PRESERVING OUR DIGITAL HERITAGE

Plan for the National Digital Information Infrastructure and Preservation Program

A Collaborative Initiative of the Library of Congress

APPENDICES

First Printing, October 2002

Contents

Preface

The appendices for the National Digital Information Infrastructure and Preservation Program (NDIIPP) Plan provide important background and supplementary materials. The appendices are diverse, but together illustrate the planning process and provide a rationale and justification for the Plan's recommendations.

Appendix 1: Consultation with Concerned Stakeholder Communities identifies the diverse group of experts and stakeholder communities that the Library of Congress consulted in the development of the plan. These experts and stakeholders came from the public and private sectors and are members of the archival, cultural, new media, and technology communities. They were essential in providing advice on national strategies for the long-term preservation of digital materials, including identifying barriers to opportunities for building a preservation infrastructure, rights and access management, and exploring models of cost-efficient sustainability and archiving.

Appendix 2: Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving. The Library of Congress and the Council on Library and Information Resources commissioned a series of six environmental scans focused in areas of digital collection development: electronic journals; e-books; digital sound recordings; digital video; digital television; and Web archiving. The environmental scans provide a baseline for understanding emerging issues that will mark the digital landscape in the future and highlight complicated and diverse challenges facing the cultural custodians responsible for preserving digital content. These scans were previously published in April 2000 by the Council on Library and Information Resources and the Library of Congress.

Appendix 3: Digital Preservation in the United States: Survey of Current Research, Practice, and Common Understandings. Dan Greenstein, Director of the Digital Library Federation, and Abby Smith, Director of Programs at the Council on Library Information Resources, summarize the activities in the United States under way that are designed to address the variety of preservation challenges—technical, legal, and social—and the changing roles and responsibilities of preservation stakeholders.

Appendix 4: Council on Library and Information Resources Survey on Digital Archiving. Author Dale Flecker reports on the findings of a Digital Library Federation survey and about its plans for digital archiving. The survey indicates that there are research libraries active in managing digital resources that constitute a logical and enthusiastic set of potential partners for the Library of Congress in the creation a plan for digital preservation.

Appendix 5: National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and Related International Activity. Author Neil Beagrie provides an overview of selected national and multinational initiatives in digital preservation, highlighting the fact that digital preservation requires collaboration of multiple stakeholders, both within the cultural and archival community as well as outside it.

Appendix 6: Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment. Author June Besek explains the relationship and impact of copyright management on preservation, highlighting copyright rights and exceptions and issues potentially involved in the creation of a nonprofit digital archive. Ms. Besek also identifies several areas that would benefit from further research.

Appendix 7: It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation. The Library of Congress and the National Science Foundation convened a workshop in April 2002 to define critical issues in preservation. The workshop brought together 51 people from government agencies, academia, and industry with expertise in computer science, mass storage systems, archival science, digital libraries, and information management to develop a research agenda. Margaret Hedstrom was a member of the organizing committee and provides an Executive Summary of the workshop findings.

Appendix 8: Highlights of the Library of Congress's Scenario Learning Process on the Future of Digital Preservation. Scenario planning helps organizations craft adaptive strategies in a climate of high uncertainty. In scenario planning, an organization creates a small number of detailed stories—or scenarios—about how the future might unfold based on different outcomes of critical uncertainties in the external environment. The organization then uses these scenarios as a platform from which to identify a high-level vision of a desired future state.

Appendix 9: Preliminary Architecture Proposal for Long-Term Preservation. Author Clay Shirky outlines a framework to support the technical functions of the NDIIPP. The proposed layered architecture is conceived to be modular, scalable, and compatible with a high level of technological change over time. The most important function of this document is to provide an initial direction for the specification and development of a national infrastructure by the Library of Congress and its public and private partners.

Appendix 10: Criteria for Projects. There are near- and long-term activities and investments, in addition to a number of initiatives already under way in many contexts, that can be mobilized as part of a nationwide system. A suite of projects and investments is required both to leverage federal investments effectively and to build functioning systems that will be positioned to take advantage of technological advances as these become appropriate. Collectively, these projects respond both to the design criteria and to the investment criteria.

APPENDIX 1

Consultation with Concerned Stakeholder Communities

Consultation with Concerned Stakeholder Communities

The Library of Congress has undertaken a series of activities involving concerned stakeholder communities. Representatives of major information technology companies, research and university libraries, not-for-profit and philanthropic institutions and foundations, and other federal agencies are members of the National Digital Strategy Advisory Board (NDSAB); these individuals are identified in Table 1. A series of convening sessions and workshops in the fall, winter, and spring of 2001–2002 offered representatives of a broad range of interests opportunities to participate in the planning process (see Table 2). Three meetings, involving about 70 representatives of stakeholder groups, were held in the first two weeks of November 2001. Two more structured scenario-planning workshops were held in February and April 2002, and a focused technical planning meeting was held in March. As part of the preparation for these sessions, a series of confidential interviews were conducted (see Table 3) and six environmental scans were commissioned and have been published (see Appendix 2). In parallel with these meetings and workshops, the Council on Library and Information Resources (CLIR) commissioned surveys of the members of the Association of Research Libraries (ARL), which represents the major research libraries, and the Digital Library Federation (DLF), which represents the libraries on the forefront of adopting digital technologies for their collections and services (see Tables 4 and 5).

Findings from these meetings, workshops, and surveys have informed the planning process in several important ways. First, the collective results established a baseline of knowledge. What is the level of interest and expertise? Which projects are under way? Which organizations would be willing to collaborate with the Library and within what type of framework? Second, results were used to plan successive activities or to initiate further studies. For example, concerns expressed by participants in the November 2001 meetings led to commissioning a white paper on copyright (see Appendix 6). Participants in the November meetings were asked to identify priority issues. Attendees at the workshops in winter and spring 2002 were asked to contribute to formulating the agendas. The NDSAB offered useful feedback based on interim reports. Finally, the consultation process became part of the basis for formulating the investment strategy and to defining the criteria for the portfolio of pilot projects that may be undertaken during successive phases of the National Digital Information Infrastructure and Preservation Program (NDIIPP). The process also identified potential pilot/experimental projects and partners.

The major findings and themes from the consultation process include the following:

Finding 1:

A number of digital preservation initiatives are under way in university, library, not-for-profit, and commercial venues. These are typically limited to the institution or organization and deal with heterogeneous information: dissertations, coursework, and teaching materials. The entertainment industries, notably the record labels, public television and cable television, also maintain internal archiving projects as dimensions of their asset management systems, but there is little industrywide collaboration and awareness of the importance of digital preservation varies. There is potential for coordination in some key areas, e.g., standards, best practices, selection, and collection development. Several universities and university libraries have explicitly expressed interest in collaborating with the Library in this effort.

Finding 2:

Technology informs almost every aspect of long-term preservation. It is not widely believed that there will be a single solution or that solutions can be achieved solely through technological means. Technological complexities vary across formats, but there is consensus around the following challenges: media and signal degradation; hardware and software obsolescence; volume of information, which surpasses the capabilities of current management strategies; strategies for data migration and emulation; urgency because of imminent loss; and the importance of distinguishing between an archived master and derivative works suitable for distribution through different transmission means. It is also important to begin working with material, both to capture valuable but highly ephemeral items and to test possible technical solutions.

Finding 3:

Roles, including relationships among federal libraries, university and research collections, and public libraries, are linked to collection management and acquisition policies. Issues include: redundancy of collections, which enhances security, versus unnecessary duplication, inefficiency and waste; the Library as a standards/best practices body and broker among systems and institutions; the Library as a potential collector of “last resort”; the importance of preserving work deposited for copyright; funding libraries with strong topical collections to digitize these materials within a system of interworking libraries.

Finding 4:

Commercial and industry representatives agreed on the following: There is a *business* case for collaborative long-term archiving of digital content; the challenge is to work out an architecture and set of policies that balance economic and cultural interests. Some discussants emphasized the importance of managing the archive (format and software obsolescence, storage media deterioration, signal degradation, playback, metadata, technical standards, and best practices) while others emphasized issues related to library services (e.g., intellectual property rights management, cost recovery, access).

Finding 5:

An appropriate balance between the economic rights of rights holders and the importance of public access and use of the collections is needed. Concerns were voiced about the Digital Millennium Copyright Act and its implications for libraries and archives. Other issues include privacy and confidentiality. As a result of questions raised by this process, a white paper on copyright has been commissioned and is presented in Appendix 6 of this document.

Finding 6:

Suggestions for sustainability included: requiring funding as a condition of acceptance of collections; public funding; partnerships and collaborations with other public and private organizations; incentives, for example, in the form of tax credits for contributed materials or contributed uses of materials that remain privately owned.

Finding 7:

Both the National Archives and Records Administration (NARA) and the National Library of Medicine (NLM) have active digital preservation initiatives under way. The National Agricultural Library has also initiated work in this area. In different ways, all three agencies have begun to cope with the managerial, organizational, and technical issues that arise when dealing with distributed systems and information flows between and among cooperating libraries.

Table 1. National Digital Strategy Advisory Board (NDSAB) Members

Members of the NDSAB represent a range of interests and include representatives from the information technology industry, publishing, philanthropic foundations, Department of Commerce, National Science Foundation, Office of Science and Technology Policy, National Institute of Standards and Technology, National Archives and Records Administration, other national libraries in the U.S. and internationally (National Agricultural Library, National Library of Medicine, British Library), major university and research libraries, and the National Academies.

Name	Affiliation
Executive Committee	
Jim Barksdale	Barksdale Management Corp.
James H. Billington	Library of Congress
Laura Campbell	Library of Congress
John W. Carlin (and Lewis Bellardo)	National Archives and Records Administration
Don Evans (and Elizabeth Prostic/ Tom Pyke)	U.S. Department of Commerce
Mario Morino	Morino Institute
Michael C. Ruettgers	EMC Corporation
Richard Russell	White House Office of Science and Technology Policy
John F. (Jack) Sandner	Chicago Mercantile Exchange
Broader Advisory Board	
Lynne Brindley	The British Library
Nancy Eaton	The Pennsylvania State University
Eleanor G. Frierson (alternate)	National Agriculture Library
James Gray	Microsoft
Margaret Hedstrom	University of Michigan
Clarence L. Irving	Irving Information Group
Glenn Jones	Jones International, LTD
Brewster Kahle	Alexa Internet
Stephen M. Griffin	National Science Foundation
Donald A.B. Lindberg (and Betsy Humphries)	National Library of Medicine
Clifford Lynch	Coalition for Networked Information
Victor McCrary	National Institute of Standards and Technology
Carol Mandel	New York University Library
Charles Phelps	University of Rochester
Richard S. Rudick	John Wiley & Sons
Richard Sarnoff (and Larry Weissman)	Random House
Donald J. Waters	The Andrew W. Mellon Foundation
William A. Wulf	National Academy of Engineering
Consultant to the NDSAB	
Deanna B. Marcum	Council on Library and Information Resources

Table 2. Participants in Meetings and Workshops

Participants in the various meetings and workshops represented a cross section of interests, including information technology industry, entertainment (motion pictures, commercial and noncommercial radio, broadcast and cable commercial and noncommercial television), publishing, philanthropic foundations, the Department of Commerce, the National Science Foundation, the White House Office of Science and Technology Policy, the National Institute of Standards and Technology, the National Archives and Records Administration, and other national libraries and major university and research libraries. Individuals were identified through a series of methodologies including: snowball interviews with purposive starts, referrals by professional associations, referrals by participants and personal contacts within relevant communities.

Activities included three meetings in November 2001, which included about 30 participants each with representation from within the Library of Congress as well as a cross section of concerned groups; two scenario planning workshops in February and April 2002 of similar size and composition, which included new participants as well as people who had been invited to earlier sessions; and a workshop of about a dozen technical experts who focused on issues associated with the technical architecture. This technical session took place in March 2002. Finally, there was a meeting in Hollywood in June 2002 to which individuals in the motion picture industry were invited.

Name	Affiliation
Darcy Antonellis	Warner Bros.
Wendy Aylsworth	Warner Bros.
Stephanie Barish	University of Southern California
Michael Barrett	Kodak
Mick Bass	Hewlett Packard
Meg Bellinger	Preservation Resources/OCLC
Pieter S. H. Bolman	Elsevier Science
Roma Bose	PricewaterhouseCoopers
Scott Bowen	Artesia Technologies
Michele Boxley	American Institute of Architects
Stewart Brand	Global Business Network
Dick Brass	Microsoft Corporation
Tom Broido	Music Publishers' Association
Terry Brown	Society of Illustrators
Joseph B. Bruns	WETA-TV and FM
John Clippinger	EcoCap
Chris Coldewey	Global Business Network
Michael Cornfield	George Washington University
Grover Crisp	Sony
Steve Crocker	Longitude Systems
Robin Dale	Research Libraries Group
Elizabeth Monk Daley	University of Southern California
Malcolm F. Davidson	Sony Music Entertainment
Troy Dow	Motion Picture Association of America
George Dyson	Western Washington University
Chris Ertel	Global Business Network
Mike Fahey	EMC Corporation
Theodore H. Feder	Artists Rights Society
Eileen Fenton	JSTOR

Table 2. Continued

Name	Affiliation
Dale Flecker	Harvard University Library
Edward O. Fritts	National Association of Broadcasters
Eleanor Fye	Microsoft Corporation
Kevin Gage	Warner Music Group
Harold D. Gangnath	PricewaterhouseCoopers
Nadina Gardner	Heritage Preservation
Tom Garnett	Smithsonian Institution Libraries
Carlos Garza	Recording Industry Association of America
Peter Gordon	PricewaterhouseCoopers
Daniel Greenstein	Digital Library Federation
Garrett Gruener	Alta Partners
Georgia K. Harper	Univeristy of Texas at Austin
Brett Harvey	American Society of Journalists and Authors
Karen Hunter	Elsevier Science
Jennifer Insogna	EMI Music Publishing
Nat Irvin II	Wake Forest University
Steven Jones	University of Illinois at Chicago
Michael A. Keller	Stanford University
Kevin Kelly	All Species Foundation / Wired Magazine
Mark Kelly	Defense Intelligence Agency
Marsha Kinder	University of Southern California
Arthur Klebanoff	Rosetta Books
Jack Lacy	Intertrust
Adam Lee	BBC
Edrolfo Leones	Walt Disney Company
Catherine Levene	New York Times Digital
Dick Lindheim	ICT, Paramount
Nina Link	Magazine Publishers Association of America
Peter Lyman	University of California, Berkeley
Dave MacCarn	WGBH, Boston
Robert Madden	Devon Jacklin Photography
Philip Brook Manville	Saba Software
Peter Marx	Universal Studios
Maureen Matheson	American Foundation for the Blind
Richard May	Warner Bros.
David Miller	NIMA
Alan Mink	National Institute of Standards and Technology
John Lewis Needham	ebrary
Michael R. Nelson	IBM
Richard P. O'Neill	Highlands Group
Walter Parkes	Dreamworks
Angela Peters	Association of American Publishers
Adam Clayton Powell III	Freedom Forum
Sallie Randolph	American Society of Journalists and Authors
Larry Reger	Heritage Preservation
David Rodgers	University of Michigan
Alex Roland	Duke University
Alexander Rose	Long Now Foundation

Name	Affiliation
Hilary B. Rosen	Recording Industry Association of America
David Rosenthal	Sun Microsystems Laboratories
John Schline	Penguin Putnam Inc.
Michael Schrage	MIT Media Lab
Peter Schwartz	Global Business Network
Richard Sergay	ABC News
Jonathan Spalter	Vivendinet
Clay Shirky	New York University
Abby Smith	Council on Library and Information Resources
Robert (Bob) Spinrad	Technical Expert (formerly Director of Xerox PARC)
Michael Spinella	American Association for the Advancement of Science
Christopher Sterling	George Washington University
Edward Tenner	Princeton University
Ralph Terkowitz	Washington Post
Kenneth Thibodeau	National Archives and Records Administration
Mary Lou Tillotson	PricewaterhouseCoopers
Jack Valenti	Motion Picture Association of America
Hal Varian	University of California, Berkeley
Howard Wactlar	Carnegie Mellon University
Ken Wasch	Software and Information Industry Association
Christopher Wera	Cable Center
Paul J. West	Universal Music Group
Alison M. White	Corporation for Public Broadcasting
Christopher Williams	Kodak
Lee Zlotoff	Auras Unlimited

Table 3. Experts Interviewed

Interviews were conducted in preparation for the sessions in November 2001 as well as the workshops in winter/spring 2002. The cross section of interests reflected in the interviews are reflective of the stakeholder groups previously identified. The interview strategies varied from relatively open-ended in the fall 2001 phase to more focused approaches in the winter/spring 2002 phase. The format also varied, depending on the interviewee's availability and preference; essentially three levels of communication were employed: full interview of 30–60 minutes with verbatim transcript; briefer telephone conversations of 15–30 minutes duration with notes; and e-mail exchanges.

Name	Affiliation
David Brin	Independent Author
Lynne Brindley	The British Library
John Seely Brown	Xerox PARC
John Carey	Columbia University
Roger Cass	Economist and Consultant
Steve Crocker	Longitude Systems
Elizabeth Monk Daley	University of Southern California
Malcolm Davidson	Sony
Nicholas DeMartino	American Film Institute
Esther Dyson	EdVentures
Nancy Eaton	Pennsylvania State University
Colin Franey	EMI
Elizabeth Frayzee	AOL/Time Warner
Carlos Garza	Recording Industry Association of America
James Gray	Microsoft Bay Area Research Center
James Hindeman	American Film Institute
Robert Kennedy	C-Span
Marsha Kinder	University of Southern California
Arthur Klebanoff	Rosetta Books
Jack Lacy	Intertrust
Edrolofo Leones	Walt Disney Company
Allen Mink	National Institute of Standards and Technology
Michael R. Nelson	IBM
Martin Nissenholtz	New York Times Digital
Adam Clayton Powell III	Freedom Forum
Ray Roper	Printing Industries of America
Richard Rudick	John Wiley
John Schline	Penguin Putnam
Steve Weber	University of California, Berkeley/GBN
Larry Weissman	Random House
Paul West	Universal Music
Woodward Wickham	MacArthur Foundation
Chris Williams	Kodak
Troy Williams	Questia

Table 4. Association of Research Libraries (ARL) Members Surveyed

ARL consists of more than 120 member institutions that represent the major research libraries in North America. At the request of the Library of Congress, the Council on Library and Information Resources (CLIR) requested ARL to poll its members concerning their activities in long term preservation of digital content.

Institutions

University of Alabama
Arizona State University
Boston College
Boston University
University of British Columbia
Brown University
University of California, Berkeley
University of California, Davis
University of California, Riverside
University of California, San Diego
Case Western Reserve University
University of Chicago
University of Colorado
Columbia University
University of Connecticut
Cornell University
Dartmouth University
Duke University
University of Florida
Georgia Tech
Harvard University
University of Houston
University of Illinois, Urbana
University of Iowa
Iowa State University
Johns Hopkins University
University of Kansas
Kent State University
Louisiana State University
McGill University
University of Massachusetts
Massachusetts Institute of Technology
University of Michigan
Michigan State University
University of Minnesota
National Agricultural Library
National Library of Canada
University of Nebraska
New York University
University of North Carolina
North Carolina State University
University of Notre Dame
Ohio University
Ohio State University
University of Oklahoma
Pennsylvania State University
University of Pittsburgh
Rutgers University
University of South Carolina
Southern Illinois University
University at Albany, SUNY
University at Buffalo, SUNY
SUNY Stony Brook
Syracuse University
University of Tennessee
Texas Tech University
University of Toronto
Vanderbilt University
University of Virginia
Virginia Tech
University of Washington
Washington State University
University of Waterloo
Wayne State University
University of Western Ontario
Yale University
York University

Table 5. Digital Library Federation (DLF) Respondents to Survey

DLF is a consortium of about 30 research libraries that are on the forefront in the adoption of information technologies to extend their collections and services; it is housed within the administrative umbrella of the Council on Library and Information Resources (CLIR). At the request of the Library of Congress, CLIR requested DLF to poll its members concerning their activities in long-term preservation of digital content.

Institutions
California Digital Library
University of Chicago
Cornell University
Emory University
Harvard University
Indiana University
University of Michigan
University of Minnesota
New York University
Pennsylvania State University
Stanford University
University of Texas
University of Washington
Yale University

APPENDIX 2

Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving

Contents

Previously published in April 2002 by the Council on Library and Information Resources and the Library of Congress.

ISBN 1-887334-91-2

Copyright 2002 in compilation by the Council on Library and Information Resources and the Library of Congress. No part of this publication may be reproduced or transcribed in any form without permission of the publishers. Requests for reproduction or other uses or questions pertaining to permissions should be submitted in writing to the Director of Communications at the Council on Library and Information Resources, 1755 Massachusetts Avenue, NW, Suite 500, Washington, DC 20036.

