the Task Force draft report and invited international collaboration. We need a high level point of contact in the United States to help identify and facilitate international collaborative efforts.

Best Practices

9. Commission follow-on case studies of digital archiving to identify current best practices and to benchmark costs in the following areas:

a. The design of systems that facilitate archiving at the creation stage.

The recently issued "Electronic Federal Depository Library Program: Transition Plan FY1996-FY1998" proposes to replace the existing depository library program for documents with one providing electronic access. The Superintendent of Documents will be responsible for ensuring long-term access to this digital information. What steps is the Federal Government taking at the point of document creation to facilitate the task of preserving access? Outside the realm of government, how are publishers redesigning the creation process to support their electronic publishing programs? What software are they using and how have they influenced software producers to modify their development of their products?

b. Storage of massive quantities of culturally valuable digital information.

There is little good experience yet in storing in digital form massive quantities of materials traditionally regarded as culturally valuable, such as books and serials. Organizations have, however, developed large digital archives for other kinds of culturally important information. Examples include the archives of census data, remote sensing satellite imagery, weather data, or commercial data such as insurance or medical records. What can be learned from experience in these areas about the means and costs of ensuring the longevity of digital information?

c. Requirements and standards for describing and managing digital information.

Descriptive information about the content of digital objects, their origins and provenance and their management over time is critical for both long-term preservation and future use of digital information. Standards and best practices for describing and managing digital information are needed to track changes in ownership or control over digital objects throughout their life cycle, to administer intellectual property rights, and to document any changes in the format and structure of digital objects that may ensue from migration. A responsible digital archives must provide to its users what it knows about the provenance and context of its objects so that users can make informed decisions about the reliability and quality of the evidence before them. Standards bodies, professional associations in the archival, library and information technology fields, and the constituencies represented by the Commission on Preservation and Access and RLG thus need to collaborate in an evaluation and expansion of descriptive standards and practices so that they satisfy the special requirements of digital preservation and access.

d. Migration paths for digital preservation of culturally valuable digital information

Data migration is a common, if difficult, practice as businesses and other organizations preserve their essential business records through successive changes in hardware and business management software. Cultural archives that have been collecting digital objects have also had to begin migrating them as the hardware and software on which they were created has become obsolete. What is the range of experience of different organizations with archiving different types of content? What can be learned and generalized from these experiences? How do strategies compare among different organizations for archiving similar materials. Are there economies of scale that could be achieved by combining efforts across digital archives? What are the costs of the different strategies employed? What strategies have failed? In what ways have practices improved over time?

Given the analysis in this report and its findings, we expect the Commission and the Research Libraries Group to pursue these recommendations on a national and, where appropriate, an international front, and to generate dialogue, interaction and products that will advance the development of trusted systems for digital preservation. There are numerous challenges before us, but also enormous opportunities to contribute to the development of a national infrastructure that positively supports the long-term preservation of digital information. Such an infrastructure is a desirable outcome that will benefit us only if we conceive and structure it to benefit those served by our successors' successors.

Notes

[1] The Task Force on the Archiving of Digital Information is indebted to Margaret Adams and Thomas Brown of the Center for Electronic Records at the National Archives and Records Administration for essential information in this brief account of the preservation of the 1960 Census data and its aftermath. Responding to a query from the Task Force, they reviewed the correspondence, memoranda and records schedules of the Center for Electronic Records and prepared a useful historical narrative about the actual treatment of the 1960 census data. See Adams and Brown (1996).

In a separate correspondence addressed to the Task Force, Adams emphasizes that the story about there being only two machines in 1985 that could read the 1960 census data is "apocryphal" and observes that custodians did not, because of machine obsolescence, lose any census data that they intended to save. However, she notes that custodians did form their intentions based on the best knowledge that they had at the time (Adams 1995). The Report of the Committee on the Records of Government (1985: 88 n. 40) makes a similar point. It argues that "in the early years of data processing, archivists and historians failed to recognize the potential use of many machine-readable data files" and so routinely destroyed some of those files pertaining to earlier censuses.

In the early days of data processing, custodians of data records simply had no basis to anticipate potential uses that seem obvious today to us in hindsight. There is a potent research topic here. As the Task Force goes on to argue in the next sections of this report, archiving is, in the end, a matter of human organization and the exercise of best judgment. The power and danger of technological obsolescence is that it sometimes requires judgments from actors and organizations on a schedule that may not match their willingness and competence to make them.

[2] Margaret Hedstrom, the former Chief of the State Records Advisory Service in the New York State Archives and Records Administration, brought this example to the attention of the Task Force.

[3] The term “refreshing” is not a term that is used widely in the literature on archival theory and practice. Throughout this report, the Task Force follows the sense of the word as defined in its charge to mean “to copy the records onto newer media and into newer formats” (see Appendix 1). This sense of the word is not the only one in common use, however. As David Gracy (1995) notes in a comment to the Task Force, “refresh” is frequently used in data processing contexts to refer to records stored logically in a database, rather than to information stored physically on a digital medium. In those contexts, the phrase “to refresh the data” means to rebuild the database, replacing old data with new data, and so conveys a sense of destruction rather than preservation. The Task Force clearly wants to avoid the implications of this alternative sense of the term and so follows a

definition that is closer to that of Van Bogart, who defines refreshing of tape media as the "periodic retensioning of tape, or the rerecording of recorded information onto the same tape (or different tape) to refresh the magnetic signal. In the audio/video tape community, refreshing generally refers to retensioning of the tape, but it can also refer to the copying of one tape to another" (1995: 33).

[4] Margaret Adams (1995) of the National Archives and Keith Parrot (1995), who writes on behalf of the Towards Federation 2001 Management of Material in Electronic Format Working Party in Australia, both suggest that the Task Force give more attention to accumulated experiences with digital information, particularly in the library and archival communities. One can readily acknowledge the impressive amount of work within those communities over several decades on the preservation of digital materials. See, for example, Bearman and Hedstrom (1993), Conway (1994, 1996a,b), Conway and Weaver (1994), Dollar (1976, 1992), Fishbein (1974, 1975), Hedstrom (1991, 1995), Hedstrom and Kowlowitz (1988), Kenney (1993), Mohlhenrich (1993), Neavill (1984), O'Toole (1989) and United States National Historical Publications and Research Commission (1991). After reviewing the evidence, one can still hold, as the Task Force does, that the development of widely trusted systems for the preservation of digital information is still in its infancy.

