IT’S ABOUT TIME

Research Challenges in Digital Archiving and Long-term Preservation

Sponsored by

The National Science Foundation

Digital Government Program and Digital Libraries Program

Directorate for Computing and Information Sciences and Engineering

and

The Library of Congress

National Digital Information Infrastructure and Preservation Program

August 2003

It's About Time:
Research Challenges in Digital Archiving and Long-term Preservation

Final Report

Workshop on Research Challenges in Digital Archiving and Long-Term Preservation

April 12-13, 2002

Sponsored by

The National Science Foundation

Digital Government Program and Digital Libraries Program

Directorate for Computing and Information Sciences and Engineering

and

The Library of Congress

National Digital Information Infrastructure and Preservation Program

August 2003

Acknowledgements

I owe thanks to many organizations and individuals who encouraged me to organize this workshop. I am grateful to the National Science Foundation (NSF) and the Library of Congress (LoC) for funding the workshop.[1] I received guidance and support from Steve Griffin, Larry Brandt, Valerie Gregg, and Sue Stendebach of the Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering. In addition to providing partial funding for the workshop, the LoC offered meeting space for the Planning Committee and logistical support for the workshop. Laura Campbell, Associate Librarian for Strategic Initiatives, and staff members Martha Anderson, Caroline Arms, and Carl Fleischhauer offered valuable feedback, helped to coordinate the workshop with NSF and other government agencies, and disseminated information about the research agenda.[2]

I owe special thanks to the Organizing Committee members: Sharon Dawes, Center for Technology in Government, University at Albany; Carl Fleischhauer, Library of Congress; James Gray, Microsoft Research; Clifford Lynch, Coalition for Networked Information; Victor McCrary, National Institute of Standards and Technology; Reagan Moore, San Diego Supercomputer Center; Kenneth Thibodeau, National Archives and Records Administration; and Donald J. Waters, Andrew W. Mellon Foundation. They were instrumental in shaping the agenda, making presentations and leading breakout groups, and in drafting and assembling the final report. All fifty-one participants deserve credit for their active participation and creative contributions.

It was a great pleasure to receive professional support from Jen Engleson Lee, SI Intern; Ann Verhey-Henke, Research Administrator; and Anne Dopkins, Faculty Secretary at the School of Information, University of Michigan. The workshop would not have been possible without them.

Margaret Hedstrom
July 8, 2003

[1] This workshop was funded by NSF Award Number 0214690. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Library of Congress.

[2] A prepublication draft of the Final Workshop was published as Appendix 7, "*It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation*," in Library of Congress (October 2002), *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*, Washington, D.C.: Library of Congress: 205-220.

Table of Contents

Sponsors

National Science Foundation

LAWRENCE BRANDT, Digital Government Program

VALERIE GREGG, Digital Government Program

SUE STENDEBACH, Digital Government Program

STEPHEN M. GRIFFIN, Digital Libraries Program

Library of Congress

LAURA CAMPBELL, Associate Librarian for Strategic Initiatives

MARTHA ANDERSON, Office of Strategic Initiatives

CAROLINE ARMS, Office of Strategic Initiatives

CARL FLEISCHHAUER, Office of Strategic Initiatives

Organizing Committee

MARGARET HEDSTROM, University of Michigan, Chair and Principal Investigator

SHARON DAWES, Center for Technology in Government, University at Albany, State University of New York

CARL FLEISCHHAUER, Library of Congress

JAMES GRAY, Microsoft Research

CLIFFORD LYNCH, Coalition for Networked Information

VICTOR MCCRARY, National Institute of Standards and Technology

REAGAN MOORE, San Diego Supercomputer Center

KENNETH THIBODEAU, National Archives and Records Administration

DONALD WATERS, Andrew W. Mellon Foundation

Executive Summary

One of the marvels of the information technology revolution is the continuous improvement in computer memory and storage performance and their simultaneous drop in cost. Thanks to what has been called "silicon scaling" the processing power of a 1980s vintage mainframe computer now fits on miniscule silicon chips that can be embedded in any number of capture devices from complex remote sensors to consumer digital cameras. Digital storage devices and media have benefitted from similar performance improvements and cost declines. Large organizations routinely add terabytes of storage capacity, and more and more individuals can afford laptop and desktop computers with tens of gigabytes of storage. One might suspect that archiving and preserving digital information would become easier and cheaper as a consequence of these improvements. But from a long-term preservation perspective, there is a dark side to the rapid growth in digital information. The technologies, strategies, methodologies, and resources needed to manage digital information for the long term have not kept pace with innovations in the creation and capture of digital information.

In April 2002, a group of computer scientists, information scientists, archivists, digital library experts, and government program managers met to examine the prospects for advancing computer and information technology research through a research program that addresses the unique challenges of long-term digital preservation. Developing an infrastructure for preserving digital information for future exploitation raises many interesting and difficult issues. The requirements for long-term preservation test the limits of current technologies and information management methodologies. Long-term digital archiving requires systems, institutions, and business models that are robust enough to withstand technological failures, shifting computing platforms and media, changes in institutional missions, and interruptions in management and funding.

This report summarizes the discussions and recommendations of the Workshop on Research Challenges in Digital Archiving and Long-term

Preservation that was sponsored by the National Science Foundation (NSF) and the Library of Congress (LoC). It discusses what is unique about archiving and digital preservation research, explains why solutions are urgently needed to prevent further loss of valuable digital information, and outlines a research agenda. The following points shaped the discussion and provided a framework for the research agenda:

- Digital archiving and preservation present unique research challenges because of concern with the long-term viability of digital information, where "long-term" may simply mean long enough to be concerned about the obsolescence of technology, or it may mean decades or centuries. Digital objects require constant and perpetual maintenance, and they depend on elaborate systems of hardware, software, data and information models, and standards that are upgraded or replaced every few years.

- Accelerating rates of data collection and content creation and the growing complexity of digital information resources tax current preservation strategies designed to archive relatively simple and self-contained collections of data and documents. Even established scientific data archives with a track record of three decades or longer cannot preserve all of the data entrusted to them using current methodologies.

- No acceptable methods exist today to preserve complex digital objects that contain combinations of text, data, images, audio, and video and that require specific software applications for reuse. Raw data is rarely useful to researchers without associated models and analytical tools.

- Libraries, archives, museums, and other cultural institutions that have preservation as part of their core mission need solutions to digital preservation challenges if they are to play a meaningful role in preserving our intellectual and cultural heritage.

- Many government agencies, private corporations, not-for-profit organizations, and even private citizens are now concerned with preserving their own digital information assets. Because digital information is vulnerable to alteration, erasure, and technology obsolescence, digital preservation methods are needed that can span a wide range of operating environments.

- People are an essential part of the process. Today’s curatorial processes for selection, description, and management of digital collections, while remaining important, are very labor intensive. There is a need to understand, evaluate, redesign, and automate many archival and preservation processes to drive down the costs of long-term preservation and to scale them up to the size and complexity of the digital archiving challenge.

- Future users of digital archives will have different needs, expectations, technologies, and analytical tools from those of the communities that initially created the digital content. This raises challenging research questions in the areas of metadata, semantics, and knowledge management technologies that will enable future reuse of collections in digital archives.

- The funding and business models for digital preservation differ considerably from common business models because an archival repository is expected to preserve digital resources even though their value and usefulness may not become apparent until well into the future. In this respect, the economic models for digital preservation resemble the economics of public goods. There is a critical need for research and development of economic and business models that will sustain digital preservation through many downturns in business cycles. New business models are needed to make digital preservation affordable to individuals, government agencies, universities, cultural institutions, and society at large.

- The challenges of maintaining digital archives over long periods of time are as much social and institutional as technological. Even the most ideal technological solutions will require management and support from institutions that in time go through changes in direction, purpose, management, and funding.

What is at stake?

This section introduces a few scenarios to illustrate the nature of the digital archiving research challenges.

Who owns the digital archiving and long-term preservation problem?