Preface

Libraries traditionally have formed a preservation safety net for materials that will be transmitted to subsequent generations of information seekers and scholars. For paper-based documents, provision of adequate storage conditions was the best means to help ensure that materials would remain readable far into the future.

With the advent of digital technology, many knowledge creators do their work on computers. Some of that knowledge may be printed on paper, but much of it, particularly databases, geographic information, scientific data sets, and Web sites, exists only in electronic form. At the same time, traditional forms of publications have changed significantly and, as a result, create new challenges. For example, publishers of electronic journals license their content to libraries, but libraries do not own that content and they may not have rights to capture digital content to preserve it.

What organizations or systems will provide the needed preservation safety net for electronic materials? Recognizing the importance of this question, the U.S. Congress in December 2000 appropriated funds to the Library of Congress (LC) to spearhead an effort to develop a national strategy for the preservation of digital information. Understanding that the task cannot be accomplished by any one organization, Congress wrote into the appropriations language a requirement that LC work with other federal, scholarly, and nonprofit organizations to discuss the problem and produce a plan.

The staff of the Library of Congress immediately scheduled a series of conversations with representatives from the technology, business, entertainment, academic, legal, archival, and library communities. LC asked the Council on Library and Information Resources to commission background papers for these sessions and to summarize the meetings. The resulting papers, along with an integrative essay by Amy Friedlander, are presented in this document.

The responsibility for preserving digital information will be distributed broadly. Our hope is that information gathered by the Library of Congress will benefit all who are working on this issue.

Deanna Marcum,
President, CLIR

Laura Campbell,
Director, National Digital Library Program
Library of Congress

Summary of Findings

AMY FRIEDLANDER
Center for Information Strategy and Policy
Science Applications International Corporation

The late twentieth century saw the beginning of the age of digital information in corporate archives, the creative arts, financial markets, medical information, and scholarship, among other venues. How the United States chooses to preserve and manage its digital information affects core issues in key industries—from medical textbook publishing to entertainment and to future scholarship in science, technology, and the arts and humanities. It profoundly affects how the future will come to know our present and is, therefore, integral to the nation's identity, now and to come. In this terrain, the Library of Congress (LC) has chosen to open its investigations with a series of probes into six principal areas in which the LC faces collection-management issues: large Web sites, electronic books, electronic journals, digitally recorded sound, digital film, and digital television. This chapter summarizes what a series of interviews and papers, conducted and written during the late summer and early fall 2001, revealed about a complex and shifting landscape.

Formal 30-minute interviews and shorter conversations and e-mail exchanges were conducted with individuals who represent a range of interests and organizations across publishing, film, entertainment, news, electronic books, computer science, libraries, corporate research, nonprofit organizations, professional and trade associations, and academe. Their names and primary affiliations are listed on page 24. (Note that corporate representatives frequently sit on the boards of nonprofit and cultural organizations, and many communities therefore inform their perspectives.) Most people talked about several concerns and formats; thus, we have abandoned any efforts to characterize responses exclusively by format (e.g., e-books or e-journals, Web sites, digital film, digital TV, digitally recorded sound), profession, or organization.

Information gained from the interviews was complemented by six “environmental scans” that were intended to provide baseline information for concerned groups outside the library, preservation, and archival communities. Their intent was to define the basic issues while illuminating the concerns brought by the library, preservation, and archival communities.

Not surprisingly, there is a range of opinion and emphasis placed on different issues across communities. In the following pages, we summarize some of the key findings.

“Born Digital” Versus Digitized

The scope of the effort was defined to encompass material that is “born digital,” that is, objects that have been created in digital form rather than converted from analog to digital. This distinction, however, was not consistently useful to interviewees or to the writers. Historic film or news footage may be embedded in a newly created digital educational project. Re-release of entertainment products partly or wholly in digital form, either as new editions of older works or as reused elements in an otherwise-new work, further blurs the distinction. The production process itself is not hermetically sealed analog or digital. “Materials collected or generated for a television show,” wrote the team from the WGBH Educational Foundation, “may consist of a great threaded mesh of digital and analog components, so tightly bound together that, at any point in their life cycle, one may serve as surrogate for another.” A similar case can be made for radio broadcasts, and many persons in the recording industry agree that preservation of a digitally recorded sound product should include its packaging—the notes, artwork, and photograph of the artist, for example. Even on the Web, many sites offer digitized versions of print works; for this reason, archiving the Web itself can be seen as encompassing both born-digital and digitized materials. One publishing executive argued that “digital” should be thought of as a medium in which content was both created and made accessible to the public. However, another publisher cautioned that the distinction between “digitized” and “born digital” is very important because it relates to the concept of completeness, and that accompanying that concept are notions of “copies,” “versions,” and other ideas critical to managing works and their associated rights.

The Scope

The notion of scope arose at many levels, from the definition of the object to the extent of the effort. Several people inside and outside the library community urged planners to consider the scope of the effort carefully, including such factors as what was selected for the collection (even if it were a single collection), its longevity (10, 100, or 1,000 years), and its purpose (preservation, limited access, or public access). From a practical point of view, given the sizes of the resources, selection seems particularly important in film, television, and the Web. The Web is complicated by the fact that only part of it is publicly accessible and by unresolved issues over rights. It is not clear, for example, that a Web site may be “harvested” for purposes of preservation without the knowledge and permission of the various rights holders. (In the case of an interactive Web site, the range of potential rights holders extends well beyond those involved in its creation.)

Several people in both the technical and the arts communities urged attention to "ephemera" as well as to "published" works (the definition of "publication" is being contested). Others believed the effort would do well to focus on published materials subject to copyright and to which the LC has a clear mandate. A number of respondents in film, television, and sound noted that again, the distinction between publication and ephemera is blurred. For example, a historic radio broadcast that is captured by the listener may contain aural information that reflects its relatively poor reception at the time; retaining that quality goes to the traditional mandate of preserving the experience, which might not be reflected in the script or in a studio recording. Similarly, only a very small percentage of the material shot is actually used in the commercial release of a film, yet digital video disc (DVD) releases have provided new life for outtakes and other associated production materials. The relative utility of material changes over the cultural life of a film or a performance; the first public release does not necessarily capture all of its aesthetic or future scholarly value. There is a substantial economic incentive, since enhancing a DVD release is one strategy for combating piracy.

The notion of scope also surfaced at the level of the artifact or item. Discussions of Web sites, e-books, e-journals, and digital television make clear the difficulty of drawing boundaries among these items. Within the Web itself are emerging distinctions between the "surface" Web and the "deep" Web. E-books and e-journals download content from the Web to their respective formats and include hyperlinks to the Web for ancillary augmentation. The advent of interactive television also invites new forms of multimedia that combine resources built for the Web with those created for broadcasting in digital form. Moreover, an item that appears seamless to the user is frequently a composite document. Formats as well understood as electronic scholarly journals are built as multimedia objects in which the constituent elements may include text, images, animation, or advertisements, each of which may be encoded in a different format. Finally, several people from the arts communities emphasized the importance of collecting the version of the object that the creator (e.g., the director of a film) considered final in the format that he or she considered final.

There are complexities to notions of "authorship"; many of these are not new to digital but are magnified by the circumstances under which digital products may be distributed and used. These complexities are related to the complicated intellectual property considerations that surround digital information. Even in a format as carefully studied as is that of electronic scholarly journals, creation and deposit can involve numerous stakeholders, and the number of interested parties multiplies in sound, television, and film, in which individuals and entities have traditionally had rights in the processes of creation and distribution. Frank Romano points out that the e-books world is witnessing changes in traditional roles and functions; for example, writers can self-publish and thus become distributors, while software companies can behave like publishers. Similar shifts and realignments can be seen in some metadata discussions, where, as Peter Lyman notes, both computer scientists and librarians are putting forth different yet overlapping views of how the systems might work.

Technical Issues Associated with Long-Term Storage

Early in the interview process, one of the technical experts cautioned planners not to "underestimate" the importance of and differences among formats. There was, nonetheless, a consensus around the basic issues, if not necessarily around solutions. The issues, which include technical obsolescence and standards, metadata, information security, and the overall architecture of the system, are by no means discrete. For example, standards affect creation as well as preservation. As one scholar of film and new media pointed out, the evolution of his organization's Web site represented a patchwork of changing and evolving standards. Several writers pointed out that the issue is not only making sure that bits survive but also ensuring the preservation of a technical environment that will permit future retrieval of the information, the work as envisioned by the author or creator, and the experience of the user.

The longevity of the storage medium was a consistent concern, as was signal degradation and software obsolescence. One technical expert urged that degradation be compared with the process by which a photograph ages. The image fades; the medium on which the image is printed also disintegrates. There are methods for error detection; however, at some point, there is concern that the integrity of the digital object is compromised.

One solution is migration from one medium to another. However, there are discussions over whether to use sampling/compression strategies (particularly if the object is made available in, for example, Joint Photographic Experts Group [JPEG] or Motion Picture Experts Group [MPEG] format), the extent to which migrating the information introduces errors if the data are resampled, and the implications of migrating formats for version control and integrity. When a digital work is migrated (e.g., from MPEG^{n} to MPEG^{n+1}), perhaps in very short order given the rapid development of the technology, what is the original work? In the case of recorded sound, for example, would improvements to fidelity resulting from more sophisticated software technology compromise the integrity of the original, since it is no longer truly the artist's treatment of a work and misrepresents the recording technology of the time?

At least one technical expert did not consider this to be a serious problem but did acknowledge that the rules for the successive formats must be retained. On the other hand, the team from the WGBH Educational Foundation noted that while standard archival practices call for refreshing the data through migration and emulation, these strategies might be inadequate for "handling the intricacies, interdependencies, and sheer volume of television content." For film and television, this has resulted in attention to selection and collection policies inside traditional libraries as well as other organizations and has highlighted the importance of metadata as a management tool.

Playback

Playback—usually associated with the equipment or software that enables users to re-create the performance of a film, for example—was seen to be a particular problem for e-books as well as for digitally recorded sound and film. For example, certain early tapes are no longer accessible because the equipment to read them no longer exists or is hard to find. Playback affects any effort to enable future users to re-create the work (however defined) as it was originally experienced. Issues associated with playback can be expanded to operating systems, browsers, and so on. Solutions vary from emulation to maintaining collections of relevant hardware and software so that an archive or archiving system of digital content can imply preservation of certain kinds of equipment as well. Particularly for e-books, where so much of the design is predicated on screen size, re-creating the experience for future users implies access to the device that was intended to display the content.

Standards and Technical Obsolescence

The rapid obsolescence of some formats, as well as the plethora of standards, was widely considered to be a barrier both to creation and to preservation. Those who had opinions on open versus proprietary standards favored the former because they were believed to facilitate management of the archive and its content. This applies to a broad range of issues, from operating systems to markup language, compression, and fonts.

Information Security

Before September 11, 2001, few people consulted had strong opinions on information security, but those who did thought that it was important as a guarantor of trust. One technical expert did not see the information security needs of an archive as being different from its general needs, or that, for example, the mission of the archive added a layer of concern. Another technical expert cautioned that “security” means a number of things in this context, including robustness and safety of the storage, privacy, and copyright control. The interviewees recommended that discussions of security be kept “simple and clear” to reduce ambiguity, unnecessary conflict and, perhaps, undue emphasis at this point.

With respect to confidentiality and privacy, several people noted different dimensions and concerns that arise when the procedures associated with managing the archive go digital. One example that was offered was the information typically provided on copyright registration concerning the authors, who might use a pseudonym or who might wish to keep their own addresses, or the addresses of their agents, from general use (Salman Rushdie was the example offered). There are overlaps between this kind of information and the information included in metadata. At least one person cautioned against excessive restriction, arguing that too many restrictions inhibit accountability.

Proposals for Storage Architecture

Those who addressed technical issues tended to favor distributed rather than centralized systems, because the former would accommodate a high degree of "local" variation within shared protocols. There were also calls for interoperability, which would make it possible for information to be shared across platforms and among vendors. One publisher thought it was important that the LC do the development in-house, avoid proprietary software, and use commercially available tools because this approach would facilitate future upgrades to the system. Two architectural approaches were set forth: one for e-journals (see chapter by Dale Flecker), which has been fleshed out in some detail, and a more rudimentary approach that looks at the problem of preservation from a broad perspective in which the LC is one of many entities that might be involved. Discussions are ongoing about the extent to which content may be partitioned as a layer that is separate from formats, metadata, applications, and access policies, mechanisms, and controls. But, as one technical expert noted, the technology is likely to be developed outside of the traditional library community by other interests. The LC has an important role as "stimulator of initiatives and a consumer of successful technologies," but it does not have the money or expertise to dictate an outcome. Nearly all of the people interviewed, whether or not they commented on technical issues, agreed with this comment insofar as it acknowledges the importance of the LC's imprimatur.

Metadata

Metadata, or "data about data," are simultaneously a standard, a management and access tool, and a feature of the system architecture. For example, whether the metadata are bundled into the content or are maintained separately is a question that is being discussed with respect to several formats. This is a matter that affects approaches to interoperability as well as system design. The team from Carnegie Mellon University argued persuasively for the importance of metadata to the management of the archive as well as for providing appropriate access. The chapter by Wactlar and Christel delineates in some detail the several approaches to metadata, illustrating the range of academic and commercial interests that have become involved in defining metadata. Moreover, as pointed out by Lyman in his study of archiving the Web, the metadata discussions reveal the different visions of archiving as embodied by the library and computer science communities. He writes, "The librarian tends to look at the content of the Web page as the object to be described and preserved. The computer scientist tends to look at the Web as a technology for linking information, thus looks at the Web as a system of relationships (hence the name "Web")."

One of the functions of metadata, as the various schemes have evolved since 1995, is outlining the terms and conditions of use—that is, access. This thorny issue is discussed in the next section.

Access and Rights Management

Few failed to identify intellectual property rights (IPR) management and fair use as key issues. Each of the chapters addresses IPR at some level, with perhaps the most general discussion offered in Peter Lyman's chapter on archiving the Web. The complexity of this set of issues varies across media. Thus, questions of international law hang heavily over the Web and any products that are distributed through the Web, while changing perceptions of who is or is not a public figure and the layered rights associated with recorded sound, film, and television figure prominently in discussions of those formats.

The interviews showed confusion over whether archiving for purposes of preservation could be decoupled from use. Some of this ambiguity arose from an appreciation of the mission of the LC as a repository that supports scholarship and is in some way "the nation's library." Some arose from unfamiliarity with the distinction that is common among traditional preservation circles in which use of rare objects, for example, can be calibrated and surrogates used in their stead. (This is one of the rationales both for bibliographic records and for metadata, which enable scholars to learn about an object without accessing the object itself.) Finally, there is an inherent tension in the entertainment and publishing industries: the value of a digital asset lies in providing access to it, but unauthorized access and duplication can reduce its value.

While there was near unanimity on the importance of managing intellectual property responsibly, no voices called for some version of complete lockout. Indeed, one representative from a major company with interests in several areas thought that the most important issues were both protection of intellectual property rights *and* ease of use (with appropriate accommodation for potential users with special needs). There was widespread acknowledgment of the need to find a balance between the economic needs of the creators and distributors and the legitimate uses of the works, but there was a range of opinion as to what that meant. Some suggested ways to handle management of intellectual property "behind the scenes" through technological means, which could be coupled with pricing that discouraged inappropriate use. Other proposals revolved around ways to use time, such as restricting access on the basis of estimates of time during which the owner expected to extract the economic value. However, product cycles of reuse would complicate that approach.

Several people felt that existing laws were sufficient: what is required, they maintained, is appropriate enforcement. Others believed that there was a need to clarify the law. Given that the Web is an international phenomenon, attention to international law is particularly important. As of this writing, terms such as "copy," "publication," "performance," and "public figure," whose definitions were once widely agreed on, had become subject to discussion. Still others pointed to misperceptions that were clouding the discussion in several contradictory ways: people think that information in digital form has both more value (those who tended to inflate the costs for permissions) and far less value (those who thought information should be free) than does information in analog form. Finally, a number of people, particularly in the film and entertainment industries, noted that the inflamed environment in which the discussions are taking place makes reasonable attempts at compromise very difficult.

Several people pointed out that copyright as a mechanism, which had arisen in the context of print, had already begun to fray under the stress of its application to media other than text, and that it was becoming increasingly unwieldy. For example, in film, the multiplicity of rights and permissions that affect distribution and reuse of material had derailed educational projects because it was simply impossible to unravel the layers. Recorded sound has similar layers of rights (see chapter by Peter Lyman). Finally, ambiguity over the law is itself becoming a barrier. Faculty members are wary of developing new course work for online learning in an environment in which there is no consensus about appropriate conduct and the legal ramifications of their decisions are unknown.

Individuals Consulted

Lynne Brindley
British Library

David Brin
Author

John Seely Brown
Xerox PARC

John Carey
Columbia University

Steve Crocker
Longitude Systems

Elizabeth Daley
Annenberg Center
University of Southern California

Nicholas DeMartino
American Film Institute

Nancy Eaton
Pennsylvania State University

Colin Franey
EMI

Elizabeth Frayzee
AOL/Time Warner

Carlos Garza
The Recording Industry Association
of America

James Grey
Microsoft Bay Area Research Center

James Hindman
American Film Institute

Marsha Kinder
Annenberg Center
University of Southern California

Edrolfo Leones
The Walt Disney Company

Allen Mink
National Institute of Standards
and Technology

Adam Clayton Powell, III
The Freedom Forum

Richard Rudick
John Wiley

Ray Roper
Printing Industries of America

John Schline
Penguin Putnam, Inc.

Larry Weissman
Random House

Woodward Wickham
MacArthur Foundation

Troy Williams
Questia

Preserving Digital Periodicals

DALE FLECKER
Harvard University Library

Introduction

Everyone has a vague, but few a very precise, idea of what constitutes a “periodical.” For the purpose of this paper, a “periodical” will be defined as a primarily text-oriented publication that regularly issues new content and intends to do so for the indefinite future. Digital periodicals come in many flavors; they include selections or versions of paper magazines, such as *Wired;* peer-reviewed scholarly journals; e-’zines; online newspapers; boutique electronic updates or analyses for the business executive; and trade, political, or special-interest newsletters. These may or may not exist in parallel print/paper form; the two formats may not constitute perfect substitutes for one another. The variety makes generalizations difficult; the analysis that follows will be accurate for the primary body of periodicals, but the wide variety of producers in this realm ensures that exceptions will be common.[1] Digital periodicals are sometimes based on e-mail delivery or occasionally on the use of specialized reader software, but most today are delivered over the World Wide Web, and that environment is our focus here.

In the paper era, libraries subscribed to and maintained collections of many periodicals (the Harvard libraries still receive about 100,000 active titles), and collections were highly redundant. Libraries invested in a range of activities intended to maintain the usability of what they collected: binding materials in protective enclosures, repairing damage, housing collections securely and in environments designed to prolong the life span of paper, and reformatting deteriorated materials through photocopying or microfilming. With the exception of microfilm masters, the copies of journals being saved for future generations were the same copies being read by the library’s current users. While in research libraries operations were always planned with one eye on the indefinite future, the actions that preserved materials for future generations also served to maintain them for current use.

The new world of Web-delivered periodicals is different. While libraries continue to subscribe to periodicals as they migrate to digital form (subscriptions to electronic journals number in the thousands today in most academic libraries), the service model has changed fundamentally. Libraries no longer receive and store materials locally, and subscriptions no longer provide copies but a license to access. This change has profound implications for the archiving and preservation of periodicals because it removes two key attributes of the current system:

1. maintenance of copies of periodicals primarily for users of future generations; and
2. redundancy of copies, which ensures that accidents, theft, conscious destruction, or changes in policy or priority at any given institution do not result in the complete loss of the published record.[2]

Digital materials are surprisingly fragile. They depend for their continued viability upon technologies that undergo rapid and continual change. All digital materials require rendering software to be useful, and they are generally created in formats specific to a given rendering environment. In the world of paper, many valuable research resources have been saved passively: acquired by individuals or organizations, stored in little-visited recesses, and still viable decades later. That will not happen with the digital equivalents. There is no digital equivalent to that decades-old pile of *Life* or *National Geographic* magazines in the basement or attic. Changes in computing technology will ensure that over relatively short periods of time, both the media and the technical format of old digital materials will become unusable. Keeping digital resources for use by future generations will require conscious effort and continual investment.

In the new world of digital periodicals, copies of materials are often held by a single institution, and the investments required to maintain their long-term viability must be made by that institution, which presumably owns them. Factors such as changes in the economic viability of materials, the high cost of a technical migration, a new market focus, company failure, or a reduction in available resources all cause worry about whether such continuing investments will be made. Without such investments, materials will be lost. Such concerns have led libraries to cling to paper copies, when available, even while they provide electronic versions of the same material for the daily use of their readers. This duplicate cost will obviously be problematic over time, and the issue of how to archive and preserve Web-based periodicals is widely felt to have reached a critical state.