[5] Some commentators on the first draft of the Task Force report objected that the distinction between archives and libraries was too simplistic (see Bearman 1996; Consortium of University Research Libraries 1996; Graham 1995b; Parrott 1995). For example, the definition of archives used in the earlier draft seemed unreasonably to ignore or to subordinate to the central feature of long-term preservation key archival concerns for object integrity, such as provenance and context (see, for example, Hedstrom 1995). Rather than introduce a different concept from either library or archive, such as "digital research library" or "digital heritage preservation agency," to emphasize its concern with preservation, the Task Force has opted in the present version of the report to introduce and emphasize the concept of object integrity in its definition of archive. In the next section, the Task Force elaborates the implications of the concept for the successful operation of digital archives.

[6] Defining and conserving structure and format as a matter of content integrity is not, of course, a new problem with respect to information objects. For one good overview of the issue regarding paper documents, see O'Toole (1989). See also Conway (1996b: 9).

[7] The use of the HyperText Markup Language (HTML) -- a partial subset of SGML -- has, of course, seen a phenomenal growth in recent years as a means of marking up documents and other information objects for the World Wide Web. Compared to SGML and TeX, however, it is relatively impoverished as a means of representing complex formats and structures within textual documents.

[8] The Task Force itself took this view in releasing its draft report. It presented four versions of the report -- a version in Microsoft Word, a version in the HyperText Markup Language (HTML), a version in the Portable Document Format supported in Adobe Acrobat, and an ASCII version. Each version had its particular targeted audience and its corresponding advantages and disadvantages, and neither the Task Force nor its sponsors designated any of the versions as "canonical."

[9] Margaret Hedstrom (1995) usefully observes that the difficulties that the digital environment presents for fixing information as discrete objects extends not only to formal publications but also to more ephemeral kinds of transactions as well. She writes that "our culture is inventing new forms of documentary evidence that are technologically complex, yet socially and culturally primitive. Electronic mail systems, for example, deliver a variety of messages in an undifferentiated structure, ranging from formal transactions, to deeply personal communications between friends, to anonymous postings on bulletin boards."

[10] Metadata, which refers to information about information, is sometimes used as a generic term for systems of reference. We avoid use of the term in this report because it conveys a tone of jargon and because its use in the literature is varied and imprecise. In its report on *Preserving Scientific Data on Our Physical Universe*, for example, the National Research Council (1995: 36-39) uses the term to include any and all documentation that serves to define and describe a particular scientific database. Other uses of the phrase elsewhere in the literature are closer to the more limited sense of referential systems that we use here to mean systems of citation, description and classification. The preference for the term metadata in those other cases appears to flow from the felt need to emphasize the special referential features needed in the digital environment and to distinguish those special features from those of more traditional systems of citation, description and classification. See, for example, Moen (1995), Weibel (1995) and Bearman and Sochats (1995).

[11] Clifford Lynch (1996: 142) observes that in linking names and locations in systems of citation for digital objects, we seem to be demanding a higher standard than for information objects in the print environment.

[12] See the National Research Council (1995a: 50), which argues that "the only effective and affordable archiving strategy is based on distributed archives managed by those most knowledgeable about the data." Against this view, the background paper on digital preservation prepared for the European Commission takes the position that "decentralised models in which the archiving of electronic publications is delegated to publishers or network resource providers are therefore not recommended. Preservation of electronic materials is better guaranteed by local storage under the control of the deposit library" (Owen and van de Walle 1995: 11). The apparent opposition between centralized and distributed (or decentralized) models is misleading, however. Viewed globally, a national depository model for archiving is merely a special case of a decentralized model, where the

criteria for distributed collection depend on national boundaries, rather than discipline interests or functional divisions.

[13] Several readers of the first draft of the Task Force report expressed skepticism that creators will accept responsibility for archiving digital information (Graham 1995b, Owen and van de Walle 1995: 11, and Parrott 1995). The "Draft Statement of Principles for the Preservation of and Long Term Access to Australian Digital Objects," however, follows the Task Force formulation when it asserts that "creators of original digital objects hold an initial, and in many cases a continuing responsibility for their preservation; creators have the power to facilitate or deny the continuing existence of digital information" (Lyall 1996).

[14] Some readers of the first draft of the Task Force report objected that the notion of a fail-safe mechanism unreasonably assumes that archives will have knowledge of the existence and impending demise of digital information (Parrott 1995). Moreover, they questioned the apparent negative connotations of the phrase "fail-safe mechanism": it implies efforts to "prevent" the demise of an object rather than causing it to endure (Graham 1995) and it is "essentially punitive" because it assumes that those who have accepted preservation responsibilities cannot be trusted (Parrott 1995). To overcome these negative connotations, the commentators suggested strengthening existing structures of cooperation and coordination (Parrott 1995) and abandoning the idea of a "fail-safe mechanism" in favor of establishing a "trigger" mechanism (Graham 1995).

Having considered these concerns, the Task Force remains convinced of the need for a "fail-safe mechanism" for digital archiving. In the world of paper, there are numerous ways of knowing of the existence of information objects and of their potential demise; it is no less reasonable in the digital world to assume that similar, and perhaps better, ways of knowing will develop. Moreover, the Task Force does not view a fail-safe mechanism as a negative force or as punitive to an a custodian of information. Rather, it regards the mechanism as an enabling tool for a rescuing agent and, therefore, as contributing to the general social good. If a custodian were to regard a fail-safe mechanism as a threat, then perhaps such a perception would serve the greater good of stimulating responsible archiving behavior. In any case, the Task Force intends the fail-safe mechanism only as a last resort. Such a protection of last resort would be necessary even if there were appropriate "trigger mechanisms" causing archiving behavior proactively to occur and if cooperative and collaborative structures were running smoothly. It would be necessary because custodians can ignore "triggers" to right behavior and because cooperation can break down.