Internet search engines crawl the web, copy web pages, and index them so that users have a reasonable chance of finding information relevant to their needs and interests. Large search engine companies like Google™ copy and index billions of web pages and store copies temporarily in caches for use in case the requested page is temporarily unavailable. Search engine companies are in the business of providing tools for searching and navigating the web. They are not in the business of long-term archiving of the web or even a portion of it, nor should they be expected to take on this responsibility. Who will?

The Internet Archive was founded in 1996 to preserve content distributed on the web. In seven years it has developed the largest collection of web pages in the world – about 10 billion web pages, including 200 million pages on the 2000 Election and 500 million pages related to the terrorist attacks of September 11, 2001. The archive is vulnerable to accidents, degrading storage media, and changing data formats. As a public nonprofit organization without a predictable, steady flow of resources, it is seeking stable institutional partners to collaborate in its long-term preservation endeavors.

How will we preserve the valuable digital collections created during the last decade?

During the last decade, libraries, archives, and museums, as well as many scientific, academic and cultural organizations, government agencies, private enterprises, and individual collectors, have assembled valuable collections of digital information. Under the American Memory Program, LoC has led an effort to digitize more than 100 historical collections from materials in its own holdings and from libraries, archives, and museums across the country. The seven million items in the American Memory collections are used daily by teachers, students, scholars, genealogists, and private citizens. How will these collections be carried forward economically for continuing use and enjoyment?

The Digital Libraries Initiative, sponsored by NSF, LoC, the Defense Advanced Research Projects Agency, National Aeronautics and Space Administration, National Library of Medicine, and National Endowment for the Humanities, fostered research and development for hundreds of digital libraries. Many digital library projects started as testbeds and prototypes, but they have evolved into critical research resources for almost every discipline. These resources need to be maintained into the foreseeable future to support ongoing research and teaching, and also to protect several hundred millions of dollars invested to digitize, organize, and provide access.

How will we save "born-digital" content for future reuse and enjoyment?

More and more valuable content is "born-digital" and can only be managed, preserved, and used in digital form. In the last decade, researchers have mapped significant portions of the human genome. Advances in biomedical research depend on building and preserving complex genomic databases. Research in biodiversity and ecosystems, global climate change, meteorology, and space science – to name only a few fields – is built on the ability to combine vast quantities of digital information with complex models and analytical tools.

The entertainment industry has shifted rapidly to digital masters for recorded sound, movies, and television. These provide critical resources for research, historical documentaries, and cultural coherence. Audio, film, and video recordings are replayed, rebroadcast, and reviewed as a source for entertainment and vital connections to the past. Who will take responsibility for preserving our digital cultural heritage? This is a hard challenge because effective and affordable digital preservation methods for multi-media do not exist today, and because much of this content is protected under copyright for extended periods of time. Creative artists and performers are using digital media to create new forms of artistic expression. Even private citizens are seeking ways to manage and preserve their e-mail, online accounts, and digital video and photograph collections.

E-commerce and e-government need better and more affordable digital archiving methods, tools, and technology.

Information systems used in business and government generate enormous quantities of digital information, some of which is worth saving for a long period of time. The aircraft industry depends on software systems to design, manufacture, and maintain complex commercial aircraft. For safety's sake, design specifications, records of manufacturing processes, parts inventories, maintenance records, and performance data, much of which is in digital form, must be kept as long as a particular model of aircraft is in service—a period that can exceed fifty years. The Food and Drug Administration requires pharmaceutical companies to file new drug applications electronically along with documentation of research protocols, tests, and clinical trials. These digital records have to be kept at least as long as a drug is available. Medical records that may be needed for an individual's entire lifetime are becoming electronic. Citizens'rights, such as eligibility for Social Security benefits, are documented in databases that accumulate data through each individual's working life. E-government and e-commerce could flounder if better methods are not found to identify and preserve digital records that are essential for keeping the business running and for maintaining accountability. Stringent requirements for authenticity and integrity permeate this problem.

Invest to save.

Digital archiving research challenges present the type of problem that requires national leadership and government investment. Despite awareness of the digital archiving problem, market forces alone have proven inadequate to develop and provide solutions. In some cases market forces work against long-term preservation by locking customers into propriety formats and systems, adding new features to encourage or force upgrades, and phasing out useful but unprofitable hardware, software, and services. Moreover, government will be a primary beneficiary of digital archiving research, which potentially will lead to more useful and cost-effective preservation systems for scientific and statistical agencies, cultural institutions, and other government programs (such as land management, social security, and intelligence) that need continuing access to digital information for a long period of time. Digital archiving research will also produce immediate and long-term societal benefit by preserving valuable scientific, operational, cultural, and personal information that might otherwise be lost.

Digital Archiving and Long-term Preservation Research Agenda

Given these concerns, a pressing and urgent need exists to develop better solutions for long-term digital preservation for government agencies, libraries, archives, museums, private corporations, and private citizens. Computer and information science research will be advanced by focusing on the challenges of digital archiving. Important new research opportunities have emerged to address storage and processing capacities of digital repositories, interoperability among heterogeneous archival systems, automation of ingest and preservation management processes, support for complex metadata requirements, and search across diverse digital repositories and collections. Research opportunities abound centering on questions of economic and business models for affordable and sustainable long-term preservation. Research is also needed on policies and incentives for long-term preservation and on the economic, social, and legal impediments to digital archiving.

Workshop participants urged the NSF, the LoC, and other government agencies to support a substantial research program that will enhance the state of knowledge and practice for long-term preservation of digital information. The research agenda is organized around four main themes: 1) Technical architectures for archival repositories; 2) Attributes of archival collections; 3) Digital archiving tools and technologies; and 4) Organizational, economic, and policy issues.

Technical Architectures for Archival Repositories

Specification, system and tool development, pilot implementation, and evaluation of repository models.

High-level models for persistent repositories indicate that digital archiving and long-term preservation is best handled by separating archival storage of bits (storage management) from data management, logical representations, and higher level services that can be built on top of a persistent storage architecture. Much more research is needed, along with testing and implementation, to define technical architectures for persistent archives, develop processes for long-term management, build tools that can be used to acquire archival data, prepare data for long-term storage, and manage data over several generations of technology.

Develop a spectrum of repository architectures.

No single architecture will serve all digital archiving requirements because of the diversity of data types, formats, and content that needs to be preserved, the varying size and scale of repositories, and the diverse and changing requirements of user communities. Research is needed along a spectrum of repository types ranging from persistent storage of data (the simplest case), to preservation of complex digital entities (a requirement to preserve programs or develop an ability to reproduce their results), to preservation of services (in which programs and time-varying data interact to produce the result the user sees).

Develop a spectrum of digital archiving services.

Different types of digital entities and different user communities require different services from an archival repository. Research is needed to provide services for complex rights management systems, search across heterogeneous digital collections, track the provenance of digital entities, and recreate and repurpose archived digital collections.

Alternative repository models and interoperability.

Repository architectures also span a spectrum from highly distributed layered models to self-contained repositories that provide end-to-end services. Focused research is needed at each layer of the architecture (such as physical storage, data management, logical representation, and services). Research is also needed to develop the methods and protocols that enable the different layers to interoperate with each other and that allow heterogeneous distributed repositories to exchange content and services.

Scalability and cost.

Issues of scalability and cost pervade the research topics above. Scalability involves preserving not only very large databases (in excess of 10 petabytes), but working across repositories with heterogeneous collections, scores of data formats and logical representations, and billions of entities. Even if storage costs are minimal, long-term digital archiving will not be affordable unless acquisition, description, data management, and access controls are highly automated.

Attributes of Archival Collections

Articulating and modeling of curatorial processes.

Archival collections are created through controlled curatorial processes that include selection, organization, description, quality control, and perpetual care. While many of these processes are necessary and add value to collections, the digital environment is fundamentally reshaping curatorial processes. Some traditional processes may become unnecessary, some may be subject to automation, and many will be transformed in ways that will make them unrecognizable in the near future. Research is needed to provide the innovation necessary to underpin a new set of theoretical, methodological, and technical approaches to curation of digital objects and collections.