Technical Profile of Digital Periodicals

Digital periodicals are surprisingly complex given the seeming simplicity of their paper antecedents, and the level of complexity is growing. The content of digital periodicals comes in a wide variety of technical formats, varying not just among publications, but within a single title or article. The following discussion is not exhaustive of the types of digital material that make up current periodicals, but it is indicative of the scope of complexity involved.

The core content of most periodicals is text. The text of a periodical or periodical article, however, can be created and maintained in a number of ways. Some current periodicals are composed of digital pictures of printed pages (frequently, these are then embedded in portable document format [PDF] wrappers for delivery and viewing in the Web environment). More commonly, text is encoded in one of several ways. Some simple publications encode the output of word-processing programs in hypertext markup language (HTML) for Web viewing. HTML provides a rather simplified level of content "markup," primarily oriented toward good visual presentation in today's Web browsers. More sophisticated publications, particularly those thought by their creators to be of lasting interest, are frequently encoded in standard generalized markup language (SGML) or extensible markup language (XML), both of which support much more detailed labeling of components of a textual document than HTML does. However, SGML and XML are enormously flexible, and different publishers use highly varied markup schemes (e.g., document type definition [DTD] schemas). Software to render text marked up in this way must be sensitive to the specific scheme used in the text being displayed.

A critical issue with computerized text is the character set used to represent the letters, ideographs, or other components. Standardization in the encoding of text components has progressed enormously in recent years, particularly with the development and adoption of Unicode[3] by an increasing range of technology providers. Text for most contemporary languages can be fully encoded in Unicode. However, textual documents contain more than letters and words, and many of the specialized symbols used in periodicals do not have standard digital representations, or evolving standards are not yet widely implemented for them. These include

- mathematical symbols
- chemical formulas
- archaic scripts or ideographs, such as Egyptian or Mayan hieroglyphs
- musical notations

Publications containing such extended characters or notations today use a variety of conventions for storage, and rendering software must be sensitive to these conventions when preparing text for Web display.

Periodicals contain more than simple text. Visual materials such as photographs and drawings are extremely common and can be encoded in different technical formats. Increasingly, sound and video clips are found in periodical publications, again in a variety of technical formats.

Advertisements represent particular difficulties for archiving and preservation. In paper periodicals, advertisements were usually tied inextricably to specific issues. With Web publications, although most periodical content is relatively static once published, advertisements seen in a particular context can change from minute to minute or from day to day. Advertisements can be selectively displayed for specific audiences or national communities (varying in language or in response to legal restrictions, such as those for drug advertisements). Advertisements are often delivered from a different source than the periodical itself and in fast-changing, proprietary, and challenging technical formats that try to stay on the cutting edge. Advertisements represent a rich source for historical research, and their preservation will be of interest. However, archiving and preserving advertisements will pose a significant challenge.

There are other new types of periodical content that raise technical issues. Increasingly, scholarly articles are accompanied by “supplementary materials”—files containing detailed research data, further explication of the article information, or demonstrations of points made in the article. These files contain many types of information (statistical data, instrumentation data, computer models, visualizations, spreadsheets, digital images, sound, or video) and come in a wide range of formats, usually dependent on whatever technical tools the author is using at a given moment. Journal editors and publishers frequently exercise no control over these formats, accepting whatever the author chooses to deposit. More than any other instance of periodical content, these supplementary materials introduce a rapidly growing and essentially unbounded flow of new technical formats that will pose significant difficulties for long-term preservation.

Because digital periodicals are composed of many pieces, frequently in differing technical formats, some form of relationship information is required to map the pieces into a coherent form for delivery to a user. This relationship information can take many forms: “container” formats (such as PDF) that hold explicit or implicit relationships, XML documents, metadata databases, and static HTML documents. Practices for what data are recorded and how they are structured vary enormously and are primarily based on the current rendering and delivery applications a publication uses.

One other type of periodical content warrants note. A particular strength of the Web is its ability to link distributed pieces of content, a power as frequently used in digital periodicals as in other types of Web objects. Such linkages come in many forms: some links are to other content in the publisher’s delivery system, where both the link and its target are under the control of the same organization; others are to independent sources. The latter can be of the casual reference sort (“If you are interested in this, that site over there also has relevant material”); other links to separate systems, however, are integral to the publication (e.g., Web bibliographies or pointers to data in knowledge-bases such as genetics or astrophysical databases). Some links are standard URLs, providing static addresses for specific objects on specific computers. Other links point instead to intermediary systems, capable of finding the current location(s) of the pointed-to object (the Digital Object Identifier, for example[4]). In archiving digital periodicals, it will be important to determine the best way to handle links and the level of responsibility an archive has for maintaining the ability to find independent linked-to objects referenced in archived periodicals.

Organizational Issues

The Open Archival Information System (OAIS) reference model[5] is a powerful abstract model for digital archiving that has informed much contemporary thinking and practice. OAIS defines roles for three players in archiving: creators, archive operators, and end users (see figure 1).

Creators/Depositors

In the case of digital periodicals, "creator" is not a sufficient term, because many players are involved in digital content creation, formatting, distribution, and ownership. A scholarly journal, for instance, can involve any or all of the following:

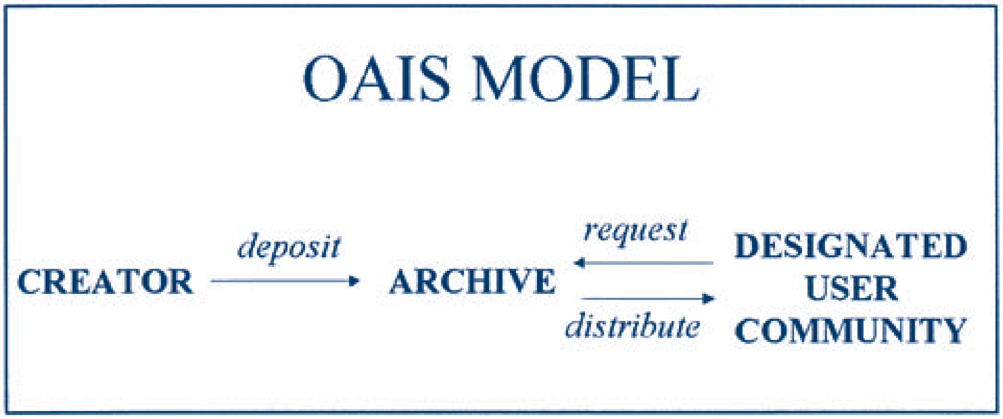
- author(s)
- copyright owner(s) of the included material (e.g., photographs, drawings)
- scholarly society that owns the journal
- publisher responsible for peer review, editing, layout, etc.
- distributor(s) providing online access to the title
- aggregator(s) that includes an article in an online compilation

At least some of these players have a role in "deposit." It may be useful to distinguish among players who have the rights, the motivation, and the appropriate technical manifestation to deposit materials and to cooperate in archiving.

Rights

The deposit of materials into an archive involves questions of ownership and rights: who is legally positioned to provide content to an archive and to negotiate appropriate licenses, if required, for archiving? Because digital periodicals are composed of many separately created pieces, the issue of ownership can be complex. Authors can vary from scholars (who generally, but not always, turn over all copyrights to the periodical owner) to publisher's employees (whose work is automatically owned by

Figure 1. OAIS model of players and roles



the employer) to free-lance writers and illustrators (whose rights vary on the basis of the nature of their contracts). Individual articles can contain separately owned objects, whose owner's rights also vary (the same picture used under the fair-use right of criticism in one periodical requires permission when used in an advertisement in another). The same article can be included in different compilations, for example, in the periodical in which it originally appeared and as an aggregated database, such as LexisNexis or ProQuest. Periodical aggregates, as well as individual titles, could be subject to archiving.

Motivation

The interests of different possible deposit agents vary with the nature of the content, intended audience, and business model associated with specific materials. Some players' concerns are purely short-term. The economic value of some products falls quickly following publication, and the audience served has little interest in anything but today's information. Such players are unlikely to want to invest in archiving or preservation of their content, but they may also have little concern if others want to do so. Other players may believe that their publications have enduring economic value and may therefore be enormously concerned about independent archives holding copies of their content and, if archiving is permitted, about the terms and conditions of access to archived content. Still others, such as scholarly societies and original authors, may want to have their materials preserved and may be willing to invest in that preservation.

Technical manifestation

A number of middlemen are often involved between the owner and the user of periodical content. In the scholarly journal example, the publisher, distributors, and aggregators all play the role of middleman. Each middleman has its own systems, and copies of periodical content contained in each system can vary on the basis of the particular nature and function of those systems. A key consideration in archiving periodical content is the location of an appropriate archival copy: in many cases, the most appropriate copy for archiving may be held by someone other than the owner.

Archive

There is an increasing belief that archiving needs to be the responsibility of institutions for which it is a core mission, rather than an ancillary operation of an organization whose central interest lies elsewhere. Digital archiving will be a technically and organizationally challenging task, and it is unlikely that a large number of institutions will have the motivation, skill, or resources to undertake the long-term archiving of digital periodicals. The great majority of periodical subscribers and readers will, over time, probably rely on a few institutions to provide storage and preservation of periodical content.

Archives are likely to differ in focus. The organization of archiving activity across institutions involves the following important issues.

Collection policy

Each archive must clearly delineate the bounds of its archiving activity. Different institutions may define their scope of responsibility in different terms: by topic, by source of publication (publisher, distributor), by designating selected individually important titles, or by defining samples to be taken across specific literatures. Some level of redundancy is desirable, particularly for titles of potential historical importance. Equally important is the issue of coverage: is an adequate portion of the periodical literature being archived for the use of future generations?

User community

Both the selection of content for archiving and the specifics of archiving and preservation practice are sensitive to the particular user community for which archiving is being done. Different user communities have different requirements as to what is saved, how it is organized and accessed, the technical formats available from the archive (e.g., the writer of popular history needs materials in a form immediately accessible in current technology, the statistical researcher may want data unaltered from the original format), and the technical and support services available from an archive. A key observation of the OAIS model is that archiving activity needs to be designed with an understanding of the specified community being served.

Relationship to depositors

An archive does not automatically have the right to copy and store the publications of any given owner. In some cases, archiving activity may fall under the blanket of copyright deposit. But even then, unless the conditions of archiving are clearly specified in copyright legislation, the owner of archived material may legitimately require a specific license covering the terms of archiving. Given the large number of publishers and owners of digital periodical content, the transactional cost of negotiating archiving agreements will have to be minimized. Among the elements that will help are community agreement on archiving parameters and conventionalized licenses for archiving.

Archiving will come at a noticeable cost. A key issue in the relationship among archives, owners, and users will be the distribution of costs. Some of the major cost elements involved, arranged roughly in order of occurrence, are as follows:

- notification/identification of content to be archived
- creation of an archival version of content
- creation of archiving metadata
- storage, monitoring, and management of the archival collection
- preservation of archived content
- service to users

These costs can be distributed to the parties in various patterns. One might wonder whether the arrangement above suggests a model of costs distributed to owners, archives, and users as one moves down the list.

Users

The OAIS model suggests that archiving is done to meet the needs of a specified user community. User communities vary not only with the nature of publications but also with the passage of time. While some periodical content continue to be used primarily as originally intended (e.g., "how to" literature, works describing events or scientific observation, literary or critical works), other kinds of uses become common over time. The historian of science or the analyst of trends uses material in ways that are different from those of the original audience of a publication.

The owners of archived content can be expected to be quite sensitive to the following two primary questions about users.

Who can access archived content?

At least while content is not in the public domain and continues to have economic value, many owners will want to limit the population that can access the archive. For example, access could be limited to

- auditors of the archive
- users with subscriptions to the archived content
- users within the walls of the archive
- users within the institutional bounds of the archive
- users making specific types of use (e.g., the archived objects could be made available to the historian of science, but not to the researcher in a pharmaceutical company)

When can content be accessed?

Many archiving discussions revolve around the idea of "trigger events," that is, conditions under which archived content becomes more widely available. A trigger event may occur, for example, when

- a given periodical is no longer accessible on-line;
- a specified time has elapsed after initial publication (this is the current policy of PubMed Central, an archiving initiative of the National Library of Medicine, which calls for deposited content to be openly available no more than one year after publication[6])
- a title changes hands

Trigger events vary from owner to owner and from publication to publication. It is interesting to note the contrasting business models in today's periodical environment that are likely to influence a time-based trigger event. Some publishers charge significant subscription fees for current issues but offer free access to back files.[7] Others, including some newspapers and magazines, provide free access to current issues but charge for access to back files. Still other business models may yet emerge.

Technical Issues

Many technical issues involved in periodical archiving will have to be faced by the various players (owners, archives, and users). Of key importance are the following.

Preserve Look, Feel, and Function?

Digital periodicals as perceived by users are composed of a complex of elements: the digital content itself, the display software used to render that content, and a variety of system functions provided by the Web site delivering the periodical. What parts of this complex should be archived? There are a number of questions raised if one were to consider archiving more than the raw content (e.g., the words, pictures, or sounds) of the publication). For example:

- Archive display formats or underlying data? Formats used for ready rendering on the Web frequently differ from the format of content in the underlying publishing system. A publisher may have text marked up in SGML or XML in its asset management system, but deliver HTML or PDF formats, or both, to users today. HTML or PDF may well be easier formats to use if one wants to faithfully recreate the original look of a publication, but many believe they will present archiving problems because the rendering software will certainly be superceded over time. The SGML or XML marked-up text will be less sensitive to technological change, but ensuring the ability to re-render it as it was originally displayed will be technically complex.[8]
- Archive periodical sites? Digital periodicals are delivered through Web sites that frequently offer a wide variety of functions, such as specific organization of content, search facilities, order forms, and communication facilities (to e-mail the editor or participate in a threaded discussion, for example). Archiving entire Web sites with all associated functionality will introduce a significant additional level of complexity beyond archiving periodical content.
- Use emulation as a preservation strategy? Emulation has been proposed by some as a means of preserving the original look and feel of digital objects. In this strategy, an archive stores not only the digital objects but also the software originally used for rendering. Because the software will depend on a specific technical environment (hardware, other software), the archive must build or acquire software capable of emulating that original technical environment, thus permitting obsolete software to run in new environments. Emulation as a preservation technique is highly controversial, with opinions about its practicality differing widely.[9]

What Content Is Archived?

Most people initially assume that periodical archiving is concerned only with the content of articles. While articles are the intellectual core of periodicals, digital periodicals contain many other kinds of information. Examples of content commonly found in scholarly journals include the following:

- editorial board
- rights and usage terms
- copyright statement
- journal description
- advertisements
- reprint information
- editorials
- events lists
- errata
- conference announcements
- various sorts of digital files related to individual articles (data sets, images, tables, videos, models)

Which of these need to be archived and preserved for the future? Some of these types of materials will pose problems for publishers. Not all of these items are controlled in publishers' asset- management systems. Some are treated as ephemeral "masthead" information and simply handled as Web site content. When such information changes, the site is updated and earlier information is lost. For example, few if any scholarly e-journals provide a list of who was on the editorial board for an issue published a year or two ago. Deciding what of all that is seen on periodical sites today should be archived and maintained will require careful consideration by archives, publishers, and users.

Should Content Be Normalized?

The variety of formats of digital objects in an archive will affect the cost and complexity of operation. To control such complexity and cost, an archive may want to normalize deposited objects into a set of preferred formats whenever possible. Such normalization can happen at two levels:

1. File formats: An archive may prefer to store all raster images in TIFF, for instance, and convert JPEG or GIF images into that format. Controlling the number of file formats will reduce the complexity of format monitoring and migration.
2. Document formats: Many publishers encode article content in SGML or XML (or plan to do so soon). Most publishers create their own DTD (or modify an existing DTD) to suit their specific needs and delivery platforms. An archive may choose

> to normalize all such marked-up documents into a common DTD, reducing the complexity of documentation, migration, and interface software.[10]

Normalization and translation always involve the risk of information loss. Archiving may well involve a difficult trade-off between information loss and reduced complexity and cost of operation.

Should a Standardized Ingest Format Be Developed?

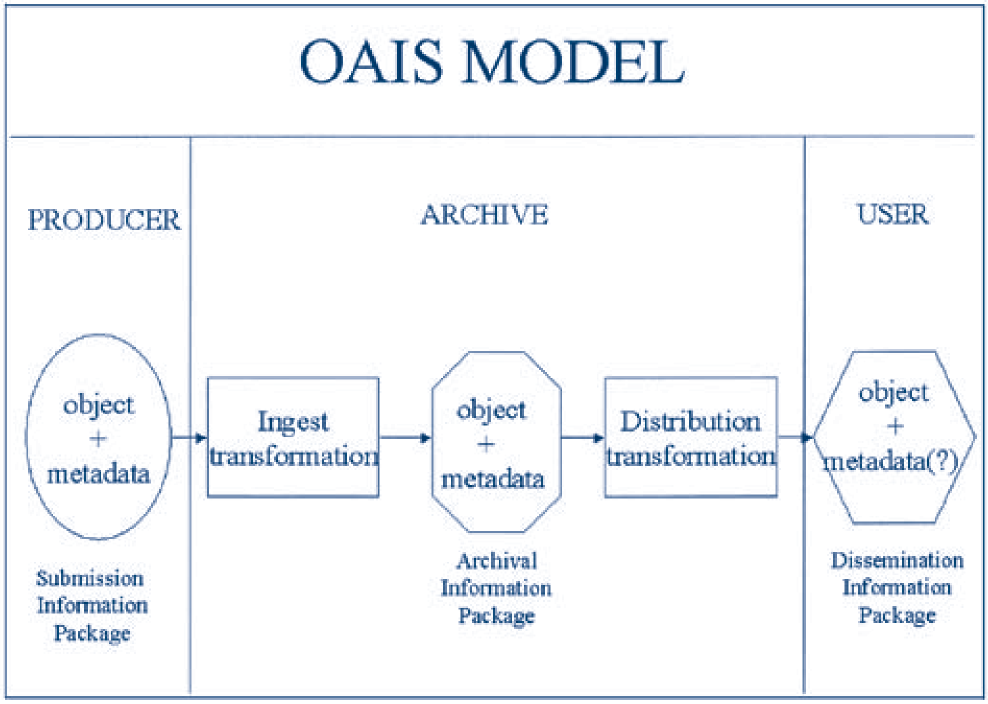
The OAIS model uses the concept of "information packages," that is, bundles of data objects and metadata about the objects that are the unit of deposit, storage, and distribution by an archive. The model allows transformation of objects as they move from one type of package to another (see figure 2).

If, as expected, any given publisher is depositing content into a number of different archives, and any given archive is accepting deposits from a number of different publishers, standardizing the format of submission information packages may reduce operational cost and complexity for both communities (although at the cost of devising and maintaining such a standard).

Preserve Usable Objects or Just Bits?

A key element in digital preservation is maintaining the usability of digital objects in current delivery technology as the environment changes over time. This process is usually assumed to be one of "format migration," that is, the transformation of objects from obsolete to current formats, although it can also be carried out through emulation, that is, maintaining current programs capable of emulating older technology

Figure 2. Information Packages in the OAIS model


OAIS MODEL
PRODUCER
ARCHIVE
USER
object + metadata
Ingest transformation
object + metadata
Distribution transformation
object + metadata(?)
Submission Information Package
Archival Information Package
Dissemination Information Package

and thus rendering obsolete formats. However the process is accomplished, the cost of preservation will be sensitive to the number and types of formats in an archive.

Digital periodicals can contain a wide range of technical formats. Whether it will be practical for archives to maintain current usability for such a diverse range of formats is far from clear. It is possible that archives will need to differentiate between formats where usability is maintained and those for which the archive only ensures that the bits are maintained as deposited and that their documentation is kept usable to support future "digital archaeologists."

Summary

There is tremendous variety in the players, content, and technology that will naturally shape any program to archive digital periodicals and make program planning difficult. However, plan we must, or face losing over time a significant portion of the formal literature of our time. If that happens, future generations will be left with a much poorer understanding of our age than we have of our nineteenth- and twentieth-century ancestors.

Further Reading

Council on Library and Information Resources, Digital Library Federation, and Coalition for Networked Information. 1999. Minimum Criteria for an Archival Repository of Digital Scholarly Journals. Available at: http://www.diglib.org/preserve/criteria.htm.

Based on the Open Archival Information System model, these criteria were developed in a series of meetings involving libraries and journal publishers.

Flecker, Dale. 2001. Preserving Scholarly E-Journals. *D-Lib Magazine* 7(9) (September). Available at: http://www.dlib.org/dlib/september01/flecker/09flecker.html.

This article describes an initiative of The Andrew W. Mellon Foundation to create several demonstration archives for scholarly digital journals, and enumerates some difficult issues raised in planning such archives.

Mark Bide and Associates. 2000. Standards for Electronic Publishing: An Overview. Available at: http://www.kb.nl/coop/nedlib/results/e-publishingstandards.pdf.