[15] The Task Force is grateful to Scott Bennett (1995) for drawing attention to the need for legal principles that could be called into play when digital works in copyright are in danger of being abandoned or destroyed. Bennett particularly cites the work of Patterson and Lindberg (1991: 204-207) on the concept of abandonment. Patterson and Lindberg, however, do not develop the concept and merely refer to it in the context of their discussion

of the fair use of copyrighted materials. The sense of the Task Force is that an interpretation of fair use, however articulate it may be, is too slender a thread on which to base a fail-safe mechanism for the protection of our cultural heritage. Alternatively, the legal deposit system currently in place in the United States is not designed organizationally as a fail-safe mechanism for the preservation of culturally significant works. In the judgment of the Task Force, the time is right for a comprehensive examination of possible fail-safe mechanisms and a thorough understanding of both their legal and organizational foundations.

[16] There are at least two models of certification. On the one hand, there is the audit model used, for example, to certify official depositories of government documents. The depositories are subject to periodic and rigorous inspection to ensure that they are fulfilling their mission. On the other hand, there is a standards model which operates, for example, in the preservation community. Participants claim to adhere to standards that an appropriate agency has certified as valid and appropriate; consumers then certify by their use whether the products and services actually adhere to the standards. In its call for certified digital archives, the Task Force has not judged the relatives merits of applying either of these particular kinds of models of certification.

[17] Inattention to even minor details can sometimes compromise all other efforts to preserve the integrity of a digital information object and can hamper efforts to authenticate the objects to future users. For example, consider the Task Force's experience with the first draft of its report. Each page of the draft contained a date. The date changed with each date of printing. During the composition phase, before the Task Force publicly released the draft, the changing date helped distinguish earlier versions of the report from one another. When it formally released the first draft, however, the Task Force did not fix the date on each page so that it would not change with each printing. Every time a reader retrieved and printed a copy of the report from its on-line location, the date printed on the report thus changed, generating much confusion about whether the Task Force had changed the substance of the draft report or not. Diane Hopkins (1995) of the World Bank kindly called the Task Force's attention to this subtle but very serious problem for preserving digital documents from a word processing environment.

[18] There has recently emerged the Standard for the Exchange of Product Model Data (STEP). It is a data representation standard intended to provide a complete unambiguous representation of data throughout its life cycle, including an archival phase. The application of STEP would facilitate the storage of data and a data model or representation, and would enable the data to be retrieved even if the software that created it is defunct or not able to run on the current computer hardware or operating system. Early implementations of the STEP standard are currently under development in industry for specific kinds of product documentation. The standard may have wider application, but these developments serve as an example of a comprehensive effort to design information systems to a standard that

considers archival issues for information objects from the point of their creation. See, for example, Herbst and Malle (1995).

[19] The National Research Council, in its study on *Preserving Scientific Data on Our Physical Universe*, is conscious of the costs of archiving but mainly points to what it regards as the chronic underfunding of archival efforts. See, for example, the working papers for the report in National Research Council (1995b: 79). There are, however, throughout the working papers and the final report a number of references that together suggest that the problem of cost is more than a simple matter of underfunding. For example, the final report (1996a: 29,31) notes that "NOAA's budget for its National Data Centers in FY 1980 was $24.6 million, and their total data volume was approximately one terabyte. In FY 1994, the budget was on $22.0 million (not adjusted for inflation), while the volume of their combined data holdings was about 220 terabytes!" The report cites this comparison as evidence of the low priority given to data archiving in budgets. Such an assessment may in fact be accurate, but surely the change in funding also reflects the effects of dramatic increases in the cost efficiencies of digital storage technologies over the period.

So what level of funding priority is appropriate? The working papers that accompany the final National Research Council report tentatively probe several alternatives. The Department of Energy, for example, is cited with approval because one of its programs provides "about 15 percent of the total program budget for data management and archiving." Elsewhere, the working papers refer to the budget of the Jet Propulsion Laboratory, which manages the archives of a joint U.S. and French satellite data system, and observes that the cost of maintaining the resulting data sets for any 10-year period "appears to average less than 1 percent of the cost of collecting the data set" (National Resource Council 1995b: 79, 91). Without a closer benchmarking and deeper analysis of the cost factors at work in these various archival programs, it is, of course, impossible to judge the relative merits of these two different measures of appropriate funding, and the final report of the National Research Council does not attempt to make such a judgment. However, the National Research Council work does suggest that the archives of the nation's scientific data offer fertile ground for systematic studies that would yield very useful benchmarks and models of archival costs.

[20] One commentator on the first draft of this report read the use of the Yale cost model here to mean that the Task Force "envisages a total replacement of one format with another" (Parrott 1995). The Task Force envisions no such thing. Although the cost projections in this model are based in part on the costs of maintaining digital objects that the Yale Library converted from its paper and microfilm collections, the plainly stated assumption is that the digital archive being modeled here is composed not of converted material but of newly published material in digital form.

[21] The Task Force is indebted to Curtis Kendrick, the Assistant Director for the Depository in the Harvard University Library, for carefully reading this

section in the first draft of the report and for providing an updated price list for the relevant Depository services cited in the Yale cost model. See Kendrick (1995). The presentation here incorporates the corrected prices. Kendrick also notes that the Harvard Depository is currently averaging about a 3% circulation rate and only operates for 250 days per year and observes that the Yale model extrapolates the Harvard unit prices to an hypothetical environment operating 360 days per year at a 15% circulation rate. Kendrick worries that the Harvard fee structure could not sustain the heightened level of activity posited in the Yale model. Although it is possible that a higher level of activity would result in higher unit prices, it is more likely that economies of scale would obtain and that unit costs would in fact be lower. In simply extrapolating the current costs, and assuming no economies (or diseconomies) of scale, the Yale model takes a relatively conservative position.

[22] In his comments on the first draft of the Task Force report, Scott Bennett of the Yale Library (1995) urges caution upon the Task Force in its use of the Yale cost model, which assumes a steady decline in hardware, software and certain operational costs over its ten-year horizon. "No trend line in technology or economics," he writes, "can reasonably be expected to continue indefinitely." The Task Force agrees that any projections beyond the 10 year horizon set in the Yale cost model would be highly suspect, and believes that the assumptions posited in the model about declining unit costs for hardware and software within the ten-year period are reasonable and plausible.