Developing appropriate preservation methods for diverse digital objects and collections.

The types of digital entities that warrant preservation range from static digital objects, to complex digital entities, to dynamic objects and streaming data that are updated continually or on an arbitrary basis. Acceptable methods are likely to vary, depending on the types of archived data and its anticipated use. Even static objects raise research questions related to scale and affordability. Preservation of complex objects presents many new challenges such as preserving the functionality, look and feel, and utility of multi-media objects, models and simulations, and analytical and visualization tools. This may require preservation of software and support for emulation. For collections of temporally changing data, methods are needed to characterize and maintain the temporal and procedural relationships between the multiple versions of digital objects.

Aggregation of items and objects into collections.

Physical collections are made up of discrete items, but this is not necessarily the case with digital collections. A digital collection can be created simply by making hyperlinks among widely distributed objects. New processes and policies are needed to build and sustain distributed collections where several organizations participate in the curatorial processes and share preservation responsibilities.

Decision models.

Decision models are needed to support many aspects of digital preservation including selection, choice of preservation formats and standards, and choice of preservation strategies (such as normalization, migration, and emulation) for different types of collections and user communities.

Digital Archiving Tools and Technologies

Acquisition and ingest.

Enhancements to web harvesting tools could incorporate selection criteria such as indicators of preferred sites based on content or quality. This would allow libraries, archives, museums, and other collecting institutions to build more coherent and well-rounded collections. Methods are also needed to align the capture rate of web harvesting tools with regular or ad hoc update cycles. Many valuable resources are not readily available through web harvesters because of access restrictions, security and privacy concerns, or because of the structure of the underlying resources. These resources are governed by formal agreements between the content creator and the repository and managed through regular or periodic transfers to the repository. Regardless of whether collections are harvested or donated, new methods and tools are needed to automate the processing of items and collections when they are acquired or ingested into a repository. Tools are also needed to verify content, extract metadata, and automatically transform disparate types of objects into the standard formats and data models that a repository can manage.

Managing the evolution of tools, technology, standards, and metadata schemas.

Even if a repository transforms its collections into a small set of standard formats, time will change formats, standards, data models, and semantic representations. Needed tools include automatic format converters, emulators, and tools to support lossless and reversible transformations. Research is needed to develop conceptual models and methodologies for tracking data provenance and for managing the evolution of metadata schemas and ontologies.

Naming and authorization.

Methods and tools are needed for unique and persistent naming of digital objects in collections and repositories that may contain billions of entities. Persistent naming conventions are also needed for their associated data models, local and global file names, attributes, and the temporal, spatial and procedural relationships among digital entities. Interoperability among naming mechanisms is essential for repositories that acquire data from many different content providers.

Standards and interoperability.

Standard and stable ways to represent text, sound, image, video, and other object components would decrease the need for frequent transformations of digital entities that are necessary to forestall obsolescence. This would reduce the risk of introducing errors or losing information over the long term, and lower the costs of archival management. There is also a need for standards to represent temporal, procedural, spatial, and other relationships and methods for interoperability among different standards.

Organizational, Economic, and Policy Issues

Metrics.

Business planning, economic modeling, and evaluating repositories are hampered by the lack of metrics for almost all aspects of digital archiving. Metrics are needed to measure the costs, benefits, and values of digital objects, and to evaluate the cost, effectiveness, and performance of preservation strategies across different types of media and content. Research is also needed to determine the maximal sustainable archive size (as a function of the access rate), the bandwidth (amount of material that can be moved forward into the future as a function of the type of storage technology), and the re-purposing rate (the amount of time needed to process the entire collection to derive new collection attributes).

Economic and Business Models.

Digital preservation resembles public goods because the primary beneficiaries of current investments may be future generations. Repositories are also vulnerable to interruptions in funding that would threaten basic storage and data management, prevent ingest of new data and collections, and reduce or eliminate user access. In order to create an economy for digital preservation, economic models and funding mechanisms are needed that provide economic viability over very long periods of time; that create incentives for developing digital archives across different types of organizations; that stimulate private research and development of digital preservation tools, technologies, and services; and that provide incentives for content owners to deposit content in digital archives. Business models that drive down the costs of digital preservation would help to incentivize preservation activities. Research in this area is critical for the development of affordable technologies and services as well as deployment of a scalable infrastructure consisting of technical solutions, practical standards, and trusted institutions.

Implementing the Research Agenda

Workshop participants proposed a variety of research modalities from theory building to testbeds, and discussed a range of project types from small group or single investigator projects to very large teams of multidisciplinary researchers at several participating institutions. Existing digital collections needing preservation may form natural testbeds. The full report provides more details and guidance on the types of projects most needed and on the potential for partnerships between researchers and government agencies; between academic and industry researchers; and between researchers and private and not-for-profit stakeholders. Mechanisms are also needed to accelerate the transfer of new knowledge into practical working digital preservation systems to prevent further loss of valuable digital collections. There is a pressing requirement for education and training in new digital archiving methods, tools, and technologies.

Conclusion: It's about time

It's about time for a concerted research effort that focuses on the unique challenges of preserving digital information over long periods of time. Working to meet the challenges of long-term preservation will add new knowledge to computer science, information and archival science, and to social and behavioral science. Moreover, digital archiving research will have immediate societal benefits by preserving important digital resources that might otherwise be lost, producing more cost-effective and sustainable models that address current archiving needs, and creating business opportunities for new technologies and services. For future generations, our efforts will create the potential for new discoveries in science and medicine, for understanding the environment, for reflecting on the past, and for inspiring new forms of creative expression.

The Workshop on Research Challenges in Digital Archiving and Long-term Preservation

Recognizing repeated concerns over the current state of knowledge about long-term digital preservation, the National Science Foundation and the Library of Congress convened a workshop entitled “Research Challenges in Digital Archiving and Long-term Preservation” in April 2002. The main goals of the workshop were to identify the research challenges in digital archiving and long-term preservation; set priorities for research based on input from stakeholders; and propose mechanisms that could build a community of researchers and foster cross-fertilization among research projects. The workshop consisted of plenary presentations and discussions of the various challenges in digital archiving as well as small group sessions to define and set priorities for research. It also provided an opportunity for experts in computer science, mass storage systems, archival science, digital libraries, and information management to discuss obstacles to preserving digital information with government managers and other stakeholders. This report presents a summary of the workshop discussions and recommendations for future research projects.

It's About Time:

Research Challenges in Digital Archiving and Long-term Preservation

> *"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."*[1] Task Force on Archiving Digital Information, 1996

Significant advances in computer and information technology have provided our society with powerful tools for creating, organizing, and distributing digital resources, while simultaneously raising new and complex challenges for long-term preservation of digital information. The need to address the question of the longevity of digital information is becoming more urgent because repositories of digital information are surpassing physical archives in both scale and significance. The ability to preserve digital material is a serious challenge for government agencies, scientific data repositories, libraries, archives, museums and other cultural heritage organizations, and any organization that needs continuing access to its own information.

What is at stake?

Although concern over society's ability to preserve digital information for near-term, intermediate, and long-term future reuse is not new, the challenges of digital archiving and long-term preservation are pressing due to the increasing centrality of digital technologies and digital information to research, government administration, e-commerce,

[1] Garrett, J. and Waters, D. (1996), *Preserving Digital Information*: *Report of the Task Force on Archiving Digital Information,* Washington, D.C.: Commission on Preservation and Access: 7.

education, entertainment, and the arts.[2] Over the course of the last decade, government agencies, research libraries, corporations, and private individuals have accumulated vast quantities of digital information, much of which remains valuable into the foreseeable future. Increasingly, organizations of all types need to ensure that the valuable digital resources they create today will be available and understandable in the future.

Much more digital content is available and worth preserving.