Commissioned by the Nedlib project (see below), this report reviews the current state of practice in using standardized formats for digital books and periodicals.

Nedlib Web site. Available at: http://www.kb.nl/coop/nedlib/.

Nedlib is a project of the European Community involving a number of national libraries. It is intended to describe a framework for electronic copyright deposit and archiving.

Springer-Verlag. Springer-Verlag joins with international library community in creating electronic information archive for mathematics. Press release, July 23, 2001. Available at: http://www.library.yale.edu/~llicense/ListArchives/0107/msg00088.html.

This notice describes an international effort to archive the literature of a specific field, mathematics.

Notes

1. It is also worth noting that the analysis in this paper is informed above all by work in one specific domain, the scholarly journal.

2. The back-up and mirroring systems used for many large-scale publications represent only a partial form of redundancy. While offering good protection against accidents and hardware failure at a specific physical location, they still leave content vulnerable to institutional failure, changes in institutional policy, conscious "amendment" (think of the Stalinist removal from photographs of those who had fallen from grace), systematic software errors, and the like. Effective redundancy requires that independent players hold copies in separate political jurisdictions, and in differing technical environments, removing the sensitivity to destruction by any single element or agency.

3. For information about Unicode, see: http://www.unicode.org/.

4. For information about the Digital Object Identifier, see: http://www.doi.org/.

5. For a general introduction to the Open Archival Information System model, see http://www.oclc.org/research/publications/newsletter/repubs/lavoie243/. For a detailed description of the model, see: http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf.

6. For information about the PubMed Central policy, see: http://www.pubmedcentral.nih.gov/about/newoption.html. There is a great deal of discussion in the scientific community about whether all scientific research literature should become freely available after a defined interval. The intent is to provide the publisher with a period of exclusive use for revenue generation. After this period, the literature would be open for use by the entire scientific community. A leading initiative in this area is the Public Library of Science proposal, described at: http://www.publiclibraryofscience.org/.

7. For example, see: http://www.highwire.org/lists/freeart.dtl.

8. Note that the "original" rendering may in fact be fleeting, as the original publisher may choose to alter and improve display of publications over time.

9. For a discussion of emulation for preservation, see the following Web sites: http://www.clir.org/pubs/reports/rothenberg/contents.html and http://www.dlib.org/dlib/october00/granger/10granger.html.

10. As part of a journal archiving project at Harvard, a consultant is examining the feasibility of creating an "archival e-journal DTD," which would be a preferred format for article deposit.

E-Books and the Challenge of Preservation

FRANK ROMANO
Rochester Institute of Technology

Introduction

The concept of electronic publishing was first articulated by Vannevar Bush of the Massachusetts Institute of Technology (MIT) in the seminal 1945 article "As We May Think." In 1991, Apple Computer introduced *Jurassic Park* as an electronic book for its Powerbook 100 laptop using the Adobe Acrobat portable document format (PDF). In 1998, the Rocket E-book was introduced, and in 1999, Simon & Schuster and Stephen King published an electronic novella that could be read on any Internet browser on virtually any computer, or downloaded to certain e-book devices. For the foreseeable future, most e-publishing will involve scientific, technical, professional, and academic information, as well as some original fiction. Librarians and others involved in digital asset management will have to preserve at least some of this material for future reference, since it is expected that original works will be created and many of these may exist only in electronic form. E-books are not a historical artifact or anomaly, but a new form of content conveyance. Growth, while steady, may be slow because of competing technical standards, digital rights management, definitional issues, and restructuring within traditional publishing, as creators, existing publishing houses, and software companies position and reposition themselves in a changing market. A critical and perhaps underestimated set of issues concerns user acceptance.

The trend toward electronic publishing has been based on factors such as the following:

- technological advances that provide increased computing functionality at lower cost (generally summarized under the name Moore's Law)

- the development of new channels of information distribution (Intranet and Internet)
- the desire to reduce costs by eliminating paper, printing, and physical storage
- the ability to search electronic files efficiently and retrieve information quickly
- the ability to reuse information in other documents and other formats (with appropriate content rights management)
- the acceptance of reading on-screen by growing segments of the population
- the convergence of text, imagery, audio, video, animation, and interactivity in new kinds of documents
- the ability of virtually anyone to become his or her own publisher
- the immediacy of content acquisition through electronic transactions and data downloading
- the demand for storage space in libraries

Since the advent of disc- and tape-based digital storage in the 1960s, we have seen the evolution and proliferation of more than 200 different data storage formats—from large- and small-diameter fixed discs, to flexible diskettes of every size, to compact and video discs. During this time, media have decreased in size and increased in storage capacity, from 1 kilobyte of data to 40 gigabytes of data, with the first terabyte discs imminent. No single format has existed for more than a decade, which has necessitated the recording and rerecording of information on new media to allow access by current computing systems. This trend has also affected the entertainment industry as it evolved from records, to tapes in cassettes and cartridges, to compact discs (CDs) and now to digital video discs (DVDs).

At the Rochester Institute of Technology, files stored on 8-inch flexible diskettes from word processors of the 1970s are unreadable—not because of their condition but because readers for that medium are unavailable. Forty-four-megabyte Syquest discs from the 1990s are about to suffer the same fate. Libraries and information repositories face a continuing challenge in maintaining files on currently supported storage hardware and media and in currently supported file formats for currently supported operating systems that require structured data organization.

Definitions

An electronic book, or e-book, is the presentation of electronic files on digital displays. Although the term “e-book” implies book-oriented information, other content can also be displayed on such devices. Static text and images are typically displayed, but moving imagery and audio are also presented. E-book files can be provided as recorded units (discs) or downloaded from digital repositories (including Web sites) to desktop computer monitors, laptop screens, portable digital assistants (PDAs or Palm™-type devices), cell phones with expanded displays, pocket pagers, or dedicated digital reading devices (also currently called “e-books”).

The e-book production cycle begins when an author creates an original work and submits it to a publisher. The publisher converts the work to one or more e-book formats and employs rights-management encryption to electronically lock the file and generate a unique decoding key. (Initially, a 40-bit encryption was used. The U.S. government now permits U.S.-only versions with 128-bit encryption, which improves security.) An e-book distributor (who may be different than the publisher) manages the protected file. The e-book publisher or distributor transfers the work to an e-book retailer, who sells the protected e-book online and offers buyers a "key" to decrypt and read the work. A buyer connects with a retailer's Web site and purchases the work, after unlocking the file with the digital rights key and downloading it to read on an e-book reading device. Some of the digital rights solutions include Adobe PDF Merchant, WebBuy, Xerox ContentGuard, Reciprocal.com, SoftLock, netLibrary, InterTrust MetaTrust Utility, LockStream.com, and others. (Rights issues are discussed in detail on pp. 49–51.)

The word "e-book" is actually a misnomer. The device can display magazine content (e-magazine) and newspaper content (e-newspaper), as well as electronic directories, catalogs, and other material. The display device is independent of the content. However, a distinguishing characteristic of books, magazines, and newspapers is the size of the page—all must adjust to the device's screen size, which is currently about the same as that of the page of a standard hardcover book.

A Web site is a collection of HTML-coded files and other files (image, audio, video) in computer code that are displayed on a screen using a browser application program. The browser (e.g., Netscape Navigator or Internet Explorer) translates the coded data into displayable typographic and image elements and presents them to the viewer. An important aspect of such sites is the ability to click on defined elements that then automatically display other Web sites ("hyperlinks"). A computer linked to the Internet functions like an e-book does and thus inherits many of the challenges associated with long-term use and preservation of Web sites.

Consider the problem of how to identify and find a Web site. Web sites have addresses so that viewers can connect from one to another. Such addresses have been used as bibliographic references or identifiers. After only a few years, those address may no longer be active. This presents another challenge to the preservation of information, because it is not expected that most e-books will be delivered via media (discs, for example) but rather through connections to the Internet or proprietary sources—wired or wireless. Thus, the content may be unfindable or unavailable for downloading. While it may be expected that libraries and other information repositories might be backups for Web-based content acquisition, libraries and information repositories will have to store such information on some form of storage media, and, unless standards evolve, they may require a plethora of different reading devices. An alternative scenario would have libraries serve as portals to any number of commercial sites. However, the likelihood of long-term preservation by commercial enterprises may not be as assured as is preservation by certain libraries.

Thus, there are three related challenges centering on (1) the location of the stored information, (2) the organization storing the information and its long-term viability and commitment to preservation, and (3) technical issues. In addition, there are questions of digital rights management; possible definitions of new "artifacts," including the notion of an e-book itself; user acceptance; and a reconfiguration of interests and equities among authors, publishers, and software firms.

The Challenge of Preservation

Preservation of electronic content will be necessary for practical purposes (i.e., for downloading current material) as well as for historical purposes. There are a number of scenarios for the delivery of this information.

- The e-book device is connected to another computer that is linked to the Internet. The user goes to a specific Web site and selects the titles required. The Web site could be that of the e-book producer, a portal that represents several publishers, a single publisher, or an academic or corporate site.
- The e-book device has a built-in modem and is connected to the Internet by a phone line directly for downloading.
- The e-book device is connected through a kiosk at bookstores, libraries, airports, or other locations for downloading.
- The e-book connects by wireless modem to the selected Web site or other location.

In every case, the e-book "title" is stored on a remote storage system and is then routed to the e-book directly or to a computer. No single data location of all e-book files will exist, and mergers and personnel changes at the hosting site may affect the long-term storage of the information. A company could decide, for example, to drop certain titles, or it could go out of business. Thus, libraries and other data repositories hold the responsibility of long-term preservation.

Computer operating systems are usually aligned to structured storage systems that record coded data. Over time, all of these aspects of the systems may change:

- recording medium (e.g., magnetic tape, disc)
- operating system (e.g., Windows)
- storage format (e.g., binary, ASCII, sound, video)
- data coding system (e.g., HTML, XML)
- metadata (e.g., bibliographic or stylistic encoding)

Dynamic Preservation

The storage of digital data will require a dynamic form of preservation, and a new definition of "archival" may have to be developed. The concept of long-term storage of a paper- or photographic-based item that remains unchanged over time may not be applicable with electronic publishing. Instead, the information will have to be rerecorded on new media to be used with existing file formats and computer operating systems as storage media degrade and systems, formats, and encoding systems evolve.

There are programs that convert from one encoding system to another. Over time, these programs will become more reliable and allow data to be reformatted to the current standard approach. But the conversion will have to take place in order to keep the information in a "current" format. Usually there is a two-year transition between one form of storage and its successor. This is both a management and a technical issue and tracks the organizational issues—the permanence and commitment of the archiving organization—cited in the previous section.

Technology Issues

The size of the page—or the screen—is the defining property of e-books. This was fundamentally enabled by the "portable-monitor" (higher-definition liquid-crystal display [LCD] screens). Capabilities vary with price, which ranges from around $150 to $600. E-books such as the Rocket Book (now the RCA Gemstar) attempt to emulate what a typical reader or student would do with a real book: highlight text, bookmark pages, browse indexes, or write notes in the margins. Most e-books (ranging from a pocket-size Palm Pilot to a device roughly half the size of a laptop computer) are capable of downloading and storing text and displaying it in a prescribed format that is intended to mimic that of a typical printed book. The text is usually displayed one screenful at a time and in most models is advanced or regressed a screenful at a time with arrow buttons. Some models do not have page numbers; in this case, a screenful of text may be considered a page. Page orientation can be adjusted with some brands. Most electronic books also have "advance" features that allow users to move quickly forward or backward as if paging through a printed book. The books are battery powered but also come with electrical adapters. Rechargeable batteries can last from 10 to 40 hours, depending on the brand and whether backlighting is used.

Screen Issues

The size and resolution capabilities of e-books vary. They can support text as well as black-and-white images such as graphs, line art, and newspaper-resolution photos. Gray-scale images are not supported with most brands. All the current e-books are black and white; there are few color models. Within two years, most models will display gray-scale images and color and also play sound and video. Most e-books come with proprietary software that is used to transfer data to and from the e-book as well as to allow downloading from Internet-based or proprietary services.

The most significant advance toward a paperless world will be portable displays —lightweight, rugged, operating for hours using lightweight batteries, with high resolution and contrast. In the late 1980s, LCDs were incorporated in the first laptop computers, and today the typical laptop computer includes a 12- to 14-inch, full-color LCD with good resolution. LCD-based flat-panel displays are smaller and lighter, use less power, and discharge fewer electromagnetic emissions than do their cathode ray tube (CRT) counterparts. There are experiments under way at Xerox Palo Alto Research Center (in cooperation with 3M) and E Ink (a spin-off from the MIT Media Lab in partnership with Lucent Technologies) and other variations on the notion of digital ink, digital paper, ultra-thin screens, flexible displays, and such.

Standards Issues

There are a number of issues and organizations involved in developing standards. These involve markup languages, identification, and metadata as well as hardware and software standards.

The hypertext markup language (HTML) and portable document format (PDF) standards continue as dominant document formats on the Internet, but are not necessarily perfect standards for information delivered on hand-held devices such as e-books. HTML displays can have difficulty with consistency and Acrobat displays the equivalent of printed pages, which may be oversized for most small devices. Both of these limitations are being addressed: HTML is metamorphosing into extensible markup language (XML) to allow more consistent reformatting on different screens, and Adobe is integrating PDF and such reformatting into future versions of PDF. Microsoft has developed Clear Type font technology for clearer, more "paperlike" reading and has announced a standard text format and operating system for Microsoft Reader. Adobe has just released its version of a more readable screen font technology called CoolType.

A PDF file is truly a portable document. It can be generated from just about any application and keeps all typographic formatting, graphics, layout, and page integrity intact. Because the PDF embeds fonts, the recipient need not have the fonts that were used by the document creator. Graphics are compressed, which allows PDF files to be very small for transmission over networks. The reader software runs on most computers and is free—downloaded from Adobe's Web site.

In 1998, the National Institute for Standards and Technology (NIST) of the U.S. Department of Commerce formed the Open E-Book Standards Committee (OEBSC) to promote a standard e-book format. The Open E-Book Publication Structure, developed by OEBSC, defines the format for content converted from print to electronic form. The Electronic Book Exchange (EBX) Working Group is establishing copyright protection and distribution standards. The Open eBook Forum (OeBF) is an international, nonprofit trade organization whose mission is to promote the development of the e-publishing market. The Open eBook Authoring Group, made up of the major e-book reader manufacturers, a few large publishers, and Microsoft, among others, released the first Open eBook Specification (OEB 1.0) in September 1999—a specification based on XML. In January 2001, the Open eBook Publication Structure Specification Version 1.01 was placed before the OeBF membership for comment. OEB 1.01 uses HTML semantics, but XML-based syntaxes.

Other standards initiatives include the Digital Audio-Based Information System (DAISY) initiative, the Text Encoding Initiative (TEI) Consortium, NISO W3C, DocBook, the International Publishers Association, MPEG, the U.S. Copyright Office, the international digital object identifier (DOI) foundation, and EDItEUR.

The Open Ebook Standards Project, led by the Association of American Publishers (AAP), several leading publishers, and Andersen Consulting (now Accenture), released the results of an intensive effort to establish recommendations and voluntary standards (AAP 2000a, b, c). Experts have been working with AAP to develop standards for numbering and metadata, and to identify publisher requirements for digital rights management, three areas critical to the growth of the market. The new standards specify a numbering system based on the Digital Object Identifier, an internationally supported system suited for identifying digital content and discovering it through network services. The numbering recommendations allow for identification of e-books in multiple formats and facilitate the sale of parts of e-books, and they also work with existing systems such as the ISBN to allow publishers to migrate to the new system.

The metadata standard has extended ONIX, the existing international publishing standard for content metadata, to include the information needed to support the new numbering system and e-book-specific data. With ONIX, publishers will be able to provide their metadata to (r)e-tailers, conversion houses, and digital rights partners. Indexing of the metadata will make e-books easier to find in online catalogs. AAP also released a comprehensive description of digital rights management (DRM) features needed to enable the variety of new products and business models publishers want to offer.

There are numerous proprietary software solutions being offered to translate digital e-book files for the many competing reader platforms. Most solutions incorporate security features to protect copyright owners (that is, the file cannot be printed or copied). It may be that reading devices may display some of all of these formats, but one or two probably will become clear standards. Publishers have already restricted their market through the use of a reading device. Unless a very inexpensive reader is developed and becomes universally available, this market cannot evolve. The information for these readers must also be standardized and pervasive. It is not that we do not have standards—we may have too many of them.

User Acceptance Issues

The AAP teamed with Andersen Consulting to evaluate the market for e-books and to define the basis of its publisher members entry into e-book publishing. In a study entitled “Reading in the New Millennium, A Bright Future for E-Book Publishing,” Andersen projected the e-book market at $2.3 billion by 2005—10 percent of the estimated $21.9-billion consumer book market in 2005. This study also highlights the importance of open standards to the success of electronic publishing because “it’s easy for consumers: any book, any source, any device” (Andersen 2000).

In December 2000, Forrester Research, an Internet research firm based in Cambridge, Massachusetts, released a report with the following projections:

- Slow growth is expected for both e-books and e-book reader devices.
- There will be strong sales for on-demand custom-printed trade books and digitized textbooks.
- In five years, 17.5 percent of publishing industry revenues ($7.8 billion) will come from the digital delivery of custom-printed books, textbooks, and e-books. Of this amount, only $251 million will come from e-books for e-book devices.
- As a result of the Web's distribution advantages, publishers will create a new publishing model called "multichannel publishing," requiring publishers to manage all of their content from a single, comprehensive repository containing modular book content and structure. (O'Brien 2000)

Virtually all recent studies predict a slow but continuous growth in the e-book market.

Publisher Issues

Publishers are implementing a range of strategies, partnerships, and experiments with delivery and packaging. AOL Time Warner Trade Publishing was one of the first traditional publishing houses to launch a digital division with the creation of ipublish.com. Random House and Simon & Schuster have also created electronic divisions. Barnes & Noble established an online e-book store, and Amazon.com has also entered the market. Electronic publisher MightyWords signed distribution partners to sell its titles on Fatbrain.com and Barnesandnoble.com; in addition, consumers may browse, purchase, and download works at Adobe.com and other Web sites.

In 1995, book publishers produced thousands of multimedia computer CDs with interactive features, pictures, and sounds, but consumers did not accept the new electronic works. Personal computers were not as pervasive; technical standards caused innumerable problems running the programs; and few personal computers had CD-ROM drives. Multimedia has grown into a significant market as standards evolved and the base of computer users expanded. Major book publishers, technology companies, online booksellers, and new e-book middlemen are investing in the future market of digital books.

Authors may see electronic books as a way to free themselves from dependence on publishers and to sell books directly to consumers. Publishers may see an opportunity to eliminate printers and bookstores. Online booksellers are moving into the publishers' business, printing digitized books themselves and selling their own electronic editions. Startup companies sell the contents of books through digital archives of thousands of books and periodicals available on-line, liberated from the constraints of time and shelf space.

Publishers now see e-books as incremental sales to computer-savvy adults and the next generation of readers. A publisher's ultimate responsibility is to get the work to the largest-possible audience and the Internet has that potential. But no one knows what an electronic book is worth. Some publishers are setting prices for e-books just below those of their printed equivalents, but others charge much less. Random House said that it would split equally with authors the wholesale revenue from selling or licensing electronic books, raising the author's share of the list price from 15 percent to 25 percent. Random House invested in Xlibris, a digital publisher that claims to issue more books in a year than Random House does. After the success of Stephen King's e-novella, Bertelsmann, Simon & Schuster, and AOL Time Warner's book division approached agents for digital rights.

Digital publishing presents an opportunity for authors and publishers to develop a much closer connection to consumers than they have in the past. There will still be retailers, but certainly the middleman component may be smaller. Some publishers are already selling digital books directly to consumers as customized editions with modular contents, especially in the educational market. McGraw-Hill's Primis Custom Publishing division has a Web site that lets instructors select chapters and excerpts from a digital archive to build their own personalized electronic volumes. Instructors order directly and bypass campus bookstores.

Random House's Modern Library Classics division sells electronic editions of its books directly to readers through links to literary Web sites such as those devoted to William Shakespeare or Jonathan Swift. Time Warner sells e-books through links to its own Web site. Barnesandnoble.com publishes and prints its own digital books. Barnes & Noble and Barnesandnoble.com have invested in several digital publishing and bookselling startup companies, including Fatbrain.com, iUniverse, and MightyWords.com. The company has installed print-on-demand systems in its warehouses so that it can begin printing and binding copies of books available from publishers as digital files. Book wholesaler Ingram Book Group's Lightning Source pioneered print-on-demand for runs as low as one book.

Amazon.com offers a distribution channel for authors who want to self-publish either print or electronic editions. Startup companies are also building an alternative sales channel for the contents of digital books, as part of large online archives that let readers search through texts as well as browse their titles. Each of the main e-book contenders is pursuing a different strategy and competing for publishers' digital books.