[23] Fully distributed storage models, like the one envisioned here for digital materials, do exist for paper and microfilm materials. The Center for Research Libraries serves its members in allowing them to distribute among themselves the costs of storing paper or film materials. The preservation program for brittle books also operates on a distributed storage model in that it assumes no duplication in the creation and storage of masters that will serve the needs of the entire community. One commentator on the first draft of the Task Force report suggested that these kinds of distributed storage models for paper and microfilm would, for the purpose of cost comparison, provide a true parallel environment to the distributed model posited for digital materials (Bennett 1995). The fully distributed models for storing paper and film, however, work only for materials that are highly specialized and for which there is little demand. The models are not otherwise widely applied because the means of distributing paper and film among the nodes in the distributed network are so costly and cumbersome. The proposed comparison, while interesting theoretically, would contribute little to the argument of the Task Force.

[24] The JSTOR project, sponsored by the Andrew W. Mellon Foundation, is one such creative venture that is seeking to demonstrate the organizational economies of scale unique to digital environment. See Bowen (1995).

References

Ackerman, M. S., and R. T. Fielding (1995) "Collection Maintenance in the Digital Library." *Proceedings of Digital Libraries '95*, Austin, Texas, June, pp. 39-48. Also available at [URL:http://csdl.tamu.edu/DL95].

Adams, Margaret O. (1995) "Comments on draft report on digital archiving." Electronic mail correspondence addressed to John Garrett and Donald Waters, November 2.

Adams, Margaret O. and Thomas E. Brown (1996) "Historical Narrative on Data From the 1960 Census." Unpublished report. Washington, D.C.: Center for Electronic Records, National Archives and Records Administration. April 3.

American Geological Institute 1994 *National Geoscience Data Repository System. Phase II: Planning and Pilot Study Proposal*. Alexandria, Virginia: American Geological Institute.

Atkinson, Ross 1986 "Selection for Preservation: A Materialistic Approach." *Library Resources & Technical Services* 30 (October/December): 341-353.

Barlow, John Perry (1994) "The economy of ideas." *Wired* (March): 84-90, 126-129.

Bearman, David (1989) *Archival Methods*. Archives and Museum Informatics Technical Report, vol. 3, no. 1. Pittsburgh: Archives and Museum Informatics.

Bearman, David (1995) *Archival Strategies*. Pittsburgh: Archives and Museum Informatics.

Bearman, David (1996) "RLG/CPA Report." Electronic mail correspondence addressed to Margaret Hedstrom, February 19-26.

Bearman, David, and Margaret Hedstrom (1993) "Reinventing Archives for Electronic Records: Alternative Service Delivery Options." In Margaret Hedstrom, ed. *Electronic Records Management Program Strategies*, Archives and Museum Informatics Technical Report, No. 18, pp. 82-98.

Bearman, David, and Ken Sochats (1995) "Metadata Requirements for Evidence." Draft paper dated October available at [URL:http://www.oclc.org:5046/conferences/metadata/requirements.txt].

Bennett, Scott 1995 "Comments on Draft Report." Electronic mail correspondence addressed to ARCHTF-L, October 23.

Bowen, William 1995 "JSTOR and the Economics of Scholarly Communication." Unpublished paper. The Andrew W. Mellon Foundation. September 18.

Cole, Timothy W. and Michelle M. Kazmer (1995) "SGML as a Component of the Digital Library." *Library Hi Tech*, 13(4): 75-90.

Committee on the Records of Government (1985) *Report*. Reprinted, Malabar, Florida: Robert E. Kieger Publishing Company, Inc., 1988.

Consortium of University Research Libraries (1996) *Response to the CPA/RLG Draft Report "Preserving Digital Information."* Leeds, England: Consortium of University Research Libraries.

Conway, Paul (1994) "Digitizing Preservation: Paper and Microfilm Go Electronic." *Library Journal* 119 (February 1): 42-45.

Conway, Paul (1996a) "Selecting Microfilm for Digital Preservation: A Case Study from Project Open Book." *Library Resources and Technical Services* 40(1): 67-77.

Conway, Paul (1996b) *Preservation in the Digital World*. Washington, D.C.: Commission on Preservation and Access.

Conway, Paul and Shari Weaver (1994) *The Setup Phase of Project Open Book*. Washington, D.C.: Commission on Preservation and Access.

Creque, Stuart A. (1995) "Why Johnny Can't Read His Data." *Wall Street Journal*, June 5: p. A14.

Davis, Stephen P. (1995) "Digital Image Collections: Cataloging Data Model and Network Access." In: Patricia A. McClung, ed. *RLG Digital Image Access Project: Proceedings From an RLG Symposium Held March 31 and April 1, 1995, Palo Alto, California*. Palo Alto, CA: Research Libraries Group,. pp. 45-59. Also available at [URL:http://www.columbia.edu/cu/libraries/inside/projects/diap/paper.html].

Denning, Peter J. and Bernard Rous (1995) "The ACM Electronic Publishing Plan." *Communications of the ACM* 38(4): 97-103.

Dollar, Charles M. (1976) "Computers, the National Archives, and Researchers." *Prologue* 8(1): 29-34.

Dollar, Charles M. (1992) *Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Methods.* Macerata: University of Macerata Press.

Duranti, Luciana (1995) "Reliability and Authenticity: The Concepts and Their Implications." *Archivaria* 39 (Spring): 5-10.

Eisenbeis, Kathleen M. (1995) *Privatizing Government Information: The Effects of Policy on Access to Landsat Satellite Data*. Metuchen, N.J.: The Scarecrow Press, Inc.

European Commission (1996) *A Study of Issues Faced by National Libraries in the Field of Deposit Collections of Electronic Publications. Report of the Workshop held in Luxembourg, December 18, 1995.* Luxembourg: European Commission, Directorate General XIII-E/4, February.

Fishbein, Meyer H. (1974) "Report on the Committee on Data Archives and Machine Readable Records -- Fiscal Year 1973." *APDA* 1(2): 5-10.

Fishbein, Meyer H. (1975) "ADP and Archives: Selected Publications on Automatic Data Processing." *American Archivist* 38(January): 31-42.

Garrett, John R., *et. al*. (1993) "Toward an Electronic Copyright Management System." *Journal of the American Society for Information Science* 44(8): 468-473.