Federal scientific agencies have invested billions of dollars to develop and deploy satellites, remote sensing devices, and other instruments that send terabytes of data daily to ground-based receiving stations for processing and analysis. Distinguishing long-term global climate change from natural variations requires detailed observations over long periods of time. In ecology, court records have been useful in establishing long-term changes in ecosystem types. In atmospheric chemistry, old stellar spectra have been used to establish changes in the chemical composition of the atmosphere. Recently, National Aeronautics and Space Administration investigators used a combination of data from current satellites and from satellite instruments launched in the early 1980s to discover important and unexpected anomalies in tropical radiation that were not expected by current models of atmospheric variability. In the future, even longer time series of Earth observations will be required to establish the true variability of this system—and of unexpected changes and cause-and-effect relationships that could not be exposed reliably without this long-term record.

In biomedical research, teams of investigators in universities and in the burgeoning bio-medical industry rely on access to massive databanks, such

[2] Several government-sponsored inquiries, numerous government-sponsored reports, and various advocacy efforts have drawn attention to the challenges of long-term digital preservation. See for example, United States Congress, House Committee on Government (1990), *Taking a Byte out of History: the Archival Preservation of Federal Computer Records*, Washington, D.C.: U.S. G.P.O.; National Academy of Public Administration (1991), *The Archives of the Future: Archival Strategies for the Treatment of Electronic Databases: a Study of Major Automated Databases Maintained by Agencies of the U.S. Government,* Washington, D.C.: The Academy; National Research Council (1995), *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 (2001), *LC21: A Digital Strategy for the Library of Congress,* Washington, D.C.: National Academy Press; and National Research Council (2003), *Building an Electronic Records Archive at the National Archives and Records Administration*, Washington, D.C.: National Academy Press.

as the human genome databases, to identify the basis for myriad diseases and to accelerate the discovery of effective remedies or cures. Many of these scientific databases are created and accumulated through government investments and some are irreplaceable at any price. Raw data rarely is sufficient. Meaningful analysis also requires preservation of models, simulations, and visualization tools. In most research communities, results are published electronically. In many fields, researchers want not only the published results, but also the data on which they were based, so that the science can be validated, reproduced, and extended.

Digital collections are growing at a rate that outpaces our ability to manage and preserve them.

Agencies at all levels of government are burdened by both rapid growth in the amount and complexity of the data they collect and by rising demands from the public for access to digital resources. The National Oceanic and Atmospheric Administration manages rich archives of space, atmospheric, oceanographic, and weather data, but with intake of data doubling every year archiving and access capabilities cannot keep up.[3] Similar concerns are echoed in statistical agencies, such as the Census Bureau, the Bureau of Labor Statistics, and the National Center for Health Statistics, that build longitudinal data sets that researchers exploit to understand trends in employment, wages, pubic health, housing, and education, or to analyze the effects of government policies and large public investments in improving health and social welfare. State and local government agencies maintain repositories of digital data such as voting registrations; deeds, easements, and other property records; and policies, ordinances, and permits to ensure citizens’ rights, protect private and public property, and manage public assets. Private firms also collect and use digital data as part of their core business mission, such as for the development of new drugs by pharmaceutical companies; for exploration of resources by oil, gas, water, and mining companies; and for market and sales analysis by retailing companies. Digital video, sound, animation, and accompanying text are

[3] Department of Commerce, National Oceanic and Atmospheric Administration (August 2001), *The Nation’s Environmental Data: Treasures at Risk. Report to Congress on the Status and Challenges of NOAA’s Environmental Data Systems,* Executive Summary, Washington, D.C.: U.S. Department of Commerce: 3.

major assets of the entertainment industry as well as important cultural assets to the public.

Many types of organizations and even private citizens face digital preservation challenges.

Digital resources are no longer limited to scholarly use of specialized databases. E-government and e-commerce are built on electronic transactions stored in databases that have to be maintained at least long enough to satisfy auditing, taxation, compliance monitoring, and other accountability requirements. Medical records that track individual medical histories are becoming digital. Cultural heritage resources and vital evidence of human creativity in the digital age is produced, shared, and enjoyed in digital form primarily through the World Wide Web. Even private citizens are accumulating valuable digital resources, from e-mail with friends and family to digital photographs. In fact, few sectors of society are not touched in some way by the need to enhance the longevity of digital resources and reduce the risk of catastrophic loss.

Background and Current Situation

Cultural institutions that embrace preservation as a core part of their mission face digital archiving challenges. Congress has asked the Library of Congress to lead the effort to develop a National Digital Information Infrastructure and Preservation Program.[4] The National Archives and Records Administration is planning to develop and implement the Electronic Records Archives[5] to preserve the permanently valuable records of the federal government. Research libraries and publishers are grappling with the challenges of preserving electronic journals, e-books, and online collections. A rapidly increasing portion of the nation’s published record, historical documentation, scholarly communication, and creative expression exists only in digital form. Solutions to the challenge of digital archiving

[4] See http://www.digitalpreservation.gov/ndiipp/

[5] See http://www.archives.gov/electronic_records_archives/index.html

and long-term preservation are critical to the future of cultural institutions that serve as stewards and custodians of intellectual and cultural heritage.

The process of research and the prospects for major breakthrough in scientific discovery are being transformed by ready access to digital data. Rapid advances in scientific research depend on continuing access to vast data repositories along with tools for analysis. A recent report from the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure[6] envisions revolutionary changes in science and engineering research, based largely on the ability of researchers to share and combine raw data from many sources and to utilize powerful tools for analysis, visualization, and simulation. To realize this vision, scientists will need hundreds of highly curated digital repositories capable of storing massive data streams from satellites and other instruments. The new generation of repositories will need to be maintained indefinitely into the future to protect very large investments in the initial collection and processing of data and to enable continuous data mining and repurposing. With investments in remote instruments, high performance computing, collaboration tools, and data grids, scientists are poised to accelerate the pace of data gathering and discovery by orders of magnitude.

Archives need to become the repositories of intellectual capital. This emphasizes the view of archives as information and knowledge repositories. The goal of an archive is to make information and knowledge content as readily accessible as possible, and to make it easy to repurpose collections for new uses. Examples where the preservation of intellectual capital is already important include universities and research laboratories. Every group or organization needs to think of persistent archives as the mechanism to preserve information and knowledge for use by current and future generations.

[6] http://www.cise.nsf.gov/evnt/reports/toc.htm

What are the unique research challenges in digital archiving?

Preserving digital information is much more difficult than preserving traditional paper, film, and audio-video information. First and foremost, the long-term perspective raises distinctive challenges. Digital archives aim to preserve data for decades, centuries, or even longer. Yet the storage media, input and output devices, programming languages, software applications, and standards that are necessary to retrieve and interpret digital information are revised and replaced every few years. This is why a recent reference model for archival systems defined **long-term** as:

> A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future.[7]

Given today’s rapid innovations in information technology, digital information created today may be inaccessible or impossible to interpret a few years from now.

Physical preservation of digital information can be managed by copying a bit stream from one medium to another. The more challenging problem is “logical” preservation – the ability to reconstruct streams of bits in a meaningful way that computers and humans can interpret, use, repurpose, and understand at any arbitrary point in the future. Logical preservation is a difficult problem because people and organizations are using increasingly complex software tools, data models, semantics, and concepts to capture, represent, display, and analyze many different types of information. Formats, standards, software, and semantics evolve at different rates for different types of digital information, adding to the complexity of designing and selecting effective long-term preservation strategies.

[7] Consultative Committee for Space Data Systems (January 2002), CCSDS 650.0-B-1: *Reference Model for an Open Archival Information System (OAIS),* Blue Book, Issue 1, Washington, D.C.:, CCSDS Secretariat: 1-11.

When long-term preservation spans several decades, generations, or centuries, the threat of interrupted management of digital objects becomes critical. Unlike many physical objects that can withstand some period of neglect without resulting in total loss, digital objects require constant maintenance and elaborate "life-support" systems to remain viable. Redundancy, replication, and security against intentional attacks on archival systems and against technological failures are critical requirements for long-term preservation, as are issues of forward migration. The challenges of maintaining digital archives over long periods of time are as much social and institutional as technological. Even the most ideal technological solutions will require management and support from institutions that go through changes in direction, purpose, and funding.