NetLibrary sells electronic books to libraries via online access to the digital version of the book on their computer servers. Users can search the contents of books in the online collection, but they cannot copy or print the books. Public and university libraries and some corporations are now customers. Questia and Ebrary, as well as other e-publishers, are negotiating with publishers and authors to enlarge their collections. Questia sells access to an archive of digital books for a subscription fee, with a variety of research tools, including links connecting footnotes in one book to text in another. Random House, McGraw-Hill, and Pearson's Viking-Penguin have invested in Ebrary, which lets readers search and browse freely through digital books and magazines, but charges a fee to print pages, copy text, or download content.

Digital Reader Issues

The future of digital publishing will also be shaped by the competition among three technology companies hoping to set the standards for publishing and reading books on screens. Microsoft, Adobe Systems, and Gemstar–TV Guide International are working to convince publishers and readers that their format is the most secure from copying, convenient to use, and easy on the eyes. Microsoft and Adobe Systems produce competing software programs intended to make reading on a screen easier on the eyes, and both have announced alliances intended to strengthen their respective positions.

Gemstar's format is used on portable appliances, such as the Rocket e-book, instead of a laptop or desktop computer. Adobe Systems has by far the largest share of the digital publishing software market. Customers have downloaded more than 180 million free copies of Acrobat Reader software for reading and printing digital documents. Gemstar holds patents on the technology to read digital books on specialized hand-held devices. Gemstar's latest generation, built under the RCA brand by Thomson Multimedia, is priced at $300. Gemstar's system avoids both personal computers and the Internet. Online bookstores sell electronic books for Gemstar's format, but to download the digital texts, consumers must plug their devices into phone lines and dial directly into Gemstar's computer servers. Users of the devices can only store and retrieve their books on Gemstar's server. Devices that apply Gemstar's electronic book patents could be used as personal organizers, wireless pagers and phones, and generalized portable entertainment devices for text, video and sound, making the habit of reading an entry into the PDA and multimedia arena.

Microsoft and Amazon.com opened an electronic bookstore that distributes free copies of Microsoft's Reader software. Amazon.com sells electronic books for a variety of formats, including Adobe's. Microsoft makes no money from its Reader software but does receive a small commission on the sale of electronic books in its software format. Microsoft started a similar cooperative marketing venture with Barnesandnoble.com with the release of a new version of its Reader software.

On-demand Printing

Publishers are applying print-on-demand methods, and such printing is starting to change their business. Xerox, IBM, and others now sell machines that in minutes can churn out single, bound copies of paperback or even hardcover books. The output is virtually indistinguishable from that of traditional printing presses.

In traditional printing, hundreds of copies must be produced to make a print-run cost-effective. This constraint does not hold for on-demand printing; as a result, some low-selling books that would have passed out of print are staying in print longer, and a few books that might not have found publishers now have done so. The Perseus Books Group installed print-on-demand equipment in its warehouse near Boulder, Colorado, to print slow-selling titles in small batches instead of letting them fall out of print. The National Academy Press in Washington, D.C., did the same. New printing technology helps fulfill demand for special-interest titles created partly by online bookstores. Some publishers order print-on-demand editions of some of their books through Ingram's Lightning Source digital publishing division, and the bookseller Barnes & Noble has installed machines in its warehouses to print books on demand.

The early indications are that electronic books are most likely to take off at the two extremes of the book market: with readers of popular novels, fiction such as romances and science fiction, and with readers of educational and business texts.

E-book Publishing

The term "e-book publisher" refers to a business in which a provider enables authors to publish books through an online service. An author submits a manuscript, and it is published and printed as a book. A search of the Internet reveals more than 100 e-publishers, most providing books in electronic form for on-screen reading using the computer's browser or a PDF viewer. A sampling of e-publishers is listed in figure 3.

Stephen Riggio, vice-chairman of Barnesandnoble.com, has said, "You will see—very, very soon—authors become publishers. You will see publishers become booksellers. You will see booksellers become publishers, and you will see authors become booksellers." With the advent of e-publishing, book industry classifications are an anachronism (Pimm 2000).

Rights, Information Security, and Privacy Issues

Replication and intellectual property risks exist because of the relative ease with which digital data can be copied, modified, and disseminated. An important industry concern is that digital content will emulate digital music and circulate free over the

Figure 3. Sampling of e-book publishers

1st Books	www.1stbooks.com
Artemis Books	www.artemispress.com
Books Just Books	www.booksjustbooks.com
Books Onscreen	www.booksonscreen.com
BookSurge	www.booksurge.com
Digitz	www.digitz.net
Dissertation	www.dissertation.com
EBrary	www.ebrary.com
ElectricPress	www.electricpress.com
GreatUnpublished	www.greatunpublished.com
Hard Shell Word Factory	www.hardshell.com
iUniverse	www.iuniverse.com
Lightning Source	www.lightningsource.com
Universal Publishers	www.upublish.com
Zeus Publications	www.zeus-publications.com

Internet. Technology companies are positioned to insert themselves into digital publishing as electronic wholesalers, taking the place occupied by distributors of traditional books. They provide protection from copying, along with software and services to store and transmit digital books, in exchange for a percentage of revenue. These systems typically require four elements:

1. authentication of transmissions and messages to determine whether the originator is authentic, or that the recipient is eligible to receive the information
2. data integrity checks to determine that the data are unchanged from their original source
3. certification that the sender of data has delivered the data and that the receiver has received it, with evidence of the sender's identity
4. confidentiality to ensure that information can be read only by authorized entities

In the quest for security, publishers may be restricting growth of this new market. Let us use printed books as an example. The purchaser reads a book and passes it on to another reader, or sells it to a used-book store, which then sells it again. (Many of us would not have been able to afford college without this system.) Although the publisher does not receive revenue from these subsequent uses or sales, the reader may develop an affinity for the author or subject, and this may stimulate future sales. Magazines are routinely passed around. Publication pages are often copied for distribution. In effect, we have had the "Napsterization" of the publishing market since printing was invented. But this practice may now be upset. Readers of e-publications who wish to save issues for future reference may not be able to do so (the archives of *The New York Times* and *The Washington Post,* for example, charge for access) and may find that the e-book readers do not have external storage.

From the publishers' and authors' points of view, there is cause for concern. Stephen King's *Riding the Bullet* was sold exclusively on the Internet. After 48 hours, *Riding the Bullet* sold more than 500,000 downloadable copies worldwide, at a cost of $2.50 per copy. Although many initial orders were delivered in free promotions, the financial implications of King's foray into e-books are still staggering. It took fewer than two days to sell 500,000 copies without printing, shipping, storage, wholesalers and distribution middlemen, or other traditional publisher costs. However, within those same 48 hours, pirated copies were on the network.

The report *eBooks: Publishing's Next Wave or Just a Ripple?* from TrendWatch Cahners (2001), makes an important point about balancing security and distribution:

> *Periodical publishers have an interesting problem with regard to digital rights management, and that is they want to protect their content, but advertising rates in periodicals is in large part based on "pass along" copies. For example, most ad rates for large consumer publications are premised on the assumption that a single copy is passed along to five other people. If you secure a digital version of that publication, you'll ensure that someone*

pays for it, but you'll also prevent them from passing it along. How do you determine your advertising rates based on that?

Cracking the Code

The Russian firm Elcomsoft has released Advanced eBook Processor, software that enables users to convert copy-protected e-books into plain-vanilla PDF documents that can be printed, copied, and distributed easily. This software company received a cease and desist order from Adobe Systems, and had its Web site removed from the Internet. Adobe says that its e-book software copy protection is not applied by the end user but by the copyright holder. The Russian programmer was imprisoned and eventually released—a release supported by Adobe. Publishers are fearful of e-book piracy and of the thought that books could be swapped like MP3 files over the Internet. Adobe must demonstrate a secure option or it will lose the support of major publishers. But Elcomsoft also showed that it could break Microsoft protection systems. Many feel it is better to show the vulnerability of such systems in an open forum than to drive it underground. For the Russian programmer, it was not a case of hacking, but a mathematical puzzle to be solved. This reflects a tension between the values of the research community and those of the commercial community. It is not clear how the conflict will be resolved.

What Is a Book?

Why are e-book rights treated differently than printed-book rights? In the case of *Random House v. RosettaBooks*, Judge Sydney H. Stein summarized the complex issues of the trial in one statement: "Show me why an e-book is a book." The result of the ensuing argument and debate was a ruling that essentially defined e-books as a new medium of communication, like audio books. But what happens when sophisticated software converts the e-book text to spoken words with the cadence and pronunciation of Anthony Hopkins? Is this analogous to the Kurzweil Optical Character Readers of the 1970s, which scanned printed books into words and then "spoke" them to the blind with a voice synthesizer?

There is an interesting privacy issue in that book buyers (at least those who pay in cash) are generally anonymous. Amazon attracted negative publicity when it used an individual's book-buying data for promotion purposes. In many cases, e-books will be sold only to a specific device assigned to a specific individual. Civil libertarians may see the irony in the complete democratization of publishing at the expense of privacy.

From Books to Bytes

Consider that more than 400 pounds and 2 million pages of printed text can be distributed on a 1-ounce DVD, and it is clear why seven dental schools now require course materials on DVD. The disc can be replaced with updated data and played on any computer with a reader. However, the search for security is tending toward a restricted Web site or database for access to the information and temporary storage on a portable device.

Text will remain a central element in electronic books. Text will be stored in the computer with the kinds of codes that can be used for searching and indexing. Structural elements of a book's contents will be tagged with codes that faithfully map the content's intellectual structure: chapters, sections, footnotes, and sidebars. But technologists dream of pages that sing and dance—a world beyond text. Multimedia illustrations would be helpful in subjects requiring complex illustration, such as the sciences. It is expected the future e-book devices will have TV-like functionality, and that the text-based publication will be augmented with multimedia presentations. Audio, video, and animation, however, will increase the need for storage and require more sophisticated devices than mere text readers.

Libraries and other data repositories must take a more active role in shaping the future of e-publishing. Efforts are focused on standards, devices, delivery, security, and commerce; however, almost no consideration is being given to preservation.

References

Andersen Consulting. 2000. Reading in the New Millennium, A Bright Future for E-Book Publishing (PowerPoint summary of findings). Available at dec2000anderson2.ppt.

Association of American Publishers. 2000a. Digital Rights Management for Ebooks: Publisher Requirements, Version 1.0. New York and Washington, D.C.: Association of American Publishers, Inc. Available at http://www.publishers.org/home/drm.pdf.

Association of American Publishers. 2000b. Metadata Standards for Ebooks, Version 1.0. New York and Washington, D.C.: Association of American Publishers, Inc. Available at http://www.publishers.org/home/metadata.pdf.

Association of American Publishers. 2000c. Numbering Standards for Ebooks, Version 1.0. New York and Washington, D.C.: Association of American Publishers, Inc. Available at http://www.publishers.org/home/numbering.pdf.

Bush, Vannevar. 1945. As We May Think. *The Atlantic Monthly* (July):101–108

O'Brien, Daniel. 2000. *Books Unbound.* Cambridge, Mass.: Forrester Research.

Pimm, Bob. 2000. Authors' Rights in the E-Book Revolution. Available at http://www.gigalaw.com/articles/2000/pimm-2000-10.html.

TrendWatch Cahners. 2001. *e-Books: Publishing's Next Wave or Just a Ripple?* New York: TrendWatch.

Archiving the World Wide Web

PETER LYMAN
School of Information Management and Systems
University of California, Berkeley

Problem Statement: Why Archive the Web?

The Web is the largest document ever written, with more than 4 billion public pages and an additional 550 billion connected documents on call in the “deep” Web (Lyman and Varian 2000). The Web is written in 220 languages (although 78 percent of it is in English) by authors from every nation. Ninety-five percent of Web pages are publicly accessible, a collection 50 times larger than the texts collected in the Library of Congress (LC), making the Web the information source of first resort for millions of readers. Nonetheless, the Web is still less than 10 years old, and the economic, social, and intellectual innovation it is causing is just beginning.

The Web is growing quickly, adding more than 7 million pages daily. At the same time, it is continuously disappearing. The average life span of a Web page is only 44 days, and 44 percent of the Web sites found in 1998 could not be found in 1999.[1] Web pages disappear every day as their authors revise them or servers are taken out of service, but users become aware of this only when they enter a Universal Resource Locator (URL) and receive a “404–Site Not Found” message. As ubiquitous as the Web seems to be, it is also ephemeral, and much of today’s Web will have disappeared by tomorrow. The implication is clear: if we do not act to preserve today’s Web, it will disappear.

In the past, important parts of our cultural heritage have been lost because they were not archived—in part because past generations did not, or could not, recognize their historic value. This is a *cultural* problem. In addition, past generations did not address the *technical* problem of preserving storage media—nitrate film, videotape, vinyl recordings—or the equipment to play them. They did not solve the *economic* problem of finding a business model to support new media archives, for in times of innovation the focus is on building new markets and better technologies. Finally, they did not solve the legal problem of creating laws and agreements to protect copyrighted material yet at the same time allow for its archival preservation. Each of these problems faces us again today in the case of the Web.

The cultural problem. The very pace of technical change makes it difficult to preserve digital media. How many people can retrieve documents from old word processing diskettes or even find yesterday's e-mail? All documents follow a life cycle from valuable to outdated, but then, perhaps, some become historically important. Archivists often rescue boxes of documents as they are being transported from the attic on their way to the dump. But the Web is not stored in attics; it just disappears. For this reason, conscious efforts at preservation are urgent. The hard questions are how much to save, what to save, and how to save it.

The technical problem. Every new technology takes a few generations to become stable, so we do not think to preserve the hardware and software necessary to read old documents. Digital documents are particularly vulnerable, since the very pace of technical progress continuously makes the hardware and software that contain them outmoded. A Web archive must solve the technical problems facing all digital documents as well as its own unique problems. First, information must be continuously collected, since it is so ephemeral. Second, information on the Web is not discrete; it is linked. Consequently, the boundaries of the object to be preserved are ambiguous.

The economic problem. Who has the responsibility for collecting and preserving the Web and the resources to do so? The economic problem is acute for all archives. Since their mission is to preserve primary documents for centuries, the return on investment is very slow to emerge, and it may be intangible hence hard to measure. Archives serve the public interest in the very long run, with immediate benefits for only a few scholars. For this reason, they tend to be small and specialized. However, a Web archive will require a large initial investment for technology, research and development, and training—and must be built to a fairly large scale if it is continuously to save the entire Web.

The legal problem. New intellectual property laws concerning digital documents have been optimized to develop a digital economy, thus the rights of intellectual property holders are emphasized. Copyright holders have reason for caution, because the technology is so new and the long-term implications of new laws are unknown. Although the Web is popularly regarded as a public domain resource, it is copyrighted; thus, archivists have no legal right to copy the Web.

And yet it is not preservation that poses an economic threat, it is access to archives that might damage new markets. Finding a balance between preservation and access is the most urgent problem to be solved, because if today's Web is not saved it will not exist in the future.

Access is a political as well as a legal problem. The answer to the access problem, like the answers to all political problems, lies in establishing a process of negotiation among interested parties. Who are the stakeholders, and what are the stakes, in building a Web archive?

- For librarians and archivists, the key issue is to ensure that historically important parts of the documentary record are preserved for future generations.
- For owners of intellectual property rights, the problem is how to develop new digital information products and to create sustainable markets without losing control of their investments in an Internet that has been optimized for access.
- The constitutional interest is twofold: the innovation policy derived from Article I, Section 8 of the U.S. Constitution ("progress in the useful arts and sciences"), and the First Amendment.
- The citizen's interest is in access to high-quality, authentic documents, through markets, libraries, and archives.
- Schools and libraries have an interest in educating the next generation of creators of information and knowledge by providing them with access to the documentary record; this means access based on the need to learn rather than on the ability to pay.

In sum, the policy problem is to find a process for balancing these interests in the long run, including finding a means through which each of the parties can conduct and evaluate significant experiments and reach solutions that strike a balance among legitimate contending interests.

Technical Description of the Object

Howard Besser has identified five key technical problems necessary for digital preservation (Besser 2000).

1. The viewing problem is the maintenance of an infrastructure and the technical expertise necessary to make digital documents readable.
2. The scrambling problem is decoding any compression or technical protection service software protecting the Web page.
3. The interrelation problem is preserving the contexts that give information meaning, such as links to other Web pages.
4. The custodial problem is defining the standards, best practices, and collection policies that define the boundary of the work and its provenance and authenticity.
5. The translation problem concerns the way in which the experience and meaning of the Web page are changed by migrating it into new delivery devices.

When one is building a Web archive these problems translate into three questions: What should be collected? How do we preserve its authenticity? How do we preserve or build the technology needed to access and preserve it?

What is the Digital Object to be Collected?

Ultimately, the scope and scale of a Web archive will be determined by the definition of the digital object to be collected—the “Web page.” This is not a simple matter. From a user’s point of view, a Web page is the image called forth by placing a URL address into a Web reader. This operational definition is necessary but not sufficient, for an archive also must be sure that the document is *translated* in an authentic manner. In this case, authenticity means that the document must both include the context and evoke the experience of the original.

The average Web page contains 15 links to other pages or objects and five sourced objects, such as sounds or images. For this reason, the boundaries of the digital object are ambiguous. If a Web page is the answer to a user’s query, a set of linked Web pages sufficient to provide an answer must be preserved. From this perspective, the Web is like a reference library; that is, it is the totality of the reference materials in which a user might search for an answer. If so, the object to be preserved might include everything on the Web on a given subject at a given point in time, for example, the 2000 election or the World Trade Center terrorist attack. Thus, there is a temporal dimension: Must we preserve the context of the Web page at every point in time, at the time it was created, or when it was at its best? This raises the issue of quality: are we to preserve all pages relevant to a query, or just the best ones? And who is to judge?

None of these possibilities would be easy to realize, for the Web is not a fixed collection of artifacts. Today, the “surface” Web contains all of the static hypertext markup language (HTML) pages that can be accessed by URLs. Some of the surface Web, especially in the commercial sector, requires passwords or encryption keys; this area might be called the “private” Web. To archive these Web pages would require permission of the owners. The private Web is often encased in security protection services that make copying and preservation doubly difficult. Beyond these problems, surface Web pages are often generated on the fly, customized on demand from databases in the “deep” or “dark” Web. The deep Web is estimated to be 500 times larger than the surface Web. It includes huge data sources (such as the National Climatic Data Center and National Aeronautics and Space Administration databases) and software code that provides information services for surface Web pages on the fly (such as the Amazon.com software that creates customized pages for each customer). The deep Web is the information architecture that produces what we read on the surface; the surface itself exists only as long as a reader is using it. This deep Web cannot easily be archived, since the data are guarded by technical protection services. It is also potentially protected by privacy concerns, since if Amazon.com owns a profile of my use of information, it is not necessarily available for archiving without my consent. Here there are not only tensions between markets and archives but also conflicts between privacy concerns and the interest of history.

The ambiguous boundaries of Web objects are also problematic because they are compounds of design elements, including texts, pictures, graphics, digital sound, movies, and code—the list expands as innovation continues. Each of these elements has intellectual property rights attached to it, although they are rarely marked and sometimes impossible to trace. Yet, at least in principle, a digital archive would have to have permission from each of these rights holders. In the words of the National Research Council's report, *The Digital Dilemma: Intellectual Property in the Information Age,* "for the digital world, one must sort out and clear rights, even of ephemera" (National Research Council 2000, 12).

Even if the Web page could be copied technically and we knew what we wanted to preserve, Web pages are protected by copyright law. Even now there are sophisticated debates about how a Web archive should collect data: Should the default be that copyrighted information is collected and the owner has to opt out; or should it not be collected or disclosed unless the owner actively gives permission ("opts in")? This is a question that may be resolved by legislation or the courts. It is important to remember that the Web is a global document; consequently, there are likely to be many jurisdictions making laws and rules, and enforcement across national borders will be difficult without treaty agreements.

The Authenticity and Provenance of the Object Collected

Defining the boundaries of the object to be collected also requires decisions about authenticity and provenance. These decisions must be recorded as part of the archive; the preservation community calls this kind of information "metadata," or information about information, and often builds records of what is in the collection using these metadata. A standard way of recording the metadata must be created to record the historical and technical context in which the document(s) were found. Among many other facts, metadata might record answers to the following questions (Besser 2000):

- What is the name of the work? When was it created, and when has it been changed? Who created, changed, or reformatted it?
- Are there unique identifiers and links to organizations or files or databases that have more extensive descriptive metadata about this record?
- What technical environment is needed to view the work, including applications and version numbers, decompression schemes, and other files? If the Web page is generated on the fly, what database generated it, and what is known about its provenance?
- What technical protection devices and services surround it, if any?
- If the Web page contains more than text, what applications generated the sound, video, or graphics?
- What copyright information is there about each of the elements of the Web page, and what is the contact information for them?

Work to define standard answers to these and other questions is ongoing through the Dublin Core metadata project.

What Technologies Are Needed to Preserve the Web Collection?