Gavrel, Sue (1986) "Preserving Machine-Readable Archival Records: A Reply to John Mallinson." *Archivaria* 22(Summer): 153-55.

Getz, Malcolm (1992) "Information Storage." Unpublished manuscript, Vanderbilt University, February 27.

Gracy, David B. 1995 "Digital Information draft." Electronic mail correspondence addressed to Maxine Sitts of the Commission on Preservation and Access, December 11.

Graham, Peter S. (1994) *Intellectual Preservation: Electronic Preservation of the Third Kind.* Washington, D.C.: Commission on Preservation and Access.

Graham, Peter S. (1995a) "Requirements for the Digital Research Library." *College and Research Libraries*, July, 56(4): 331-339.

Graham, Peter S. (1995b) "PDI Draft." A series of electronic mail correspondence addressed to ARCHTF-L, October 5-10.

Hedstrom, Margaret (1991) "Understanding Electronic Incunabula: A Framework for Research on Electronic Records." *American Archivist* 54 (3): 334-354.

Hedstrom, Margaret (1995) "Electronic Archives: Integrity and Access in the Network Environment." In Stephanie Kenna and Seamus Ross, eds., *Networking in the Humanities: Proceedings of the Second Conference on Scholarship and Technology in the Humanities, held at Elvetham*

Hall, Hampshire, UK, 13-16 April, 1994. London: Bowker-Saur, pp. 77-95.

Hedstrom, Margaret and Alan Kowlowitz (1988) "Meeting the Challenge of Machine Readable Records: A State Archives Perspective." *Reference Studies Review* 16(1-2): 31-40.

Herbst, Axel and Bernhard Malle (1995) "Electronic Archiving in the Light of Product Liability." In *Proceedings of Know Right '95*. Vienna: Oldenbourg Verlag, pp. 455-460.

Hopkins, Diane 1995 "Correct Date of Draft Report?" Electronic mail correspondence addressed to John Garrett and Donald Waters, September 19.

Kendrick, Curtis (1995) "Draft Report Version 1.0 of the Task Force on Archiving Digital Information." Memorandum addressed to Barbara Graham, Associate Director for Administration and Programs in the Harvard University Library, October 17.

Kenney, Anne R. (1993) "Digital-to-Microfilm Conversion: An Interim Preservation Solution." *Library Resources & Technical Services* 37: 380-401.

Lesk, Michael (1990) *Image Formats for Preservation and Access: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access*. Washington, D.C.: Commission on Preservation and Access.

Lesk, Michael (1992) *Preservation of New Technology: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access*. Washington, D.C.: Commission on Preservation and Access.

Levy, David M. and Catherine C. Marshall (1995) "Going Digital: A Look at Assumptions Underlying Digital Libraries." *Communications of the ACM* 38(4): 77-83.

Lyall, Jan 1996 "Draft Statement of Principles for the Preservation of and Long Term Access to Australian Digital Objects." Canberra: National Library of Australia.

Lynch, Clifford (1993) *Accessibility and Integrity of Networked Information Collections*. Office of Technology Assessment, Congress of the United States, July 5.

Lynch, Clifford (1994a) "The Integrity of Digital Information: Mechanics and Definitional Issues." *Journal of the American Society for Information Science* 45(10): 737-744.

Lynch, Clifford (1994b) "Uniform Resource Naming: From Standards to Operational Systems." *Serials Review* 20 (4): 9-14.

Lynch, Clifford (1996) "Integrity Issues in Electronic Publishing." In Robin P. Peek and Gregory B. Newby, eds., *Scholarly Publishing: The Electronic Frontie*r, Cambridge: The MIT Press, pp. 133-145.

Mallinson, John C. (1986) "Preserving Machine-Readable Archival Records for the Millenia." *Archivaria* 22(Summer): 147-52.

Michelson, Avra and Jeff Rothenberg (1992) "Scholarly Communication and Information Technology: Exploring the Impact of Changes in the Research Process on Archives." *American Archivist* 55 (Spring): 236-315.

Moen, Bill (1995) "Metadata for Network Information Discovery and Retrieval." *Information Standards Quarterly* 7(2): 1-4.

Mohlhenrich, Janice, ed. (1993) *Preservation of Electronic Formats: Electronic Formats for Preservation.* Fort Atkinson, Wis.: Highsmith.

National Academy of Public Administration (1989) *The Effects of Electronic Recordkeeping on the Historical Record of the U.S. Government: A report for the National Archives and Records Administration.* Washington, D.C.: National Academy of Public Administration.

National Research Council (1995a) *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*. Washington, D.C.: National Academy Press.

National Research Council (1995b) *Study on the Long-Term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers*. Washington, D.C.: National Academy Press.

Neavill, Gordon B. (1984) "Electronic Publishing, Libraries, and the Survival of Information." *Library Resources & Technical Services*, 28 (January): 76-89.

O'Toole, James M. (1989) "On the Idea of Permanence." *American Archivist* 52(1): 10-25.

Owen, J. S. MacKenzie and J. van de Walle (1995) *ELDEP Project: A study of issues faced by national libraries in the field of deposit collections of electronic publications. Background Document for the ELDEP Workshop, Luxembourg, December 18, 1995*. The Hague: NBBI.

Parrott, Keith (1995) "Comments on Task Force Draft Report." Electronic mail correspondence sent on behalf of the Towards Federation 2001 Management of Material in Electronic Format Working Party in Australia and addressed to ARCHTF-L, October 31.

Patterson, L. Ray and Stanley W. Lindberg (1991) *The Nature of Copyright: A Law of Users' Rights*. Athens: University of Georgia Press.

Pelikan, Jaroslov (1992) *The Idea of the University: A Reexamination*. New Haven: Yale University Press.

Rothenberg, Jeff (1995) "Ensuring the Longevity of Digital Documents." *Scientific American*, 272 (January): 42-47.

United States Census Bureau (1993) "Levels of Access and Use of Computers: 1984, 1989, 1993." *Current Population Survey Reports*. Population Division, Education and Social Stratification Branch: [URL:http://www.census.gov/ftp/pub/population/socdemo/computer/compusea.txt]

United States National Historical Publications and Research Commission (1991) *Research Issues in Electronic Records.* St. Paul: Minnesota Historical Society.