Digital preservation challenges are also unique because of the scale and diversity of information that has long-term value. Digital objects worthy of preservation include databases, documents, sound and video recordings, images, and dynamic multi-media productions. These entities are created on many different types of media and stored in a wide variety of formats. Despite a steady drop in storage costs, the recent influx of digital information and its growing complexity exceeds the archiving capacity of most organizations. This is largely due to the fact that digital archiving and long-term preservation entail much more than storing large quantities of raw bits of data. Digital collections require curation and processing to ensure their longevity, protect their integrity, and enhance their value for use in the future. Current preservation methodologies require intensive human intervention that is not affordable in the long run and that will not scale to the massive size of many digital collections. Moreover, there are no well-developed methodologies for preserving many of today's complex data types and formats.

The development of infrastructure for digital archiving is strongly driven by the need to support multiple communities. Each community has its own requirements that will influence the content, organization, design, and services of digital archives. Given the wide variety of digital materials that need to be preserved and the different stakeholders in preservation processes and outcomes, it is apparent that no single approach will address all digital preservation challenges. Rather, research is needed on a spectrum of solutions ranging from tools that private citizens can use to preserve their digital photograph collections to large repository systems for scientific databases, imagery, and recorded sound. To build an infrastructure that will make long-term preservation of digital collections affordable and sustainable, we also need to identify concepts, methodologies, and technologies that can satisfy common requirements for storage, management, and access.

Research Agenda for Digital Archiving and Long-term Preservation

Workshop participants discussed research challenges in four focus areas: 1) Technical Architectures for Archival Repositories; 2) Attributes of Archival Collections; 3) Digital Archiving Tools and Technologies, and 4) Organizational, Economic, and Policy Issues.

Each of these discussions yielded important research questions. Research challenges in each area are summarized below, followed by a description of some cross-cutting issues.

Focus Area 1: Technical Architectures for Archival Repositories

Organizations with digital preservation needs are attempting to design, build, and procure repositories that will preserve digital information even though the components of the repository will evolve over time. No persistent repository systems exist today, although there are useful models, design principles, and components.[8] Much more research is needed to define technical architectures (blueprints) for persistent archives, develop processes for long-term management, build tools that can be used to acquire archival data, prepare data for long-term storage, and manage data over several generations of technology.

[8] Examples of useful models include *Reference Model for an Open Archival Information System (OAIS);* Shirky, C. (2002), Appendix 9, *Preliminary Architecture Proposal for Long-Term Digital Preservation,* in *Preserving our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*, Washington, D.C.: Library of Congress; and Research Libraries Group (2002), *Trusted Digital Repositories: Attributes and Responsibilities*: *An RLG-OCLC Report,* Mountain View, CA: RLG. While useful, all of these models require further elaboration and evaluation.

Developing technical architectures for persistent repositories is challenging because of the differing size and diversity of digital collections. Scientific data repositories typically store very large collections of observational or experimental data in a small number of formats or data types. Much of the development of repository models has been oriented to large and relatively homogenous data sets. Even with relatively homogeneous data, it is unclear whether and how persistent archives can be designed to preserve the very large databases of thousands of terabytes or a few petabytes. Cultural institutions, such as the Library of Congress and the National Archives and Records Administration, will need technical architectures that can manage and preserve tens of thousands of heterogeneous collections, some of which will be very large. Private citizens will need either tools that are inexpensive and easy to use or access to affordable archiving services.

Experts in digital archiving accept the notion that the physical storage of digital data should be separated from the logical representation of collections in a repository. Repositories can then build services on top of a persistent storage architecture to accommodate the need of various user communities and satisfy the requirements of different types of data. Therefore, it is important to characterize a spectrum of digital repository models and their associated research issues.

Data model driven architecture—This model is used to preserve a specific type of data for future reuse. This is the simplest type of repository to implement because it preserves one type of data and requires support for one data model. At a minimum, the archive must guarantee the preservation of data into the future.

Research Issues

- capacity and scalability of multiple petabyte repositories
- management of relationships among entities in the repository
- effective means for validation of the content
- management of collections across an evolving software infrastructure
- effective means for refreshing data on aging media
- methods for replication of collections or entire repositories regardless of size
- methods for automated acquisition, quality control, and description

Archives of derived data products—Many archives store data products that are the result of further processing of the original raw data. An example of this type of archive is the remote sensor data collected by the National Aeronautics and Space Administration. Processing is applied to the original sensor data to create multiple levels of derived products, including application of calibration mechanisms, conversion from sensor data to physical quantities, and transformation to alternate coordinate systems. Likewise, statistical agencies process raw counts to adjust for sample size and weighting. This raises the challenge of tracking the processes that were applied to the original data to produce the derived products so that the results can be validated and replicated.

Research Issues

- development of mechanisms for tracking provenance information and for describing the derived data products in relation to the original data
- methods for preserving the processing algorithms, the mathematical expressions of the related operations, and associated software
- methods for describing each additional derivation and its relationship to previous derived products and to the original data
- development of methods to automate the management of such repositories for purposes of scalability and affordability

Archives of temporally changing data—These archives preserve data that is continuously changing, either through regular additions that are streamed into the archive, through ad hoc actions taken by content creators, or in conjunction with workflow processes. Archives

of temporally changing data will have to characterize and maintain the temporal and procedural relationships between the multiple versions of digital objects. A related problem is the management of a data collection that can change over time, either through extension by addition of new entries or through revision by creation of new versions.

Research Issues

- definitions, methodologies, and tools for time-based capture and representation of digital objects and collections
- methods for taking useful snapshots of dynamic databases
- identification of models and methods to represent temporal and procedural relationships
- methods to generate the required technical and descriptive metadata at a rate faster than the archive update rate

Archives of evolving data—Preservation and management of many types of digital information requires transformation of the original data to new formats or canonical forms. Archives may convert data to simplify the ability to present or render it in the future. This requires processing the data. To guarantee authenticity, archives that process data must be able to document all processing operations. The procedural knowledge inherent in the processing steps must be preserved, along with the versions of the data. Thus version management is a required capability of this type of persistent archive.

Research Issues

- definition and characterization of transformation processes so that they can be automated
- methods to document transformations to original data
- methods and tools for reversible transformations

Controlled access repositories—Many repositories preserve data under access restrictions due to intellectual property rights, privacy and confidentiality, or national security classification. Rules governing access may inhere in the item being requested, may result from a relationship between the item and the requestor, or be a combination of the two.

Controlled access repositories must manage restrictions over time periods that exceed the lifetime of individuals and even of organizations. Research questions derive from stringent requirements for auditability, authentication, and complex rules governing access.

Research Issues

- approaches to the persistent management of authentication over long periods of time
- methods to reconcile access restrictions and permissions using complex rule bases

Repurposed archives—In the most general sense, every access to an archive is a form of repurposing, as it entails a unique reinterpretation of the collection by the user. The extent to which the user succeeds in this effort is dependent upon the availability and quality of descriptive information (metadata) for the collection. Over time, users will need to interpret metadata in different ways, including remappings to new schemas, ontologies, and concept spaces.

This spectrum of interpretive possibilities illustrates the need for research that more closely examines the relationship between the purpose of the archive, the types of data and information that it acquires, and the needs of its user community.

Research Issues

- management of schemas, ontologies, and concept spaces over time
- methods to utilize descriptive metadata to a new semantic context
- tools to map from an existing concept space to a new concept space

Repository architectures—Repository architectures also span a spectrum from highly distributed layered models built on data grid technology to self-contained repositories that provide end-to-end services.[9] In the layered

[9] Data grids provide the interoperability mechanisms needed to manage data across heterogeneous storage and information repositories. Applications of data grid technology are being made for the National Archives and Records Administration (prototype persistent archive) and the National Science Foundation (discipline specific data grids in astronomy, earth system science, education curricula).

model, different organizations might specialize in particular preservation functions. A library or an archive, for example, might contract with a service provider for physical storage and data management but take responsibility for selection, organization, and end-user services. On the other hand, a single entity might maintain a self-contained and relatively autonomous repository that supports all archival functions from basic physical storage and maintenance of bits to end-user services. Such a model might be appropriate for preservation of highly specialized collections that serve one particular community.