Technologies to reproduce the Web object—however defined—must be preserved, including the hardware and software necessary to access the information in an authentic context or to recreate it. This is difficult in the best of cases. Have we authentically preserved a computer game if we preserve only the graphics, or must we preserve the look and feel of the game in use? Every solution changes the context of information in ways that affect its authenticity. One strategy tries to preserve the original equipment; another uses contemporary technology to emulate the original "look and feel" of the information in use; still another migrates the digital signal to new storage media.[2]

Migration is not just a technical problem. Storage media for digital documents are not yet stable for long-term preservation. Magnetic storage media such as tape and discs eventually deteriorate. Moreover, hardware and software eventually become obsolete, hence very expensive to preserve and operate. A Web archive must migrate from one technical environment to another as generations of technology succeed one another. Nevertheless, under today's law such migration could be a violation of copyright law because it involves copying the signal from one medium to another.

These problems are typical of those that occur in the early stages of every innovation, when getting to market quickly is more important than is perfecting the product. Digital information products are not designed for longevity, and even if they were, it is likely they would become obsolete quickly. As a consequence, the technologies of digital preservation are complex and expensive. The problems are understood far better than are the solutions at this point, but it is already clear that a Web archive will require substantial investment in technological infrastructure and technical research and development, and that commercial entities are unlikely to lead this effort unless there is short term economic value in doing so.

Organizational Issues

Both archives and libraries collect, organize, preserve, and provide access to the documentary record. The distinguishing function of archives is to preserve the integrity of documents for the long run.[3] Preservation for centuries invariably requires new technologies; hence, the Council on Library and Information Resources and other organizations are investigating long-term storage and migration of data.[4] While the technical problem of preservation is difficult, it is well understood. The problem of access, by contrast, involves legal and economic issues that have not yet been adequately explored. While print archives provide a useful model, the economic and legal environments surrounding print are quite different from those surrounding digital documents (National Research Council 2000, 113–116).

Economic and legal issues cannot be separated. In 1998, the Digital Millennium Copyright Act (DMCA) gave copyright owners rights to protect their works in digital formats. The DMCA implements the 1996 WIPO Copyright Treaty and WIPO Performances and Phonograms Treaty. Among the purposes of these treaties was harmonizing copyright policy around the world to encourage global commerce in digital information.

As a public policy, the DMCA was focused upon making the Internet safe for intellectual property. If digital information is easily moved from place to place on a network, such movement is considered to be copying and is protected by copyright. If Internet information is easily accessed, making it difficult for a rights holder to control distribution, the DMCA encourages the development of technical protection services (such as encryption) by making it illegal to develop technologies to break them.

For printed information, copyright policy has balanced information markets with public goods, such as education, the First Amendment, and libraries to provide access to information.

- The *first-sale* doctrine allows libraries to circulate copyrighted works to library patrons. In the digital realm, however, information may be licensed by contract rather than sold under copyright. With licenses, the provisions of the contract determine the uses that are allowed, which are unlikely to include library circulation or fair use. While printed works may also be sold with "shrink-wrap" licenses, the print market has not accepted them as readily as have markets for digital information.
- The *fair-use* doctrine allows for copying for personal educational purposes, within limits that are designed to protect information markets from damage. Here again, if licenses govern commerce in digital information, these copyright provisions do not govern the contractual agreement reached between buyer and seller.

The Digital Dilemma makes a constructive case for extending the fair-use doctrine to digital information in the future (National Research Council 2000, 137–139).

The rationale for the market approach, embodied in the DMCA, was twofold. First, new information markets are expensive to develop, and from the industry perspective, public interest doctrines such as first sale and fair use are taxes on this investment. Second, the global scale of the Internet means that millions of copies can be made and distributed in seconds, causing economic damage that cannot be repaired. Thus, while copyright laws governing print place emphasis upon ex post facto remedies such as litigation, the DMCA emphasizes prevention. Every digital copy, perhaps even copies made temporarily for system management purposes, thus requires the permission of the copyright holder. The DMCA explicitly allows archives to make digital copies of print works for the purpose of preservation.

To prevent illegal copying, the DMCA encourages the use of technical protection services such as encryption by making it illegal to use software to break them, and also making it illegal to develop and distribute such software. Software developers feel that this provision raises free-speech issues and perhaps property issues if it makes it illegal for the owner of a legal copy to make a backup. Congress recognized the complexity of some of these issues, empowering the LC to advise Congress whether this provision in Section 104 prevents noninfringing uses of certain classes of copyrighted works.[5]

What is the impact of these new legal regimes upon archives? Print archives are permitted to collect copyrighted materials and copy them for preservation purposes. For example, it is legal to copy print materials from one medium to another as part of a migration strategy over time, but it may not be legal to do so with digital collections, or to reformat them (e.g., from CD-ROM to a hard disk).

Differences between the production and distribution of printed and digital works raise additional legal issues for Web archives. When something is published in the print world, it is registered for copyright; thereafter, the laws governing it are largely unambiguous. On the Internet, it is not always clear when something has been "published." At this point, it is not clear to most users whether placing information on the Web places it in the public domain or under copyright protection. *The Digital Dilemma* concludes that the Web is copyrighted in principle, but notes public confusion on the issue and explores ambiguities that make it unclear whether archives have the right to make preservation copies and preserve them using migration strategies.[6]

In the print world, it has been possible to develop a copyright regime that balances the needs of markets and those of archives. The Internet makes it difficult simply to transfer copyright doctrine from the print to the digital environment. Yet many of the problems for the Web archive outlined earlier seem to be unanticipated consequences of laws intended to support the digital marketplace and might, in principle, be resolved by negotiation. This process might begin by discussing the possible damage to the marketplace caused by long-term archives and seeking solutions.

Implications for Long-Term Preservation

The most urgent task at this point is to create an organization capable of managing the process of building a Web archive, including negotiating to solve these problems. Inevitably, a Web archive will be a new kind of organization, one that responds to the problems and interests surrounding the Web. It may not be a place at all—it may be a function distributed among institutions over many locations on a global network.

The starting point for building a Web archive is to envision organizational strategies to manage this process. Two organizational strategies are emerging—one from the archival and library professions and the other from computer scientists. These strategies are not opposites and are not mutually exclusive, but contrasting them helps frame the strategic choices.

One library and archival strategy for organizing digital archives is presented in *Preserving Digital Information,* a report of the Task Force on Archiving of Digital Information (1996), published by the Commission on Preservation and Access and the Research Libraries Group. In contrast, Brewster Kahle's for-profit Alexa Internet and nonprofit Internet Archive might be used to illustrate the computer scientists' vision for organizing the Web archive.

Two Technical Strategies

Which profession should develop digital archives—librarians or computer scientists? In other words, who owns this problem?

- One technical strategy is offered by the library community, which has developed sophisticated cataloging strategies. The MARC record is used to build print library catalogs that may be searched by users to identify the best information resources. MARC records include fields to describe every aspect of printed documents; the Dublin Core metadata project is defining a standard for cataloging digital documents.
- Computer scientists funded by the National Science Foundation (NSF) Digital Library program are developing a second model. While the Dublin Core is designed to enable searches of library catalogs of digital collections, the NSF digital library projects are developing search engines that directly parse the digital documents themselves.

Records identify the best information source described in a catalog, while search engines and data-mining technologies go to the source itself. Each has its advantages. The point is that these technologies are optimized for two different kinds of archive. The computer science paradigm allows for archiving the entire Web as it changes over time, then uses search engines to retrieve the necessary information. An archival catalog supports high-quality collections built around select themes, saving only the Web sites judged to have potential historical significance or special value, and describing these special qualities in collection records and catalogs that could be searched.[7]

This is a fundamental debate about the nature of the Web as a technical object as well. The librarian tends to look at the content of the Web page as the object to be described and preserved. The computer scientist tends to look at the Web as a technology for linking information—a system of relationships (hence the name “Web”). This implies not only a difference in scale: it is a difference in philosophy. Should Web archives include everything or only carefully selected samples? Should the end user make decisions about the quality of the Web page, or should they be made by a selector who chooses which Web pages to save?

Preservation Powers

Copyright requires that copies of a published work be deposited in the LC, and the National Archives has the legal responsibility for archiving federal documents. In each case, responsibility is clearly located in a funded institution. How do the librarian/archivist and computer science models solve this organizational problem?

Preserving Digital Information (1996) proposes that the digital archive begin with principles such as the following:

- The copyright holder has initial responsibility for archiving digital information objects to ensure their long-term preservation.

- This responsibility can be subcontracted or otherwise voluntarily transferred to others, such as certified digital archives.
- If important digital objects are endangered because the owner does not accept responsibility for preservation, “certified digital archives have the right and duty to exercise an aggressive rescue function as a fail-safe mechanism” (Task Force on Archiving of Digital Information 1996, 20). Clearly, this “rescue function” would require a revision of the Copyright Act to create such a right and duty. Alternatively, the task force suggests the creation of a system of legal deposit, on the model put forth by a European Union proposal, to require publishers to place a copy of their published digital works in a certified digital archive. The word “certified” is important, for it refers to a professional and legal code of conduct so that access to the archive would not be misused.

The strengths of this proposal are that it creates clear institutional responsibility for the Web archive (“certified”) and describes necessary legislation to extend proven print models (such as deposit) to the digital realm. However, the proposal has not gathered political support, and the model relies upon already-scarce library subsidies for economic support.

Alternatively, consider the model of Alexa Internet and the Internet Archive. Alexa Internet is a for-profit corporation that measures the quality of Web pages by tracing consumers’ use of the Web. These measurements are made using an enormous Web archive, built by Alexa Internet using Web “spiders” (robots or agents) that roam the Web copying everything they find, unless forbidden entry. In this model, commercial use provides a viable economic base for the creation of the Web archive; note that Yahoo!, Google, and other search engine companies have also built large Web archives for commercial purposes. Alexa Internet then turns over the Web archive to the nonprofit Internet Archive, which provides for long-term preservation of the digital archive.

This linkage between corporate archives and nonprofit philanthropic archives is not unprecedented: many print archives have been built through philanthropic gifts from corporations or their owners after the economic value of the collection has faded. It relies upon the philanthropic vision of individuals, which may seem unreliable but may be more realistic than the legal establishment of a last-resort rescue power. However, it is problematic in that its funding depends upon the sustainability of a dot.com business model. Moreover, it is not clear that it is legal for a Web crawler to copy the Web without permission; Alexa Internet proactively copies, but removes Web pages from the archive upon request of the creator or copyright holder (an opt-out strategy).

The models developed by librarians and computer scientists are not opposites; in fact, they overlap in significant ways. Each relies upon a partnership between the for-profit and nonprofit realms, for in practice the digital archive is much more likely to rely upon the voluntary transfer of preservation responsibility from the copyright holder to certified archives than a controversial rescue power. Alexa Internet is an example of a philanthropic transfer from a commercial entity to an archive. Each model ultimately relies upon the resolution of legal ambiguities concerning the right to copy the Web. To some extent, each uses an element of eminent domain over copyright, the digital archive in its rescue power and Alexa Internet in its opt-out philosophy.

Access and Market Failure

Preservation does not threaten markets, but access might. How can the Web archive protect markets from the potential damage of competition from illegal copies preserved by the nonprofit sector? Four current practices might help to provide a solution to this problem.

1. *Delay.* The archive can delay making the archive available to the public until the economic value of the copy has been extracted. For example, Alexa Internet holds the tapes of the Web archive for six months before releasing them to Internet Archive. The length of the delay is an important subject for negotiation, since different kinds of content have different economic value cycles.
2. *Opt out.* The copyright holder can opt out of the archive. First, the Web crawler or robot making the copy can be automatically excluded from the Web site. Second, even if the crawler copied the item, the owner could ask that it be removed. This would allow the default to be that the Web is preserved, accomplishing the goal of the *Preserving Digital Information* task force, yet provide space for the owner and the archive to negotiate an agreement about the terms of access, if any.
3. *Restricted access.* The archive can restrict access to the collection to those judged by the copyright holder to pose no threat, a category that might include scholars.
4. *Motive.* On the model of the Fair-Use doctrine, the archive user could be required to have an educational motive and sign an agreement that the use of the archive would be restricted to certain purposes.

These ideas are not comprehensive; they are described only to suggest that current practices offer fertile ground for discussion.

Unresolved Issues

Every law ultimately relies upon the perception of citizens that it is fair. Within this general cultural approval of the legitimacy, a political consensus must be built among those with significant stakes in the issues. Often this kind of consensus begins with an agreement about a fair procedure for resolving differences; an example is the Conference on Fair Use (CONFU) process, which attempted to build a consensus that defined the Fair-Use policy.

The building of a public consensus will depend in this case on developing a shared understanding of digital information. Web pages clearly have intellectual and economic value, but thus far the new kinds of value created by Web pages, and digital information generally, have not been well described. The questions to be resolved include the following:

- How do the creators of intellectual property use information? Specifically, what is the role of Fair Use in creating new information? Is copyright law the best way to govern the role of digital information in the creative process, or is the public interest best served by an emphasis upon innovation, that is, the output of the creative process?
- What value comes from distributors or publishers in a networked environment? This is clear in print, but digital commerce is still in a highly experimental state of development, making the market value of digital commodities difficult for consumers to understand.
- Consumers give value to any commodity, in a sense, by sustaining markets that ultimately justify investment in innovations, but this relationship is unexpectedly novel in the case of Web pages. For example, Web pages collect information on users and often place cookies on readers' Web browsers. This information has commercial value, both enabling more customized services to be provided to the consumer, and, it is hoped, building brand loyalty and justifying advertising rates on Web pages. In this sense, we might now try to understand the consumer's role in the value chain and to define how the consumer adds value to information.

Old intellectual and organizational paradigms are not easily adapted to new digital markets because they do not describe them well; thus, they constrain innovation in markets that are still evolving. Ultimately, legal and policy frameworks for the digital economy must be consistent with the citizen-consumer's own experiences if they are to be perceived as legitimate.

If the social and political framework for the Web archive is still evolving, so, too, are other key elements. These include the following:

Evolving Technology

The Web has grown to global scale very rapidly; it may represent the fastest diffusion of a new technology in human history. At the same time, the technology of the Web has not stopped evolving. Even now, significant evolution is occurring as, for example, new architectures replace static Web pages with customized Web pages generated on the fly. Because innovation is not linear, the development of the Web is unpredictable. For stakeholders, the best option is to participate in the new organizations that, if they do not govern the future of the Web, at least attempt to analyze and influence its direction. To participate in discussions about the technical future of the Web, it is worthwhile to follow the discussion of the World Wide Web Consortium.

Evolving Law

Copyright law protects the entire Web. However, the Web is global, and a practice that is legal in one jurisdiction may violate the law in another. For this reason, Web law needs to become harmonized, which suggests that international treaty making (e.g., the WIPO treaty) may be as important as is national legislation.

Evolving Economic Issues

The Web began as software for the exchange of documents among scientists and researchers, using an Internet that was subsidized for education and research purposes. Today the Internet is increasingly commercial, and the Web has been the subject of vigorous investment as a technology for the digital economy. The search for sustainable business models for Web business has undergone a rapid evolution, ranging from Web advertising models to banner ads, sponsorship ads, subscription models, and business to consumer (B2C) enterprises. Investment in these enterprises and technologies has slowed for the moment because there is little sense that viable economic models have been identified.

Public Policy

In recent years, responsibility for information policy leadership at the federal level in the United States has been moved from the Department of Education to the Department of Commerce, because the Internet is seen as a medium for commerce and international economic competition. At the same time, the public sector policy governing the Web has been focused on e-government, requiring government agencies to develop Web resources and to move from print to Web publishing. Thus, at one pole the market was treated as the best way to deliver content onto the Web, while at the other pole, the public good was defined solely in terms of online government information. There is a space between these two poles, where a broader concept of the public interest could be developed. This is a space that might be called "innovation policy," and that is the ground upon which a Web archive policy, among other innovations, might be created.

References

Besser, Howard. 2000. Digital Longevity. In *Handbook for Digital Projects: A Management Tool for Preservation and Access,* edited by Maxine Sitts. Andover, Mass.: Northeast Document Conservation Center.

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

Lyman, Peter, and Hal Varian. 2000. How Much Information? Available at: http://www.sims.berkeley.edu/research/projects/how-much-info/.

Lyman, Peter, and Howard Besser. 1998. Defining the Problem of Our Vanishing Memory: Background, Current Status, Models for Resolution. In *Time and Bits: Managing Digital Continuity,* edited by Margaret MacLean and Ben H. Davis. Los Angeles: Getty Information Institute and Getty Conservation Institute.

National Research Council. 2000. *The Digital Dilemma: Intellectual Property in the Information Age.* Washington D.C.: National Academy Press.

Rothenberg, Jeff. 1999. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation,* Washington, D.C.: Council on Library and Information Resources. Available at: http://www.clir.org/pubs/abstract/pub77.html.

Sanders, Terry. 1997. *Into the Future: Preservation of Information in the Electronic Age.* Film. 16 mm, 60 min. Santa Monica, Calif.: American Film Foundation.

Task Force on Archiving of Digital Information. 1996. *Preserving Digital Information.* Washington, D.C.: Commission on Preservation and Access and Research Libraries Group. Available at: http://www.rlg.org/ArchTF/tfadi.index.htm.

Web sites noted

Alexa Internet. http://www.alexa.com
Dublin Core. http://dublincore.org
The Internet Archive. http://www.archive.org
World Wide Web Consortium. http://www.w3c.org

Notes

1. Numerical descriptions of the Web are based on data available in fall 2000. These data sources were originally published on the Web, but are no longer available, illustrating the problem of Web archiving. However, the original sources are reproduced in detail in Lyman and Varian 2000, and are available at http://www.sims.berkeley.edu/research/projects/how-much-info/internet/rawdata.xls. Some of the source documents are available on the Internet Archive's "Wayback Machine" at http://www.archive.org/.

2. A comprehensive description of the technical issues in digital preservation is provided in Rothenberg 1999. Migration is discussed on page 13, and emulation on pages 17–30.

3. For functional descriptions of the terms "digital library" and "digital archive," see Task Force on Archiving of Digital Information 1996, page 7.

4. The Council on Library and Information Resources has published numerous papers on digital preservation. See http://www.clir.org.

5. In August 2001, the Copyright Office at the Library of Congress released the DMCA Section 104 Report, available at http://www.loc.gov.

6. See the more detailed discussion in National Research Council 2000, 113–119.

7. On the issue of the quality of information, see, for example, Conway 1996.

Preservation of Digitally Recorded Sound

SAMUEL BRYLAWSKI
Recorded Sound Section
Motion Picture, Broadcasting and Recorded Sound Division
Library of Congress

The views and opinions expressed herein are those of the author and do not necessarily reflect those of the U.S. Government or the Library of Congress.

Introduction

In 1878, Thomas A. Edison speculated publicly on the possible uses of his phonograph, the first device for recording and playing back sound. Among the 10 applications he predicted were recording music, aiding business dictation, preserving reminiscences (oral histories), creating talking books for the blind, and recording educational lectures. Today, all of Edison's predictions have come true, and uses not imagined in the nineteenth century are common. Every day, thousands of hours of sound are produced and disseminated by radio, compact discs (CDs) and cassettes, and the World Wide Web. People throughout the world, in all economic strata, depend on recorded sound for entertainment, information, and intellectual stimulation.

The twentieth and twenty-first centuries are documented and recorded by sound and image as well as by words. We perceive much of the world through packaged and broadcast images and sounds. Our experiences today, and those of the last 100 years, are documented in these media for the study and enjoyment of generations to come. Sound recordings carry the voices and music that have shaped a century—voices of one's own family as well as of politicians and other well-known persons. Recorded music in archives includes unique aural documentation of indigenous peoples; the varied jazz, sacred music, and popular and folk songs that form the roots of contemporary rock; and the multimillion sellers themselves. Broadcast radio news collections document historical events and how they were presented to the public.

The great challenge to the librarians and archivists who are entrusted with preserving our culture for posterity is to determine which, and how much, of the thousands of hours of sound recorded daily to retain. Similar challenges have always faced caretakers of culture. However, with so much sound now available, through many media and in many formats, they have become more complex. That these sounds are now predominantly digital makes the challenges more formidable and the opportunities more extraordinary.

Sound has been recorded digitally since the 1970s, when pulse code modulation (PCM) became an accepted method of recording by audio engineers and producers. Today, digital recording techniques and processes contribute to nearly every recording made or distributed. Digital sound, however, has evolved in meaning as it has proliferated in use. In the consumer marketplace, compact audio discs, World Wide Web audio streaming, MP3 sound files distributed through the Web, and DVD audio discs all fall under the rubric of "digital audio," yet they have been created to varying standards and in a wide variety of formats (Schoenherr 2002). Today, a digital recording is as likely to be a computer file, with no tangible attributes, as it is to be a compact disc or digital audio tape (DAT).