The University of the State of New York, *et al.* (1988) *A Strategic Plan for Managing and Preserving Electronic Records in New York State Government: Final Report of the Special Media Records Project*. Albany: New York State Education Department.

Van Bogart, John W. (1995) *Magnetic Tape Storage and Handling: A Guide for Libraries and Archives*. Washington, D.C.: Commission on Preservation and Access.

Waters, Donald J. (1994) "Transforming Libraries Through Digital Preservation." In Nancy E. Elkington, ed. *Digital Imaging Technology for Preservation: Proceedings from an RLG Symposium held March 17 & 18, 1994*. Mountain View, CA: Research Libraries Group, pp. 115-127.

Waters, Donald J. (1996a) *Realizing Benefits from Inter-Institutional Agreements: The Implications of the Draft Report of the Task Force on Archiving of Digital Information*. Washington, D.C.: The Commission on Preservation and Access. Also available at [URL:http://arl.cni.org/arl/proceedings/127/waters.html].

Waters, Donald J. (1996b) "Archiving Digital Information." A presentation to the OCLC Research Library Directors' Conference, Dublin, Ohio, March 12, 1996.

Weibel, Stuart (1995) "Metadata: The Foundations of Resource Description." *D-Lib Magazine* (July). [URL:http://www.dlib.org/dlib/july95].

Wiederhold, Gio (1995) "Digital Libraries, Value and Productivity." *Communications of the ACM* 38(4): 85-96.

Appendix 1

Commission on Preservation and Access and Research Libraries Group

Task Force on Archiving of Digital Information Proposed Charge

Preamble

Continued access indefinitely into the future of records stored in digital electronic form cannot under present circumstances be guaranteed within acceptable limits. Although loss of data associated with deterioration of storage media is an important consideration, the main issue is that software and hardware technology becomes rapidly obsolescent. Storage media becomes obsolete as do devices capable of reading such media; and old formats and standards give way to newer formats and standards. This situation holds both for electronic records derived through conversion from some analog form (paper, film, video, sound etc.), and for records that originated in electronic form.

It has been proposed that one solution to this problem is to "refresh" the stored records at regular intervals, that is, to copy the records onto newer media and into newer formats. While this approach is simple in concept, implementation raises a number of issues, most of which are not technological. How, for example, can we guarantee that owners of electronic records will faithfully pursue such a refreshing mandate indefinitely into the future? Does the very nature of this question imply the need to contract such tasks to one or more organizations who can be relied upon to carry the refreshing torch forward? There are also important legal, economic, cultural, and technical questions.

Charge

The Commission on Preservation and Access and the Research Libraries Group join together in charging a Task Force to:

- Frame the key problems (organizational, technological, legal, economic etc.) that need to be resolved for technology refreshing to be considered an acceptable approach to ensuring continuing access to electronic digital records indefinitely into the future.
- Define the critical issues that inhibit resolution of each identified problem.
- For each issue, recommend actions to remove the issue from the list.
- Consider alternatives to technology refreshing.
- Make other generic recommendations as appropriate.

The Task Force may also wish to envision possible end-states that portray an environment in which technology refreshing is accepted as a routine approach; and scenarios for achieving such end-states. An important goal is to understand what might constitute "best practices" in the area of technology refreshing.

The Task Force shall consult broadly among librarians, archivists, curators, technologists, relevant government and private sector organizations, and other interested parties.

The Task Force is requested to complete an interim report by May, 1995 or thereabouts that can be circulated widely among interested communities to obtain feedback as input to a final report to be completed summer, 1995. This final report will constitute the key product of the Task Force.

Appendix 2

This appendix presents the underlying, detailed analysis for the cost model presented in the main body of the report. The model, developed at the Yale University Library, compares the relative costs of storage and access in a paper-based library and a projected digital archives of image-based documents. Here is a table of the principal assumptions and variables used in building the model.

Assumptions and variables

General		
Volumes		
AnnualVolumesAcquired	200,000	
Usage		
AnnualCirculationRate	15%	*=Calculate browses and circulations of newly acquired materials
AnnualVolumesCirculated	30,000	*=AnnualVolumesAcquired*AnnualUseRate
Other		
Inflation	4.00%	*=per year
InterestRate	7.00%	*=for financing
WorkingDaysPerYear	360	
Depository Model (based on Harvard Depository rates)		
Storage Costs		
StorageCostPerLinearFoot	$2.78	*=per year
VolumesPerLinearFt	13	*=approximately
StorageCostPerVolume	$0.21	*=StorageCostPerLinearFoot/VolumesPerLinearFt
Depository Use		
DepositoryUseRate	15%	*=AnnualCirculationRate
DepositoryVolumesUsed	30,000	*=DepositoryUseRate*AnnualVolumesAcquired
Retrieval		
RetrievalCost	$2.63	*=as published by the Harvard Depository Library
VolumesPerRetrieval	1	*=worst case
RetrievalCostPerVolume	$2.63	*=RetrievalCost/VolumesPerRetrieval
Courier--Base		
DepositoryDailyCourierCharge	$15.75	*=Flat fee for up to 4 cu ft per delivery
DepositoryDailyCourierChargePerVolumeUsed	$0.19	*=DepositoryDailyCourierCharge*WorkingDaysPerYear/DepositoryVolumesUsed
Courier--Per Volume		
DepositoryCourierChargePerVolUsed	$0.08	*=as published by the Harvard Depository Library
Circulation		
DepositoryCirculationCostPerUse	$0.60	*=Equivalent to the circulation cost per book (check out and return)
Delivery to Office		
DepositoryDeliveryCostPerUse	$0.90	*=Benchmarked to stack maintenance costs

Assumptions and Variables, continued

Digital Archives Model

General

HardwareCostInflation	-13%	*=((1/2)^(1/5))-1
SoftwareCostInflation	-13%	*=((1/2)^(1/5))-1
HardwareAmortizationPeriod	5	*=years
SoftwareAmortizationPeriod	5	*=years
HardwareMaintenanceRate	10%	*=of original purchase price, annually
SoftwareMaintenanceRate	10%	*=of original purchase price, annually
FTEBenefittedSalaryManagement	$55,000	*=Annually
FTEBenefittedSalarySystemsEngineer	$55,000	*=Annually