Research Issues

- focused research at each layer of the architecture (physical storage, data management, logical representation, services)
- methods and protocols that enable the different layers to interoperate
- methods that allow heterogeneous distributed repositories to exchange content and services

Focus Area 2: Attributes of Archival Collections

We know that a great deal of information is “saved” on file servers, on personal hard drives, CD-ROMs, and DVDs, and in large repositories of spinning disks and tapes. Archived collections have additional attributes that enhance their quality, trustworthiness, interpretability, and longevity. Archival collections do not just happen when someone clicks on the “save” icon. Rather, archival collections are created through controlled curatorial processes that include selection, organization, description, quality control, and perpetual care. They require that individuals or organizational entities assume a formal role of stewardship. Just as the development of infrastructure for digital archiving is strongly driven by the need to support multiple communities, it is also strongly driven by the requirements to preserve many diverse types of complex objects and collections.

Research topics span a range of conceptual, policy, and technological issues. Specific research issues may revolve around the particular characteristics of different types of collections, such as electronic journals, multi-media entertainment objects, scientific databases, or models and simulations. Common concerns, such as metadata requirements, authenticity, and process and decision models, cut across many different types of objects, although they may be addressed through means that are specific to a particular type of object or collection.

Preservation of static digital objects—Although not ideal, methods exist today to preserve simple, static digital objects in widely used formats. These methods include reformatting data into standard formats when it is ingested into a repository, periodically transforming digital entities from obsolete to current formats (migration), and using emulation to run obsolete application software. Even simple static objects present research issues related to scale and affordability.

Research Issues

- developing tools to handle the processes of reformatting, migration, and normalization with minimal human intervention
- methods to evaluate the effectiveness of different formats, data models, and metadata schemas from a preservation perspective
- methods for automatic error detection and correction during reformatting processes

Preservation of complex digital objects—The next challenge is to develop means for preserving complex objects and dynamic objects that change on a regular basis. Complex objects contain core content, but they also have additional features such as formatting, visual aspects and graphics, monochrome and color images, and sound and video. Interactive documents provide users with the opportunity to set preferences or make choices that determine how an object appears or behaves. Dynamic objects and streaming data are updated continually or on an arbitrary basis. Acceptable methods are likely to vary, depending on the types of archived data and its anticipated use.

Research Issues

- identification of which aspects of complex objects are worth preserving and how different features affect the authenticity, utility, and aesthetics of objects for particular user communities.[10]
- definitions of acceptable levels of information loss
- methods to preserve software and to run executables on future computing platforms
- methodologies to "fix" a view of an object that is frequently changing
- criteria for determining the number of versions necessary to preserve a meaningful sense of an object's evolution over time
- methods, tools, and technologies that can manage complex objects over time

Aggregation of items and objects into collections—Physical collections are made up of discrete items, but this is not the case with digital collections. A digital collection can be created simply by making hyperlinks among widely distributed objects. This presents many challenges for archiving because different organizational entities control the content, format, quality, and availability of linked materials. With the emerging need to capture materials from the web for aggregation into collections, definitions are needed to establish boundaries around collections and to determine how many levels of linked items need to be preserved.

Research Issues

- methods to establish boundaries around collections that gives them coherence and an identity
- processes for developing and sustaining distributed collections where several organizations participate in the curatorial processes and share preservation responsibilities
- development of collection-level metadata schemas that describe attributes common to all items in a collection and provide for inheritance of metadata from the collection to the item level
- means to formally express the attributes of collections so that higher-level services can operate with them

[10] Note that preserving the "look and feel" of complex objects often requires preserving software.

Decision models—Long-term preservation does not imply that everything is worth saving even if the cost of storage is minimal. Selection decisions are not based on an arbitrary notion of what is valuable. Most libraries, archives, and museums have well established collecting policies for physical items based on the mission of the institution and knowledge of its user community.

In the digital realm, selection decisions are more complex. Organizations building digital collections must be concerned not only with what content to preserve, but also with which features and functionality users will need. In libraries, selection decisions are changing because an increasing amount of the content libraries deliver to users is physically held in publisher repositories. This raises concerns over who should assume responsibility for long-term preservation (publishers or libraries) and when, if ever, the obligations to acquire and preserve published material should shift from the content providers to a library or an archive. Furthermore, collecting policies that were designed for physical materials do not encompass new types of digital objects and collections such as web sites and multi-media objects.

Decision models are needed to support many aspects of digital preservation. Research is needed on cost/benefit models to support decisions about which preservation formats and standards to use, which preservation strategies (such as normalization, migration, or emulation) are most effective, and feasibility and benefits of extracting or adding metadata.

Research Issues

- formal models for selection decisions to automate aspects of the selection process
- cost and benefit analysis of selection at a fine level of granularity versus saving everything from a particular content creator, research process, or organization
- metrics for measuring the quality and fidelity of preserved digital objects

Focus Area 3: Digital Archiving Tools and Technologies

Although many useful information management tools exist, preservation is different because of its long-term time scale. Digital repositories will have to manage the evolution of tools, technologies, standards, and metadata schemas over time. Tools to automate many aspects of the preservation process are needed because human costs are the most expensive element of archiving and they are likely to increase while processing and storage costs decline. Moreover, current methods that require extensive human intervention do not scale to the size, diversity, or complexity of the digital assets being generated today.

Acquisition and ingest—Many powerful tools exist to locate, harvest and copy digital information for preservation. Internet search engines and web archiving projects acquire material available on the World Wide Web by using automated robots and web crawlers. With enhancements, this basic technology could be tuned to support acquisition and future preservation of vast digital resources. But many valuable resources are not readily available because of access restrictions, security and privacy concerns, or because of the structure of the underlying resources. Preservation of such resources often must be handled through formal agreements where the content creator transfers responsibility for preservation to a repository.

Research Issues

- methods to incorporate selection criteria into harvesting tools such as indicators of preferred sites based on content or quality
- methods to align the rate of capture with regular or ad hoc update cycles
- methods to automate the processing of items and collections for preservation at the point that they are acquired or ingested into a repository such as tools for automated indexing, metadata extraction, validation, and quality control
- tools to automatically transform disparate types of objects into the formats, standards, and data models that a repository can manage over the long-term
- tools to document these transformations automatically

Naming and authorization—Managing the identity of preserved digital objects over time is a challenge for digital archives because the identifiers assigned to digital objects can be changed easily and the technologies for naming and tracking digital objects evolve. Moreover, persistent naming conventions are needed to describe digital entities ranging from the components of the data model, to local file names for the digital entity, to global file names used to assemble distributed collections, to attribute names used to build collection catalogs, and to relationship names used to describe properties of the collection.

Research issues

- methods for unique and persistent naming of archived digital objects in collections and repositories that may contain billions of entities
- persistent naming conventions for data models, local and global file names, attributes, and relationships
- tools for automatic certification and authentication of preserved digital objects
- version control methods
- interoperability among naming mechanisms used by different content providers

Standards and interoperability—Standards for data formats, data models, metadata, and many other aspects of digital information are useful for long-term preservation. However, standards change over time and digital repositories are likely to contain objects that conform to many different standards. Archived digital entities will have to be migrated to new standards in the future. The migration can be viewed as lossless if the new standard provides a superset of the features of the old standard. Unfortunately, this is not always true. A goal for standard encoding formats is the creation of lossless feature conversions when migrating between standards. The new standard needs to provide a way to encode the operations and features of the prior standard.