For example, the sound collection of a large library might include 78-rpm jazz recordings on shellac and vinyl long-playing discs and re-recorded on R-DAT cassettes, as well as the published recordings of a contemporary rock band recorded on compact audio discs, with unpublished recordings of the same band on MP3 files. The library might hold a group of vintage radio dramas on instantaneous analog discs that have been reformatted for preservation on open-reel analog tapes. An oral history collection or other field research recording might be found on the Sony digital MiniDisc format. The audio reserves service room of a university library might be holding a collection of MP3 files recorded from contemporary radio talk show broadcasts streamed on the Web.

With the development of the World Wide Web have come new digital sound formats and delivery systems that offer archivists, as well as home consumers, a wider variety of recorded sound, instantaneously, than in any time in history. MP3 files, sound files created by an algorithm that highly compresses (reduces) the amount of data required to convey the audio information, proliferate on the Web, illegally as well as legally. MP3 files commonly consist of "home-recorded" tracks by aspiring popular music groups; illegally distributed commercially owned recordings of contemporary and older popular music groups; and spoken-word and music recordings made available free or offered for sale by legal owners or licensees. In addition, thousands of individuals and corporations offer music, spoken-word recordings, and radio programming over the Web as "streams"—continuous sound delivered from Web sites to which users have no choice of content other than deciding which site to monitor. Whether these sound recordings are going to be maintained for posterity or only for the next 10 years, if they are to persist, it will be as digital recordings of some type.

Types and Rights

Major sound archives hold many conventional forms of commercially produced analog sound recordings, such as 78-rpm “coarse-groove” discs, 33 1/3-rpm long-playing “microgroove” recordings (LPs), and cassette tapes. Whether of music or the spoken word, such recordings are usually the aggregate creation of several parties. These creators have varied rights to the use of the recordings. Copyright in the sound recording itself is usually held by the corporation that issued the recording, i.e., the record label. Most recordings are representations or performances of an “underlying work,” a musical composition or literary text that is protected by its own copyright. A royalty based on sales or use is paid to the holder of the copyright in the underlying work.

While these may be the only copyrights per se in the recording itself, other rights may be inherent in the work. Printed materials included in the packaging, both textual and graphic, may be protected by copyright, again including underlying rights as well as protection for new matter. Vaguer and more complex are the possible rights in recordings held by trade union members and other artists who contributed to the recorded work. American Federation of Musicians or other union recording contracts with record companies may call for additional fees to the union for uses beyond single-unit retail sale. The rights of recording artists to the sound recordings on which they are heard is currently a subject of conflict between some artists and their record companies. Points of contention include royalties due from new media uses and the ownership of recording masters.

Many archives’ most significant holdings are not commercially produced recordings but are unpublished recordings of various types. Such works include radio broadcast recordings, television sound tracks, “live” musical or dramatic performances, ethnographic field recordings, and interviews. It is in these recordings in which rights issues are most complex and in need of study, and perhaps adaptation, as they relate to preservation. When a for-profit or nonprofit corporate body, such as a broadcast network/station/producer or a music producer, creates these unpublished recordings, that body often owns the rights to the recording. As with commercially distributed published recordings, unpublished recordings are usually interpretations of music or literary underlying works that are commonly protected by copyright. Because the recordings were intended to remain as unpublished works when they were originally made, the producers were very unlikely to have entered into any contractual agreements with their co-creators, such as members of creative trade unions (musicians, actors, writers, and announcers), authors of underlying works, or interviewees. In some recordings, such as unauthorized tapings of live performances (“bootlegs”), none of the contributors to the work, including the producers, was aware that a recording was being made.

In the United States, federal copyright protection was not available for sound recordings until 1972. However, state and common laws protect these recordings until the year 2067, no matter when they were created. This means that, in effect, the law grants greater protection to sound recordings than to print materials. Determining exactly which parties hold the rights to a pre-1972 recording can present significant challenges, because no centralized registration exists as it does for post-1972 federal copyright protection.

Audio Acquisitions

The radical transformations that have made digital formats the predominant form of sound recording have made available to the public more types of sound recordings, and greater numbers of hours of audio, than ever before. As a result, research library administrators responsible for collection development policies must regularly reevaluate their long-range goals as well as their day-to-day acquisitions. No longer are acquisitions limited to physical items offered by retailers and in catalogs, or bought on their behalf by contracted purchasing representatives. Rather, librarians and archivists face a plethora of technologies, platforms, and genres.

Compact Discs: The First Digital Audio Revolution

In the consumer arena, the digital audio revolution began in the early 1980s, when the compact audio disc format was introduced. Public adoption of the CD format burgeoned beyond anyone's expectations. The public, and libraries, were attracted by the lack of surface noise and hiss that was commonly heard on LP and 78-rpm records and cassette tapes and by the CDs' touted invulnerability to normal wear. The sound on compact discs was criticized by audiophiles, collectors with high-end playback equipment, and other consumers, but most consumers never heard their arguments or the aural evidence. In fact, the 44-MHz, 16-bit sampling rate, or amount of compression, selected by the creators of the compact discs was a compromise that sacrificed sound quality at the expense of time capacity of the discs. As would be the case in the late 1990s with even more radically compressed MP3 audio files, convenience and cost proved to be more important to consumers than high fidelity was. Nonetheless, years after the introduction of the compact disc, manufacturers' claims of its indestructibility have been debunked. Archives that plan to make their holdings permanent will have to reformat CDs just as they will audio tapes and other fragile media.

Initially, the content of compact discs replicated that of the LP discs they would supersede. However, record companies gained significant profits from the re-release of older catalog issues, in addition to new releases. This new market for "old" holdings paralleled the growth in numbers of re-releases of motion pictures on video tape, which was occurring at the same time. Companies rediscovered the value of their archives of older intellectual property. In many cases, they discovered that they had prematurely destroyed their own masters under the mistaken assumption that there was no "aftermarket" for them. The convenience and lack of background noise on CDs prompted the public and libraries to recreate their holdings of LP discs and replace them with CD reissues.

Serious sound archives dedicated to documenting the history of music and sound recording continue to acquire LP and 78-rpm discs for their unique repertoire and their audio quality. Stored properly, these discs will last many years, but they deteriorate from repeated playback. Moreover, high-quality disc playback equipment is expensive. It is becoming more difficult to acquire the hardware to play these recordings adequately.

With compact discs came myriad recording reissues. The complete recording careers of hundreds of notable classical, jazz, blues, and rock artists have been thoroughly documented on thousands of CD reissues. These discs and sets have enabled libraries to build research-level, encyclopedic collections of important musicians and recording artists. These are recordings that libraries might not have obtained otherwise, either because of inaccessibility or the expense of obtaining and maintaining the original records.

Two important points related to reissues must be emphasized. The first is that most comprehensive jazz, blues, and classical reissues are produced outside of the United States in countries where older recordings are no longer protected by copyright. In most European countries, the copyright on a sound recording is 50 years from the original date of recording. In the United States, it is 95 years from the date of recording for post-1972 recordings and, possibly, until the year 2067 for pre-1972 recordings. (It is usually only the recording that has entered the public domain overseas. The underlying works—i.e., the musical compositions—are protected by longer copyright terms and the royalties due on them are often paid.) Most jazz and blues reissues sold in the United States are, technically, illegal imports. However, as the 50-year span enters the rock-and-roll era, it will not be unusual to see stricter enforcement of the U.S. law or pressure on European countries to change their laws to conform with those of the United States.

The second point is that the profusion of reissues presents challenging selection and preservation issues to libraries. Although liberal foreign copyright laws enable publication of thousands of previously out-of-print recordings, the quality of these reissues varies greatly. While the producers of comprehensive reissues make thorough searches to locate one copy of every recording an artist has made, the copy used is often generations away from the master recording and is in only mediocre condition. To compensate for the condition of the source recordings, many producers of reissues misrepresent the original recordings with signal processing: overuse of noise reduction, sound equalization, and limiting tools in order to reduce the surface noise found on the source. The result is a quiet recording that distorts the richness of sound on the master recording. When the time comes to preserve these recordings, it will be very difficult and time-consuming to select the best source material from the abundance of available issues.

New Means of Digital Audio Distribution

Compact discs brought significant changes to archives, but these changes pale in comparison with those that digitally created and distributed sound files will bring. Today, many archives are rethinking their acquisitions policies, preservation techniques, and delivery systems. The sheer number of new audio materials made available through the World Wide Web is astounding. The greatest attention has been paid to MP3 files legally and illegally traded through peer-to-peer networking programs such as Napster. Music publishers and record companies halted the use of Napster as a source of free copyrighted music, but the program's popularity has resulted in the development of authorized paid subscription services that intellectual property holders hope will take its place. This phenomenon will have ramifications for library acquisitions. There is promise for more thorough audio acquisitions programs facilitated by streaming sites, as well as subscription services offered by Web companies.

In general, post-1960 radio broadcasts are represented more sparsely in archives than is any other contemporary mass medium. Popular public radio broadcast series have long been available for sale on audio cassettes, but few other radio broadcasts are available to libraries or the public. Before radio broadcast streaming over the World Wide Web, one could acquire commercial radio broadcasts by tape recording them or by subscribing to a service that sold recorded samples of a station's "sound"—that is, its mix of disc jockey patter, public service announcements, and station identification and advertisements. Programming archives are held by public radio production and distribution companies, such as National Public Radio and Minnesota Public Radio, but few popular commercial broadcast radio series are collected systematically or preserved in any manner. Twenty years ago, a popular radio talk show that featured nationally renowned guests offered its archive to the Library of Congress (LC). The LC turned down the collection, and the tape collection was subsequently destroyed.

Radio on the World Wide Web

A large number of radio broadcasts, contemporary and vintage, are streamed on the Web. By one estimate, more than 2,500 radio stations stream all of their programming. This figure was from before April 2001, when a strike was called by the American Federation of Television and Radio Artists (AFTRA), which is demanding supplemental payments to its members for streaming of radio advertisements in which they appear. In addition to individual stations, more than 30 radio networks stream over the Web, according to the *Radio and Internet Newsletter*.

Computer software, such as that sold by High Criteria, Inc., enables streamed audio to be recorded and converted to WAV or MP3 files. Streaming is not intended to be recorded, or fixed, by the user. The laws and licenses that govern streaming were designed with the assumption that its use is ephemeral. It is unknown whether recording streamed audio for archival purposes is legal. However, under the provisions of the American Radio and Television Archives law, which was enacted in 1976 to support an archive of American broadcasting at the LC, the Library may be allowed to acquire streamed audio of radio broadcasts.

The costs of streaming broadcast radio over the Web include license fees to the copyright holders such as music publishers' representatives and the Recording Industry Association of America, which represents record companies, and hardware and networking costs. Some of these fee structures were still being negotiated at the end of the summer of 2001. A solid framework for the profitable streaming of commercial audio has not yet emerged; however, a number of digital audio subscription services offer unique and important programming that may prove to be profitable sooner than streamed commercial radio will. The company Audible.com offers monthly subscriptions to daily radio programs, audio versions of national magazines and newspapers, three original programs, and hundreds of books and lectures. The content is delivered through the Web to subscribers as one of three proprietary audio file types. It is not known whether any public archive holds copies of the Audible.com programs other than those derived from public radio sources. Audible.com is one of several services that now sell spoken-word audio as computer files. The company claims to have 28,000 hours of audio, produced by 160 content partners.

Another firm, Real Networks, offers a subscription service in collaboration with major league baseball. The service enables those who pay a monthly fee to hear a live radio feed of every major league baseball game. It also allows subscribers access to an archive that includes recordings of every major league game of the season. It is not known whether any public archive would be interested in holding every baseball game radio broadcast of a season, but it would not be unusual for an archive to want to hold a home team's season. Likewise, a research library with strong baseball holdings might want to build a representative collection of every baseball announcer working in the major leagues.

The Web has also given rise to what might be called "private streaming" radio stations. Several Web companies (e.g., Live365.com and Shoutcast.com) enable individuals to stream audio segments of their own choosing, organizing and advertising their programs under a variety of themes. Such indigenous radio stations, often unaffiliated with any companies or organizations, exploit the narrowcasting potential of the Web. Archives will want to document this trend and possibly preserve the programming of stations issuing very unusual content. Much of the programming on these private stations concentrates on common hit music, which archives are unlikely to preserve in this format.

Web audio might also be systematically archived under the auspices of the U.S. Copyright Office, under the mandatory deposit requirements of copyright law. As subscription publications, popular radio programs such as "All Things Considered," "Fresh Air," and "Car Talk," as well as the daily *New York Times Audio Digest* and *Audible Los Angeles Times* are probably subject to legal demand by the Copyright Office. It might be argued that streamed Web content is subject to the same requirements.

New Modes of Business

Libraries and archives whose missions include documenting contemporary music and broadcasting face great challenges with respect to materials selection. A sampling of Web streaming sites might fulfill these mandates and adequately document the trend of audio being distributed exclusively as Web streams. However, independent musicians (that is, those not affiliated with a record label) now use the Web to distribute their recordings. Web sites include tens of thousands of MP3 files available for free sampling or for downloading for minimal payment. As with Web radio sites, music distributed on the Web can be targeted to audience niches. In theory, profits can be made on only moderate sales. Musicians tout the Web's potential for directing their work to audiences, thus circumventing record label middlemen, whom, they believe, neglect performers without mass appeal and reduce musicians' earnings. At this time, the outcome of efforts by musicians and others to recast traditional modes of music distribution is unknown. So much music was available free, through services such as Napster, that it remains to be seen how many people will be willing to pay for obtaining music files from the Web.

Two Web music subscription services, MusicNet and PressPlay, are being introduced by the five major record companies. Vitaminic, an Italian commercial Web distributor of music from independent labels and musicians, claims to manage songs by 20,000 artists and is in operation currently, as are many smaller sites created to serve independent musicians. Through these services an enormous amount of music will be available to subscribers, which may include libraries; however, the audio fidelity of the files available for download will not be of high quality. The files are likely to be compressed MP3, Windows Media, or other file formats, with significantly less sonic quality than audio fixed on a compact disc or LP. The companies that manage the sites featuring independent music will not hold higher-quality copies of the music. Nor are the companies likely to maintain archives of music they no longer sell, especially licensed content. For example, MusicNet distributes more than 3,000 "live" concerts, otherwise unpublished, which may be accessed by subscribers who pay an additional premium. If the artists terminate their contract with a site, or if the site goes out of business, how will the music be preserved, and by whom?

In coming years, hundreds of thousands of music files are promised to be available exclusively through the World Wide Web. No single library will be capable or desirous of preserving this abundance of content. Only a small fraction of the popular music groups whose work will be made available through these new means will ever receive national recognition. Some of this music will be of interest to research libraries and archives. Some libraries will desire music that is progressive or that contains sophisticated topical or literary song lyrics. Libraries with a localized mission or constituency, such as those associated with historical societies or state universities, might choose to document comprehensively local musicians whose songs and music are on the Web. Harvesting these songs will be difficult. The challenges of selection are nearly overwhelming. However, the library community might aid subscription Web music sites by collaborating in the design of indexes to the sites and using those indexes to build collections. Artists who add song files to a Web site currently categorize their work by genre for inclusion in the sites' directories. Libraries might work with sites to encourage documentation of regional designations as well, to aid in the search for music of local interest. Collaboration with music sites could also extend to preservation efforts managed jointly by the sites and libraries, with the endorsement and cooperation of the artists. Archives can assist in assuring the preservation of high-fidelity copies of contemporary music. The widespread adoption of heavily compressed MP3 files indicates that high fidelity audio is not a priority for many digital music enthusiasts, so much music is distributed exclusively as compressed files. Yet the original recordings from which the compressed files were created are high fidelity and should be preserved in that form when possible.

Rights Management and Protections

The copyright controversies surrounding the creation and trading of MP3 files affect archives in a number of ways. The record industry's actions in response to the widespread violations of their copyrights include creation of protective digital-rights-management systems such as the Secure Digital Music Initiative (SDMI). SDMI is a digital watermark system that was developed to be read by compatible hardware in an effort to prevent illegal duplication of files. Other such systems have impeded legal uses of compact discs, including preservation. Compact disc encoding intended to prevent "ripping," digital audio extraction of compact discs, or conversion of CD tracks to MP3 files, have prevented compact discs from being played at all in CD-ROM computer drives. Because compact discs are not permanent, such anti-piracy efforts could seriously impede preservation of the discs by libraries and archives by preventing legal duplication for preservation. Many experts believe that illegal copying of compact discs and other formats will never be completely inhibited. Driven by what has been termed a "power struggle" between intellectual property owners and customers, computer hackers will always be eager to subvert antipiracy devices or programs, despite the law. Those less technically adept are likely to acquire hardware that circumvents digital duplication impediments by recording files from analog leads, either for recording on analog cassettes or re-conversion to nonwatermarked digital files. These ongoing intellectual property skirmishes are likely to make record companies and other rights holders wary of cooperative preservation projects in which files might be shared between archives.

The documentation and preservation of music and the spoken word distributed through the Web is a great challenge to libraries and archives—one that no single institution is likely to be able to accomplish on its own. It has been suggested that libraries seriously interested in preserving the profusion of files of contemporary music and other audio materials available through the Web collaborate with each other. In its study on a digital strategy for the LC, the National Academy of Sciences recommends that libraries, led by the Library of Congress, define a subset of digital materials for which to "assume long-term curatorial responsibility" (National Research Council 2000a). Such collaboration might result in the preservation of a greater percentage of available audio and reduced redundancy.

Preservation

The "Permanent" Format and Repositories

Only within the past few years have archivists begun to accept digitization as a means to preserve audio holdings that are at risk of deterioration. In the past, librarians and archivists distrusted digital media as a format to save important audio recordings. No medium has proved stable enough to be called permanent. A significant amount of data compression has been inherent in digital sound recording, including compact audio discs, and has reduced the quality of the sound being preserved, especially in comparison with high-quality analog recordings. Several factors have led to a shift toward digital preservation. The preferred preservation medium of the last 45 years is quarter-inch analog magnetic tape on 10-inch open reels. In 2001, only two major companies still produced the tape stock. Only a few companies manufacture the machines that play open-reel tapes. Ironically, many of the master preservation tapes produced in the 1970s and 1980s are deteriorating faster than are the original older media they were intended to preserve. Many brands of tape stock manufactured less than 20 years ago are subject to hydrolysis, because the binder that adheres the recording material to the backing absorbs moisture from the air. Upon playback, the tapes squeak and break down.

Ultimately, preservation reformatting will be required for all media upon which sound is recorded, since preservationists acknowledge that there is no permanent format. Most preservationists believe that resources spent to identify and develop a permanent medium are better spent building systems that acknowledge impermanence and exploit the potential of readily available technology. Digital media have the advantage of not suffering any loss of information as they are copied, unlike the generational losses inherent in the duplication of analog media such as discs and cassette tape. The future of audio preservation is reformatting audio tapes and discs to computer files and systematically managing those files in a repository.

Digital audiovisual file repositories, in wide use by European broadcasting companies, are designed to back up their data systematically on the preferred storage format of the moment, under the assumption that that format will change from time to time. The data are to be sustained through any number of shifts in design and configuration of storage format. Digital mass-storage systems (DMSS), as the repositories are called, ensure the persistence of data by validating their integrity as they are copied periodically. Such systems are complex in design and inherently dependent upon sophisticated technology that must be maintained in perpetuity. Yet, to many archivists they are liberating. The well-planned repository presumes media obsolescence, plans for it, and, according to its supporters, frees the archive community of the futile search for an affordable permanent medium.

Digital Objects and Metadata

Digital repositories such as the one proposed for the LC call for each audio recording in the repository to be represented by a set of digital files, a “digital object.” The digital object comprises the audio tracks of the recording; graphic components of the recording’s packaging, such as disc labels, dust jackets, and sleeves; and metadata (which can be partitioned into “descriptive,” “structural,” and “administrative” metadata) about the original recording and its digital files. To archivists, the print elements of a sound recording are important components in the preservation of the sound recording. Not only must they be preserved with the recording: they must be accessible to the researcher, in context, when the recording itself is played. Structural metadata identify and organize the individual files (termed "intermediate objects") of images and sound that represent a digitized item. The metadata assist the presentation of these from the digital repository. In a repository, structural metadata are called up by program scripts to reconstruct virtually the sound recording's packaging (e.g., scanned images of the covers, accompanying text) and to provide researchers with control over which audio tracks to audition.

In digital preservation programs, administrative metadata record exactly how an item is preserved: specifics of hardware used, hardware settings, and signal processing employed, including data compression rates. Administrative metadata include a limited amount of rights information for each sound recording preserved. Restrictions specific to the sound recording, such as donor information and the year the sound recording itself is expected to enter the public domain, are also recorded as metadata.