Storage--General

PagesPerVolume	215	*=per experience in Project Open Book
GBytesPerVolume	0.015	*=per experience in Project Open Book
GBytesPerPage	0.00007	*=GBytesPerVolume/PagesPerVolume; B&W, TIFF Image, CCITT Fax 4 Compression
VolumesPerGByte	66	*=TRUNC(1/GBytesPerVolume,0)

Storage Device

DigitalRoboticStorageCostPerGByte	$263.05	*= Source: Malcolm Getz, Information Storage, 1992; less 4 years deflation
DigitalRoboticStorageCostPerVolume	$3.99	*=DigitalRoboticStorageCostPerGByte/VolumesPerGByte
AmortizedDigitalStorageCostPerVolume	$0.97	*=-PMT(InterestRate,HardwareAmortizationPeriod,DigitalRoboticStorageCostPerVolume)
DigitalStorageMaintenanceCostPerVolume	$0.40	*=DigitalRoboticStorageCostPerVolume*HardwareMaintenanceRate
DigitalStorageOperationsRate	10%	*=Estimated annual cost as a percent of original cost (AmortizedDigitalStorageCostPerVolume)
DigitalStorageOperationsCostPerVolume	$0.40	*=DigitalStorageOperationsRate*DigitalRoboticStorageCostPerVolume
DigitalStorageSystemsEnginFTE	15%	*=Estimated time of an FTE devoted to systems engineering of storage operation
DigitalStorageSystemsEnginCostPerVolume	$0.04	*=DigitalStorageSystemsEnginFTE*FTEBenefittedSalarySystemsEngineer/AnnualVolumesAcquired
DigitalStorageManagementFTE	10%	*=Estimated time of an FTE devoted to management of storage operation
DigitalStorageManagementCostPerVolume	$0.03	*=DigitalStorageManagementFTE*FTEBenefittedSalaryManagement/AnnualVolumesAcquired

Storage Media

CostPerDisk	$70.00	*=per experience in Project Open Book
UsableGBytesPerDisk	1	*=1.3*80% (note: cannot fill the entire disk)
VolumesPerDisk	69	*=TRUNC(UsableGBytesPerDisk/GBytesPerVolume,0)
MediaCostPerVolume	$1.01	*=CostPerDisk/VolumesPerDisk
AmortizedMediaCostPerVolume	$0.25	*=-PMT(InterestRate,HardwareAmortizationPeriod,MediaCostPerVolume)
MediaRefreshmentCostRate	200%	*=Need to buy the disk plus operations charge for migration of data
MediaRefreshmentCostPerVolume	$0.49	*=MediaRefreshmentCostRate*AmortizedMediaCostPerVolume

Access--General

DigitalIncreasedUseRate	5%	*=Assumed increase over DepositoryUseRate
DigitalUseRate	20%	*=DepositoryUseRate+DigitalIncreasedUseRate
DigitalVolumesUsed	40,000	*=DigitalUseRate*AnnualVolumesAcquired
SimultaneousUsers	112	*=CEILING(DigitalVolumesUsed/WorkingDaysPerYear,1)

Access--Document Server

DocumentServerCost	$350,000	*=Scale of machine needed for number of simultaneous users
DocumentServerCostPerVolumeUsed	$8.75	*=DocumentServerCost/DigitalVolumesUsed
AmortizedDocumentServerCostPerVolumeUsed	$2.13	*=AmortizedDocumentServerCost/DigitalVolumesUsed
DocumentServerMaintenanceCostPerVolume	$0.88	*=DocumentServerCostPerVolumeUsed*HardwareMaintenanceRate
DocumentServerOperationsRate	10%	*=Estimated annual cost as a percent of original cost (DocumentServerCostPerVolumeUsed)
DocumentServerOperationsCostPerVolumeUsed	$0.88	*=DocumentServerOperationsRate*DocumentServerCostPerVolumeUsed
DigitalAccesSystemsEnginFTE	25%	*=Estimated time of an FTE devoted to systems engineering of access operation
DigitalAccessSystemsEnginCostPerVolume	$0.04	*=DigitalAccessSystemsEnginFTE*FTEBenefittedSalarySystemsEngineer/AnnualVolumesAcquired
DigitalAccessManagementFTE	20%	*=Estimated time of an FTE devoted to management of access operation
DigitalAccessManagementCostPerVolume	$0.03	*=DigitalAccessManagementFTE*FTEBenefittedSalaryManagement/AnnualVolumesAcquired

Access--Software

AccessSoftwareCost	$200,000	*=Investment needed for number of simultaneous users
AccessSoftwareCostPerVolumeUsed	$5.00	*=DigitalAccessSoftwareCost/DigitalVolumesUsed
AmortizedAccessSoftwareCostPerVolumeUsed	$1.22	*=-PMT(InterestRate,SoftwareAmortizationPeriod,DigitalAccessSoftwareCostPerVolumeUsed)
AccessSoftwareMaintenanceCostPerVolumeUsed	$0.50	*=DigitalAccessSoftwareCostPerVolumeUsed*SoftwareMaintenanceRate

Printing and Delivery

PrintingRateofTotalVolumes	10%	*=Estimate
PrintingCostsPerPage	$0.033	*=Per print shop at Yale

PrintingCostsPerVolume	$7.10	*=PrintingCostsPerPage*PagesPerVolume
PrintingCopyrightClearanceCharge	$2.50	*=Estimate on average
TotalPrintingCostsPerVolume	$9.60	*=PrintingCostsPerVolume+PrintingCopyrightClearanceCharge
PrintingCostPerTotalVolumesUsed	$0.96	*=totalprintingcostspervolume*printingrateoftotalvolumes
PrintedCopyDeliveryCostPerTotalVolumesUsed	$0.09	*=DepositoryDeliveryCostPerUse*PrintingRateofTotalVolumes

Depository Library Model

The model for the depository library takes the unit costs for all variables in Year 1 and inflates them annually by the assumed inflation factor. Volumes stored are assumed in the model to be equal to the number of volumes acquired. The costs for all volumes stored and used is the product of the volumes stored and used and the total unit costs of storage and access.