Research Issues

- standard and stable ways to represent text, sound, image, video, and other object components
- standard ways to represent temporal, procedural, spatial, and other relationships
- methods for interoperability among different standards
- predictors for which digital preservation standards are likely to achieve wide scale adoption over extended periods of time

Focus Area 4: Organizational, Economic, and Policy Issues

The ability to preserve digital resources for the long-term requires a deep infrastructure of technical solutions, standards, trusted institutions, affordable business models, and skilled personnel. Without affordable business and economic models, even the most effective technical solutions are likely to fail. Much digital preservation research has concentrated on technical problems and technological solutions without careful analysis of the social, organizational, and economic mechanisms that have to be in place to make preservation possible and sustainable. Digital preservation shares certain properties with other public goods such as national defense or public parks. Specifically, there are few market incentives to create public goods because they tend to be costly and it is difficult to exclude anyone from their benefits. Creating an economy for long-term preservation entails providing incentives for organizations to invest in digital archives, even though some of the benefits of investments made today may not be realized for decades.

The public goods aspect of preservation is also present in physical library, archive, and museum collections, but digital information raises a number of new concerns. The cost components of digital preservation differ from physical preservation in a number of ways. The life of many physical works can be extended with preservation treatments, transfer to stable media, and well-controlled storage environments. Digital information, however, requires ongoing perpetual maintenance. The viability of digital collections is threatened even by very brief interruptions in management or funding.

Opportunities abound for research on economic and business models for sustainable long-term repositories and on policies for distributing the costs and responsibility for digital preservation. Institutions with long-term preservation responsibilities need economic and business models to guide collection development, choose preservation strategies, define levels of service, and develop revenue streams sufficient to sustain a digital archive over very long time periods. Absent well-developed technology and tools for digital preservation, there is little empirical data on the costs and benefits of digital archiving.

Metrics—Business planning and economic modeling are hampered by the lack of metrics for almost all aspects of digital preservation. Evaluation of digital archiving is impossible without concrete measures of costs, benefits, and values of digital objects.

Research Issues

- metrics to assess the cost and effectiveness of preservation strategies for different types of content
- methods to measure the performance of various storage media over very long periods of time
- metrics to assess preservation functionality such as the maximal sustainable archive size (as a function of the access rate), the bandwidth (amount of material that can be moved forward into the future as a function of the type of storage technology), and the re-purposing rate (the amount of time needed to process the entire collection to derive new collection attributes)

Incentives for long-term preservation of digital information—The incentives for developing digital archives are likely to vary across different types of organizations and between public and private sector organizations. National and state archives are mandated by law to preserve permanently valuable government records. Research libraries and cultural institutions embrace preservation as part of their core mission. Scientific agencies need to preserve data to support research. Private corporations will need digital archives to maintain records for business purposes, to comply with auditing and accounting requirements, or to protect digital assets such as sound recordings, animation, and moving images. All of these organizations have a vested interest in tools and technologies that reduce long-term preservation costs. At the same time, private firms in the information technology sector may find a market for hardware, software and services that will support digital archiving. Ideally, incentives could be designed to encourage organizations to create optimally preservable content, develop their own archiving capabilities, provide archiving services, and build repositories.

Research Issues

- evaluation of alternative revenue streams (public subsidies, tax incentives, philanthropic contributions, endowments, fees for service, etc.)
- estimates of the size of the market for digital archiving systems and services
- market analysis of user demand for preserved digital content

Incentives for deposit of digital content in archives—Content creators need incentives to deposit content in repositories for long-term preservation. Depositors must have a very high level of confidence that a digital repository will preserve content indefinitely, will not introduce errors, and will apply rights management, confidentiality, and other access rules consistently. Research in this area is closely tied to the concept of trust.

Research issues

- identifying the minimum level of performance that depositors are willing to accept from a digital repository
- models for tax incentives or other rewards that would encourage owners of intellectual property to place content in the pubic domain prior to the expiration of copyrights

Implementing the Research Agenda

Most digital archiving research to date can be characterized as a combination of small stand-alone projects, projects to resolve immediate operational problems, and projects that were tacked onto larger research initiatives. We have reached a point where an effort focused on digital preservation is urgently needed. The effort will need to engage a sufficient number of researchers, involve government agencies and other partners with substantial digital archiving needs, and mobilize an appropriate level of investment. We anticipate that a minimum investment of $5 to $8 million per year is needed for a focused research program for the next ten years.

The ten-year time frame is essential, not only because of the complexity of the problem, but also because of the considerable time required to implement, evaluate, and test the results of research. A ten-year program would also provide a foundation for evaluating digital preservation strategies over two or three generations of computer and information technologies. We recommend that the National Science Foundation and the Library of Congress lead this initiative. These two institutions should encourage sponsorship from other government agencies, private foundations, content providers, and industry, as well as participate in active partnerships with researchers from many disciplines.

Research on digital archiving and long-term preservation is amenable to many innovative approaches. Possible research methodologies cover a whole spectrum from small, single investigator projects to testbeds involving many researchers and several participating institutions. Organizations of all sorts have accumulated large digital collections that can serve as natural testbeds. Moreover, digital archiving research may have immediate societal benefits by preserving important digital resources that might otherwise be lost, producing more cost-effective and sustainable models that address current archiving needs, and creating business opportunities for new technologies and services.

Research Scenarios

The examples below suggest the types of projects most needed.

Theory-building

- develop concepts of value, aesthetics, experience, and behavior of digital objects
- develop new theories about authenticity of digital information in a rapidly changing technological and legal context

This type of project would involve an individual investigator or small group.

Exploratory

- propose and test alternative architectures and preservation methods.[11]

Small inter-disciplinary teams could carry out this type of project.

Simulations and experiments

- simulate different policy and economic models, and compare results
- apply different methods to the same content, and compare results
- develop prototype tools for automating archiving processes and evaluate their effectiveness

Small to large teams would be required for this type of project.

Observational

- observe user behavior and perception of the same content presented in different formats in order to determine acceptable levels of information loss

An individual investigator or small group could conduct this type of project.

Testbeds

- use existing digital content in government agencies, digital libraries, and archives to develop metrics, and compare the results of different preservation methods

[11] Note that normalization, migration, and emulation are the most commonly used methods. None of these solutions is ideal, and there is a need to explore other alternatives.

- use very large collections to test the effectiveness and scalability of digital preservation tools and technologies
- use repositories with large numbers of heterogeneous digital collections to test functionality and scalability of metadata extraction tools, naming conventions, and automatic ingest process.

This type of project requires large, multi-disciplinary teams and several participating organizations.

Partnerships

Many opportunities exist for partnerships between researchers and organizations of all sorts that hold significant digital collections and face pressing digital preservation needs. Because the research challenges are complex, multidisciplinary teams are needed spanning computer science and engineering, information and archival science, economics, and social and behavior sciences. The workshop identified many opportunities for partnerships between researchers and the creators and custodians of government information in scientific agencies, defense and intelligence agencies, statistical agencies, the National Archives and Records Administration, the Library of Congress, and the National Libraries of Medicine and Agriculture. Funding for government/academic partnerships is available through the National Science Foundation Digital Government Program and Digital Libraries Program.

The National Science Foundation and Library of Congress recently entered into a strategic partnership to support research that will contribute to development of a national infrastructure for sharing and preserving digital information. Digital libraries, private firms with large corporate databases, and cultural heritage organizations are other potential partners for research projects that will address pressing preservation needs. Joint ventures between academic researchers and researchers in the information technology sector are also encouraged. Development of a digital preservation infrastructure will also benefit from related research on storage media and systems, data management, knowledge management, semantic interoperability, and information retrieval.

Centers for Digital Archiving Research

There may be benefits in one or more centers for digital archiving and long-term preservation research serving as focal points for this effort and addressing issues of technology and knowledge transfer, education and training, and capacity building.

Technology and Knowledge Transfer: From Research to Practice

The need for research on long-term preservation is driven by current pragmatic concerns about the longevity of digital information. There should be a dual focus on striving to preserve valuable digital information while also “learning by doing.” Development of more effective preservation methods and technologies will demand many years of research. But, as one participant said, “if research never ends, then archiving never begins.”

Mechanisms to accelerate the transfer of new knowledge into practical working digital preservation systems will require greater awareness and interaction between the research community and government agencies and other stakeholders. There is also a need to disseminate what is already known to avoid duplication of efforts or replicating approaches that failed. Metrics to assess improvements would be especially helpful in tracking progress and developing benchmarks. It may also be beneficial to identify a range of research projects aimed at addressing immediate, near-term, and long-term problems.

Digital archiving and long-term preservation will radically change the knowledge base and skill sets required of people who manage and operate digital repositories. It is already creating a demand for people with a mix of technical, engineering, information science, managerial, and domain expertise. There is a pressing need to re-train current personnel in the use of new process, systems, and technology; to educate the next generation of digital archivists; and to identify and train computer scientists and engineers who can design and build the tools and technologies needed to create persistent archives.

Conclusion: It’s About Time

It’s about time to launch a new research initiative. This initiative will advance computer and information science, archival and library methods, and social and organizational models in connection with digital information. This initiative will address a critical need for reliable, sustainable, and cost effective means to manage digital information essential for discovery of new knowledge. Perhaps most significantly, this initiative will build a foundation for digital preservation practices that government agencies, cultural institutions, businesses, and others urgently require.

Appendix I: Participants

MARTHA ANDERSON, Library of Congress

BRUCE R. BARKSTROM, National Aeronautics and Space Administration

MICK BASS, Hewlett-Packard Company

NEIL BEAGRIE, Joint Information Systems Committee, UK

LAWRENCE BRANDT, National Science Foundation

PETER BUNEMAN, University of Edinburgh and University of Pennsylvania

LAURA CAMPBELL, Library of Congress

ARTURO CRESPO, Stanford University

ROBIN DALE, Research Libraries Group

JON EISENBERG, National Academies, Computer Science and Telecommunications Board

DALE FLECKER, Harvard University

CARL FLEISCHHAUER, Library of Congress

EVELYN FRANGAKIS, National Agricultural Library

AMY FRIEDLANDER, Council on Library and Information Resources

ANNE GILLILAND-SWETLAND, University of California, Los Angeles

JIM GRAY, Microsoft

DANIEL GREENSTEIN, Digital Library Federation

VALERIE GREGG, National Science Foundation

STEPHEN M. GRIFFIN, National Science Foundation

MYRON P. GUTMANN, University of Michigan, Ann Arbor

RICH HARADA, High Density Storage Association and Creative Businesses, Inc.

MARGARET HEDSTROM, University of Michigan, Ann Arbor

ROBERT HORTON, Minnesota State Historical Society

BERNIE HURLEY, University of California, Berkeley

CARL LAGOZE, Cornell University

BRIAN LAVOIE, OCLC

CAL LEE, University of Michigan, Ann Arbor

RAYMOND LORIE, IBM Almaden

CLIFFORD LYNCH, Coalition for Networked Information

PETROS MANIATIS, Stanford University

VICTOR MCCRARY, National Institute of Standards and Technology

Appendix I: Participants (continued)

ALEXA T. MCCRAY, National Library of Medicine
NANCY MCGOVERN, Cornell University
KURT MOLHOLM, Defense Technical Information Center
REAGAN MOORE, San Diego Supercomputer Center
DOUGLAS OARD, University of Maryland
CHRISTOPHER OLSEN, Central Intelligence Agency
ARCOT K. RAJASEKAR, San Diego Supercomputer Center
DAVID ROSENTHAL, Sun Microsystems
JEFF ROTHENBERG, RAND
CHARLES ROTHWELL, National Center for Health Statistics
ED SEQUEIRA, National Library of Medicine
ABBY SMITH, Council on Library and Information Resources
MACKENZIE SMITH, Massachusetts Institute of Technology
THORNTON STAPLES, University of Virginia
SUE STENDEBACH, National Science Foundation
KENNETH THIBODEAU, National Archives and Records Administration
HERBERT VAN DE SOMPEL, Los Alamos National Laboratory
HOWARD D. WACTLAR, Carnegie Mellon University
DONALD J. WATERS, Andrew W. Mellon Foundation
ED H. ZWANEVELD, National Film Board of Canada

Appendix II: Workshop Agenda

Thursday, April 11

5:00 p.m.	Arrival of Out of Town Participants/Registration, **Airlie House front desk**
6:30 p.m.- 8:00 p.m.	Buffet Dinner, **Airlie Room**

Friday, April 12

8:00 a.m. - 9:00 a.m.	Breakfast, **Airlie Room**
8:00 a.m. - 9:00 a.m.	Registration, **Airlie House front desk**
9:00 a.m. - 10:30 a.m.	Opening Plenary Session, **Federal Room**
	Welcome from our Sponsors Larry Brandt, National Science Foundation, Digital Government Program Steve Griffin, National Science Foundation, Digital Libraries Program Laura Campbell, Library of Congress
	Why are we here? Overview of Goals and Objectives Margaret Hedstrom, University of Michigan
	The Bit Stream Interpretation Problem Raymond Lorie, IBM Almaden Research
	Introduction to Morning Discussion Topics Topic A: *Architecture for Repositories,* Reagan Moore, San Diego Supercomputer Center Topic B: *Attributes of Archival Collections*, Margaret Hedstrom
10:30 a.m. - 10:45 a.m.	Break
10:45 a.m. - 11:45 a.m.	*Small Group Discussions* Topic A: *Architecture for Repositories* — Group 1, **North Room** and Group 2, **South Room** Topic B: *Attributes of Archival Collections* — Group 3, **West Room** and Group 4**, Foxes' Den**
11:45 a.m. - 12:30 p.m.	Report out from small groups, **Federal Room** Reaction — Open Discussion
12:30 p.m.- 1:30 p.m.	Lunch, **Airlie Room**
1:30 p.m.- 2:00 p.m.	Reconvene Plenary Session, **Federal Room** *Introduction of Topics for Afternoon Discussion* Topic C: *Policy and Economic Models*, Don Waters, Andrew W. Mellon Foundation Topic D: *Tools and Technology*, Jim Gray, Microsoft Research

Appendix II: Workshop Agenda (continued)

Friday, April 12

2:00 p.m.- 3:15 p.m..	*Small Group Discussions* Topic C: *Policy and Economic Models* — Group 1, **North Room** and Group 2, **South Room** Topic D: *Tools and Technology* — Group 3, **West Room** and Group 4**, Foxes' Den**
3:15 p.m. - 3:45 p.m.	Break
3:45 p.m. - 4:45 p.m	Report out from Small Groups/Summary/Issues, **Federal Room**
4:45 p.m. - 5:00 p.m.	Summary/Plans for Saturday, **Federal Room**
5:00 p.m.	Close
5:30 p.m. - 7:00 p.m.	Posters and Informal Presentations, **Federal Room**
6:00 p.m. - 7:00 p.m.	Reception with cash bar, **East Room**
7:00 p.m. - 8:00 p.m.	Workshop Dinner, **East Room**

Saturday, April 13

8:00 a.m. - 9:00 a.m.	Breakfast, **Airlie Room**
9:00 a.m. - 10:15 a.m.	Plenary Session, **Federal Room**
	Reactions to the First Day Clifford Lynch, Coalition for Networked Information Reaction — Open Discussion *Instructions for Small Group Discussions* Margaret Hedstrom
10:30 a.m. - 10:45 a.m.	Break
10:45 a.m. - 11:45 a.m.	*Small Group Discussions* *What are the current priorities for research?* — Group 1, **North Room** *How might future scenarios impact research priorities?* — Group 2, **South Room** *What constitutes an infrastructure for long-term preservation?* — Group 3, **West Room** *How can research be transferred into practical applications?* — Group 4**, Foxes' Den**
11:45 a.m. - 12:30 p.m.	Reports from small groups and discussion, **Federal Room** Reaction — Open Discussion
12:30 p.m.- 1:30 p.m.	Lunch, **Airlie Room**
1:30 p.m. - 3:00 p.m.	Final Working Session, **Federal Room** Plenary discussion and Recommendations
3:00 p.m. - 3:30 p.m.	Next Steps, Organizing Committee, **Federal Room**
3:30 p.m.	Adjourn