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
Costs per Volume										
Depository Storage										
Storage Costs	$0.21	$0.22	$0.23	$0.24	$0.25	$0.26	$0.27	$0.28	$0.29	$0.30
Depository Storage Costs Per Volume	**$0.21**	**$0.22**	**$0.23**	**$0.24**	**$0.25**	**$0.26**	**$0.27**	**$0.28**	**$0.29**	**$0.30**
Depository Access										
Retrieval	$2.63	$2.74	$2.84	$2.96	$3.08	$3.20	$3.33	$3.46	$3.60	$3.74
Courier										
Basic Daily charge	$0.19	$0.20	$0.20	$0.21	$0.22	$0.23	$0.24	$0.25	$0.26	$0.27
Per Volume Charge	$0.08	$0.09	$0.09	$0.09	$0.10	$0.10	$0.11	$0.11	$0.11	$0.12
Circulation	$0.60	$0.62	$0.65	$0.67	$0.70	$0.73	$0.76	$0.79	$0.82	$0.85
Delivery	$0.90	$0.94	$0.97	$1.01	$1.05	$1.09	$1.14	$1.18	$1.23	$1.28
Depository Access Costs Per Volume Used	**$4.40**	**$4.58**	**$4.76**	**$4.95**	**$5.15**	**$5.36**	**$5.57**	**$5.79**	**$6.03**	**$6.27**
Costs for All Volumes Stored and Used (volumes stored = new volumes)										
New Volumes	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000
Estimated annual use (15% of volumes)	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000
Depository Storage Costs for New Volumes	$42,769	$44,480	$46,259	$48,110	$50,034	$52,035	$54,117	$56,281	$58,533	$60,874
Depository Access Costs for Volumes Used	$132,090	$137,374	$142,869	$148,583	$154,527	$160,708	$167,136	$173,821	$180,774	$188,005
Total Depository Storage and Access Costs	**$174,859**	**$181,854**	**$189,128**	**$196,693**	**$204,561**	**$212,743**	**$221,253**	**$230,103**	**$239,307**	**$248,879**

Digital Archives Model

The model for the digital archives takes the unit costs for the variables related to hardware and software in Year 1 and deflates them annually by the assumed value of the hardware and software cost deflation factors. Systems engineering and management overhead costs inflate at the assumed rate of inflation. Printing costs are assumed to be constant. Delivery costs are inflated from Year 1 at the assumed rate of inflation. Volumes stored are assumed in the model to be equal to the number of volumes used, which is calculated as a product of the digital use rate and the total number of volumes acquired. The costs for all volumes stored and used is the product of the volumes stored and used and the total unit costs of storage and access.

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
Costs per Volume										
Digital Storage										
Device Costs	$0.97	$0.85	$0.74	$0.64	$0.56	$0.49	$0.42	$0.37	$0.32	$0.28
Device Maintenance	$0.40	$0.35	$0.30	$0.26	$0.23	$0.20	$0.17	$0.15	$0.13	$0.11
Operations Costs	$0.40	$0.35	$0.30	$0.26	$0.23	$0.20	$0.17	$0.15	$0.13	$0.11
Media Costs	$0.25	$0.22	$0.19	$0.16	$0.14	$0.12	$0.11	$0.09	$0.08	$0.07
Media Refreshment and Data Migration	$0.49	$0.43	$0.38	$0.33	$0.28	$0.25	$0.22	$0.19	$0.16	$0.14
Storage Systems Engineering	$0.04	$0.04	$0.04	$0.05	$0.05	$0.05	$0.05	$0.05	$0.06	$0.06
Storage Management Overhead	$0.03	$0.03	$0.03	$0.03	$0.03	$0.03	$0.03	$0.04	$0.04	$0.04
Digital Storage Costs Per Volume	**$2.58**	**$2.26**	**$1.98**	**$1.73**	**$1.52**	**$1.34**	**$1.18**	**$1.04**	**$0.92**	**$0.82**
Digital Access										
Document Server	$2.13	$1.86	$1.62	$1.41	$1.23	$1.07	$0.93	$0.81	$0.70	$0.61
Server Maintenance	$0.88	$0.76	$0.66	$0.58	$0.50	$0.44	$0.38	$0.33	$0.29	$0.25
Server Operations	$0.88	$0.76	$0.66	$0.58	$0.50	$0.44	$0.38	$0.33	$0.29	$0.25
Access software	$1.22	$1.06	$0.92	$0.80	$0.70	$0.61	$0.53	$0.46	$0.40	$0.35
Software Maintenance	$0.50	$0.44	$0.38	$0.33	$0.29	$0.25	$0.22	$0.19	$0.16	$0.14
Access Systems Engineering	$0.04	$0.04	$0.04	$0.05	$0.05	$0.05	$0.05	$0.05	$0.06	$0.06
Access Management Overhead	$0.03	$0.03	$0.03	$0.03	$0.03	$0.03	$0.03	$0.04	$0.04	$0.04
Printing	$0.96	$0.96	$0.96	$0.96	$0.96	$0.96	$0.96	$0.96	$0.96	$0.96
Delivery	$0.09	$0.09	$0.10	$0.10	$0.11	$0.11	$0.11	$0.12	$0.12	$0.13
Digital Access Costs Per Volume Used	**$6.72**	**$6.00**	**$5.38**	**$4.84**	**$4.36**	**$3.95**	**$3.60**	**$3.29**	**$3.03**	**$2.79**
Costs for All Volumes Stored and Used (volumes stored = volumes used)										
New Volumes	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000
Estimated annual use (20% of volumes)	40,000	40,000	40,000	40,000	40,000	40,000	40,000	40,000	40,000	40,000
Digital Storage Costs for Volumes Used	$103,208	$90,314	$79,108	$69,371	$60,915	$53,575	$47,207	$41,685	$36,903	$32,763
Digital Access Costs for Volumes Used	$268,870	$240,109	$215,114	$193,400	$174,543	$158,176	$143,977	$131,669	$121,009	$111,785
Total Digital Storage and Access Costs	**$372,078**	**$330,423**	**$294,222**	**$262,771**	**$235,459**	**$211,751**	**$191,184**	**$173,355**	**$157,912**	**$144,549**