It is clear that the success of digital preservation efforts will rest to a significant degree on the scope and reliability of the metadata recorded. Metadata support and make possible the asset-management systems that back up and periodically duplicate digital audio files in a preservation repository. Metadata can help in limiting access to intellectual property to those with proper authorizations. As descriptive cataloging information, metadata enable people to locate what they are looking for in a repository. However, full repository systems require hundreds of metadata elements for each preserved item. At this time, populating the metadata databases is very labor-intensive—that is, expensive—and could be a barrier to the development of digital repositories. Among the recommendations that the National Research Council (2000b) made to the Library of Congress in the LC21 report is that "the Library should actively encourage and participate in efforts to develop tools for automatically creating metadata." Many believe that such tools are essential to the development of effective digital preservation programs.

Standards for preservation and repository-related metadata are now being developed. Work by the Audio Engineering Society and other organizations will result in refinements of Dublin Core descriptive metadata definitions as they relate to sound and guidelines for documentation of technical preservation information. The integration and standardization of competing metadata formats is only beginning to be addressed. In the field of audiovisual repository management, the Digital Library Federation's Metadata Encoding and Transmission Standard project (METS) is especially promising. METS is an XML-based format for structural, administrative, and descriptive metadata that builds on the object framework outlined by National Aeronautics and Space Administration's Open Archival Information System. It is designed not only to assist in the management of files within a digital repository and the presentation of those files to a user, but also to enable the exchange of files between repositories. Given the high expense of professional-quality preservation, especially digital preservation, such a standard could be particularly useful. There is little likelihood that METS or any format will be adopted universally. METS is still evolving, and commercial audiovisual digital repositories that use other metadata system are already in operation.

Standards

The standards needed for effective digital preservation are by no means restricted to metadata. There is considerable debate among preservation recording engineers, archivists, and conservators over the principles and guidelines that direct capture from analog audio sources. There is a general consensus that the digital configuration of standard compact discs (44 MHz, 16 bit) is inadequate, but debate over how high the sampling rate and word length of digital preservation should be. Many engineers and conservators argue for a sampling rate of 192 MHz and word length of 24 bits, at a minimum. The diminishing costs of computer storage space have alleviated the need to process audio data with high-compression algorithms. Some archivists advocate a sliding standard based on the nature of the source material (e.g., whether it is spoken word or music, or its frequency range). Given the frequent debates over audio standards and fervid opinions of specialists, it is unlikely that there will ever be universal agreement on standards. However, scientifically designed tests will further refine the questions debated, if not devise a resolution. The National Recording Preservation Act of 2000 directs the Library of Congress to work toward the creation of standards for digital preservation.

Most archivists now agree that the initial preservation capture of audio should be a flat transfer of the source signal. The master preservation file or recording should not include any playback curve or signal processing, such as that used to reduce analog disc surface noise. Standard equalization curves used on the analog source recordings are noted in metadata. Computer controlled playback devices can then reintroduce the equalization during playback. Recently developed digital audio workstations aid in recording this technical metadata, including the condition of the source, as well as its technical characteristics. However, most existing digital audio workstations are designed for production, not preservation transfers, and require further enhancements to meet the standards of preservationists. Many otherwise-sophisticated digital audio workstations currently available do not allow digital recording at high sampling rates, such as 192 MHz.

Conclusion: The Importance of Collaborative Approaches

At this time, there is virtually no coordination of preservation efforts between commercial archives, such as those of the record companies, and institutional archives. While this might not be surprising given their different missions, collaboration could be mutually beneficial for many reasons. According to an award-winning series of articles in *Billboard* magazine, record companies have discarded thousands of master recordings and thus hold incomplete archives of their intellectual property (Holland 1997). No central database or file of master recordings exists. Such a database was attempted in the 1990s, but companies were reluctant to share what they felt was proprietary information. Many of the major record companies' releases are held only by collectors and institutional libraries and archives. Companies and archives might wish to pursue collaborative preservation projects whereby 78-rpm and LP discs held by institutional archives are digitized jointly and companies' digital sound files are shared with archives in a controlled setting.

Such collaborative projects would not be easy to undertake. Record companies today feel bruised by the rampant swapping of music files propagated by programs such as Napster and may be reluctant to authorize the use of master files outside their domains, however strictly they are controlled. In fact, copyright laws, particularly those enacted to reduce digital piracy, now can prohibit legitimate and necessary preservation functions (National Research Council 2000a).

Whether between record companies and archives or with others, some type of collaborative approach to audio preservation will be necessary if significant numbers of audio recordings at risk are to be preserved for posterity. Hundreds of thousands of magnetic tapes and fragile discs risk being lost if they are not preserved in the next 20 to 50 years. The cost of preservation will be in the tens of millions of dollars. One particular risk of preservation programs now is redundancy. Archives capable of creating high-quality preservation master files have few means to ensure that other archives have not preserved the same files. Descriptive metadata are often derived from library catalog records that do not identify unique musical performances or do so in a nonstandardized format that is difficult to exchange. Moreover, most of the descriptive metadata now being created do not provide detail at the high level of granularity required to fully identify the musical compositions that make up a recording (for example, composers' names and dates of compositions). Publishers and performing-rights organizations do maintain such information, and it can be accessed through new technologies such as "audio fingerprinting," which enables devices to identify music selections aurally in only a few seconds, but it is not available for population of public databases.

Inadequate cataloging is a serious impediment to preservation efforts. Without full inventories and cataloging of their collections, archives are ignorant of the scope of the challenges they face and are hindered in creating comprehensive preservation plans. The problem is especially acute for unpublished holdings, such as recordings of concerts, radio broadcasts, oral histories, and ethnographic or field recording collections. Many libraries are required to devote most of their cataloging resources to published materials, for circulating collections and other materials used daily. The full scope of preservation needs can be realized only if libraries and archives can devote more resources to cataloging unique or unpublished holdings. It would be useful to archives, and possibly to intellectual property holders as well, if archives could use existing industry data for the bibliographic control of published recordings and detailed listings of the music recorded on each disc or tape. The 1970s witnessed the building of bibliographic utilities that enable libraries to share cataloging data, primarily for books and magazines. These utilities now include cataloging for hundreds of thousands of sound recordings, but the detail is grossly inadequate to manage preservation or share files. Greater collaboration between libraries and the sound recording industry could result in more comprehensive catalogs that document recording sessions with greater specificity. With access to detailed and authoritative information about the universe of published sound recordings, libraries could devote more resources to surveying their unpublished holdings and collaborate on the construction of a preservation registry to help reduce preservation redundancy.

The sharing of nearly all preserved audio files is illegal under current laws, which place restrictions on audio recordings made as long ago as the nineteenth century. If secure networks are developed and rights holders could be assured that piracy of their music would not result, special licenses or agreements with intellectual property holders might be devised to provide wider access to out-of-print and unpublished recordings. Many archivists believe that adequate funding for preservation will not be forthcoming unless and until the recordings preserved can be heard more easily by the public. Archives are interested in this issue, and they could be active partners in the creation of subscription services, which include a variety of music now wider than that available in the commercial market. Many would be willing to share their files of preserved audio files with other institutions or individuals if reciprocal agreements could be formulated legally.

Record companies are engaged intensely in providing customers with an alternative to Napster that will generate income for the record industry and prevent piracy of music. The major subscription Web sites for music will probably concentrate on contemporary music and the history of rock and roll (Surowiecki 2000). The universe of musical riches promised by celestial jukeboxes is not likely to include a wide selection of historical sound recordings that represent the full breadth of recorded music. This is certain to be true if they are not preserved and documented properly. If audio recordings that do not have mass appeal are to be preserved, that responsibility will probably fall to libraries and archives. Within a partnership between archives and intellectual property owners, archives might assume responsibility for preserving less commercial music in return for the ability to share files of preserved historical recordings.

All audio preservation is expensive; it is estimated that preservation engineers' studio time required for a recording averages three times the length of the source recording. Digital preservation holds great promise but it adds significant investment costs, such as the creation and maintenance of repositories and the generation of controlling metadata. Whether for lack of foresight or funding, libraries are not creating digital mass-storage systems for audiovisual works, which are common in broadcasting archives. We face an extraordinary dilemma: at a time when a greater range of audio is available to more people than ever before, and the means are finally at hand to preserve those sounds for posterity, we stand the greatest risk of losing them.

References

Holland, Bill. 1997. "Labels Strive to Rectify Past Archival Problems." *Billboard*. July 12 and July 19. Available at: www.chezmarianne.com/bholland/words/vault.html.

National Research Council. 2000a. *The Digital Dilemma: Intellectual Property in the Information Age*. Washington, D.C.: National Academy Press.

National Research Council. 2000b. *LC21: A Digital Strategy for the Library of Congress.* Washington, D.C.: National Academy Press.

Radio and Internet Newsletter. Available at: www.KurtHanson.com.

Schoenherr, Steven E. 2002. *Recording Technology History.* Available at: http://history.sandiego.edu/gen/recording/notes.html.

Surowiecki, James. 2000. "Can the Record Labels Survive the Internet?" *The New Yorker,* 5 June.

Understanding the Preservation Challenge of Digital Television

MARY IDE, DAVE MACCARN, THOM SHEPARD, AND LEAH WEISSE
WGBH Educational Foundation

Executive Summary

By nature and necessity, public broadcasting is a hodgepodge of media types and formats. A documentary might include moving and still images, speeches and voice-overs, sound effects, or a song. Children's programming might include a combination of live action, cartoons, musical numbers, and kaleidoscopic effects. Source material for any of these production elements might be analog (a strip of film, a track from a 78-rpm phonograph record) or digital (panoramic portraits, credit rolls, logos).

In whatever manifestations these objects previously existed, they become bits and bytes before they reach the public eye. That is an enormous amount of digital information to manage over time. A single second of uncompressed high-definition digital content would take up 150 megabytes of storage space. A minute would fill a home computer's 10-gigabyte hard drive. Although the holding capacity per unit volume doubles almost every two years, these technical advancements come at a cost: media obsolescence.

As we move into the increasingly complex digital world, those charged with preserving our television heritage have the opportunity to develop and establish better coordinated and standardized preservation policies and practices to ensure what television programs and related assets survive.

Introduction: Statement of Problem

In many respects, the dilemma of archiving digital content is the same as it was for analog: how do we preserve the substance of a medium while its physical containers decay or grow obsolete? For analog products, standard practice recommends procuring appropriate shelf space within a controlled environment. Digital objects may be handled in similar fashion—that is, as shelved artifacts—but this approach avoids examining the qualities that make digital both attractive and perilous for productions. Alternative digital-storage solutions are being marketed all the time. Each new option brings its own set of pitfalls as well as rewards. The bottom line: the storage industry has yet to solve the problem of technical obsolescence with the creation of an archive format.

Standard archival practice continues to advocate the refreshing of physical media. Refreshment strategies, which include migration and emulation, may prove effective for some types of media, but they are inadequate for handling the intricacies, interdependencies, and sheer volume of television content.

Over the past decade, television production and broadcasting have been moving from analog to digital. The analog method, which transmits sounds and pictures through continuous wavelike signals or pulses of varying intensity, is being replaced by digital capture and transmission in which sounds and images are converted into groups of binary code (ones and zeros). This transition is both complex and clouded. Materials collected or generated for a television show may consist of a great threaded mesh of digital and analog components, so tightly bound that, at any point in their life cycle, one may serve as a surrogate for another. What is analog today could be digital tomorrow. What is digital today may be stored as analog.

A look at the life cycle of a “production object” reveals myriad routes from the capture of the moving image to the airing of the broadcast. Footage is shot in a studio or on location and makes its way into a video editing system. If the source material is analog, a digital capture card converts the analog information into digital signals. Stills may be scanned from photographs and illustrations, then manipulated with software. What starts as a static image can end up as animation. A slow pan across a Civil War battlefield, a zoom into Mary Lincoln’s eyes—these become simulated camera movements, and the digital object that began as a JPEG (Joint Photographic Experts Group) or TIFF (Tag Image File Format) becomes an MPEG (Motion Picture Experts Group) video file.

Sound or audio tracks are also treated as distinctive elements in a television production. Whether it is background music, a voice-over, or the sound of water dripping, audio tracks must be maintained both as parts of the completed program and as entities unto themselves. The very same audio information might exist as a WAV file and be packaged within an MPEG.

In addition to materials that have clear analog sources, some materials may be created on desktop machines by teams of artists, designers, and computer programmers using a wide range of off-the-shelf software. A program logo, for example, may begin life as a Photoshop bitmap. It may then be transformed into an Illustrator vector graphic. This vector graphic may be imported into another application, rendered as a three-dimensional moving object, and incorporated into a show.

The very concept of a “finished program” is debatable. We have already witnessed the rising popularity of digital video disc (DVD) feature film “extras”: outtakes, cut segments, director's cuts, and alternative endings. Considering that an audience may see as little as 5 percent of the original footage shot for any given broadcast, there is an enormous long-term potential market in providing them some leftovers. What remains to be explored is the full value of the original source materials for nonfiction productions: unedited interviews or other documentary footage that lends itself to new interpretations as events unfold. We cannot predict the educational or entertainment value that audiences will derive from production materials, but current trends indicate that there is wisdom in saving it all.

How Are Items Selected for Collection and Preservation?

Radio and television broadcasting has been a major influence in shaping the political, social, cultural, and economic trends of the twentieth century. Broadcasting has heightened citizen awareness of our global community and its diversity. The broadcast industry's recordings and related production materials are primary sources for the study of history and culture. The media mirror the world; they also change our perceptions of the world and draw us into it. Television "is not just a new way of doing old things but a radically different way of seeing and interpreting the world" (Kernan 1990, 151).

Current appraisal methodologies used to select television programs for preservation suggest a hybrid of the methods traditionally applied to textual materials. Appraisal for selection requires a significant level of knowledge about the moving-image production process and analog and digital production technologies. The appraisal criteria must also take into consideration the technical and financial preservation commitment implications. The fragility of moving images and the rapid advancements in reformatting technologies complicate the ethical and practical accessioning and appraisal process.

Guidelines or standards for selecting television material for preservation are valuable resources. One of the earliest and most comprehensive international television appraisal studies was the 1983 Record and Archives Management Programme (RAMP) study, prepared for the United Nations Educational, Scientific, and Cultural Organization (UNESCO) by Sam Kula. In his RAMP report, Kula acknowledged that selection criteria tend to first meet the needs of broadcasters, and the potential for reuse of programming content is particularly important. Re-use potential also considers the intrinsic historical or cultural value of content (Kula 1990).

The Fédération Internationale des Archives de Télévision/International Federation of Television Archives (FIAT/IFTA) is a Europe-based organization of archivists who manage television archival material. FIAT developed the following criteria for master television program selection in 1996:

- material of historic interest in all fields
- material as a record of a place, an object, or a national phenomenon
- interview material of historic importance
- interview material indicative of opinions or attitudes of the time

- fictional and entertainment material of artistic interest
- fictional and entertainment material illustrative of social history
- any material, including commercial and presentational, illustrative of the development of television practices and techniques (Library of Congress 1997, 189)

Commercial and public broadcasting stations and other collecting institutions have developed their selection criteria on the basis of their institutional needs and missions. But for any collecting institution, the preservation commitment, whether for digital or analog materials, is staggering in cost and maintenance. The time has come to encourage and explore the concept of regional and national planning for the preservation of broadcast television programming.

The Library of Congress (LC) study, *Television and Video Preservation 1997: A Study of the Current State of American Television and Video Preservation,* outlines the state of American preservation practices and calls for a concerted national and regional effort to plan for the preservation of American television programming. Librarian of Congress James H. Billington says in the study's preface that "at present, chance determines what television programs survive. Future scholars will have to [rely] on incomplete evidence when they assess the achievements and failures of our culture" (Library of Congress 1997, xi).

Standard Formats for Digital Television

Standards for digital television include not only the formats for the physical media but also for the broadcast stream itself. The current analog broadcast standard, for example, has an image resolution of 525 horizontal lines and 640 vertical lines or pixels. To understand what this means, consider that a home computer monitor is likely to have a resolution of 800 by 600 or better. In contrast, the standard resolution for high-definition television (HDTV) is 1080 lines and 1920 pixels. In addition, the aspect ratio for HDTV is 16:9, while the standard for conventional TV is 4:3. As the numbers suggest, HDTV holds a great deal of promise for today's viewing audience, yet increases the amount of information available. These numbers also point to a problem: how can this extra information be transported through the same broadcast pipeline?

The Advanced Television Systems Committee (ATSC) Digital Television Standard (A-53) was devised to increase the amount of broadcast information allowable through a conventional 6-MHz channel. A finished program might be transported directly from an editing station, set up in the control room as a compressed MPEG-2 video file, and broadcast to home analog television sets, and may additionally be transferred to an archival storage system or media. Although the A-53 standard is regulated across the United States, the problems of physical storage for this material are growing more complex.

Since 1987, at least 17 digital videotape formats have come into the marketplace, and, as with analog tape, competing and incompatible formats proliferate. The format issue alone is a nightmare for collecting institutions for two reasons: (1) formats are platform-dependent to particular playback machines; and (2) physical media require constant migration to new formats.

Videotape is a notoriously fragile medium made up of three major components: the backing, the magnetic coating, and the binder that holds the magnetic coating to the backing. While the life expectancy of videotape is, at best, 15 to 20 years, time and experience have shown that the older analog videotape formats are sturdier and last longer than newer ones do.

Some digital video formats use compression. Compression can dramatically reduce the size of a data file by eliminating redundant information by taking advantage of the psycho-visual studies of human perception. Some compression techniques are proprietary. Because manufacturer's implementations vary, they produce "unanticipated consequences such as a phenomenon called 'concatenation,' in which artifacts of the compression process make it difficult to transfer content to new formats" (Liroff 2001, 8).

While the specifications for DVDs were being hammered out, hopes were high in the archival community that it might serve as an adequate preservation vehicle. Now, the consensus among moving-image archivists is more pessimistic. Though regarded as an advancement in distribution and access, the DVD, like the CD and the CD-ROM that it physically resembles, is subject to deterioration from oxidation, humidity, and physical damage. In addition, there is no guarantee that the format will not become obsolete within another generation. That said, technologies and materials might improve to the extent that the archival community might reevaluate the DVD format. Perhaps a "backward-compatible" DVD format might be developed for purely archival use.

Organizational Issues

Organizational issues concerning digital television content include asset and rights management, distribution channels, and user purposes and needs. Solutions to these issues will vary with an institution's mission. Because this is a transition period of analog to digital, traditional and nontraditional methods of dealing with organizational issues are currently used in tandem.

Asset and Rights Management

Over the past 20 years, an expanding market for production repurposing has encouraged the practice of keeping edited master programs and related production elements. Also, the advent of smaller tape formats has allowed us to store more individual items. Digital asset management (DAM) systems provide access to and storage for these rich media assets, which are digitally indexed and often associated to specific rights management information.

Digital rights management (DRM) entails tracking rights of each creating entity, controlling access, security issues, collecting payments, and distribution. A producing entity must track copyright-related data including insurance agreements, trademark issues, talent payments, licensing and market agreements, co-production payments, and financial support.

The breakdown of program material into segments is crucial to rights management. Segmentation is not only vertical but also horizontal. Attributes must be logged for each component part. For example, music or narration for a program needs to be available as a stand-alone component, if only to allow editors to remove it for rebroadcast. Rights information needs to be applied to each of these components.

Product placement through digital manipulation may factor into how we manage moving-image materials. Though highly controversial, experiments are under way in commercial television to set up product placement variables within dramatic scenes. Flexibility in product placement may be particularly lucrative when a show is licensed for syndication. For example, one version might show a can of Pepsi-Cola as a strategically placed prop. In another market, that image might be digitally turned into a can of Coca-Cola. Though it is hard to imagine the public affected by product placement, it is conceivable that just as cable markets license our programs, we may indeed see product placement as a requirement for licensing.

Distribution

There are multiple program distribution routes, including broadcast transmission, home video, satellite, cable, and Webcasting. By the year 2003, the Federal Communications Commission has mandated that all commercial and public broadcasting stations will have to convert to the digital television (DTV) transmission standard. Once digital TV is widespread, broadcast materials will exist in several versions and formats. DTV will expand broadcasting capabilities to include three formats: HDTV, multicasting, and datacasting. The highest quality will be HDTV, providing an image far superior to that available on analog sets.

Multicasting would permit multiple programs to be carried by one broadcast signal, allowing broadcasters, such as cable systems, to increase the amount of programming available as well as to target viewer demographics. It could also allow viewers to experience alternative angles of a particular broadcast. Live drama, breaking news events, and sports telecasts would benefit from multicasting.

Datacasting, as its name implies, allows data (video, audio, text, graphics, maps, and services) to be embedded in the broadcast signal for downloading into a computer or set-top box, allowing the broadcasting of ancillary materials to accompany a program. These materials may be accessible as downloadable data that may be collected and accessed through computers, or as streaming content that may be viewed on a designated portion of a television screen. Datacasting could give viewers immediate access to a wealth of supplementary material, such as cast lists, biographies, and transcripts. These features are like the “extras” that are included in many current DVDs.

New technologies continue to up the ante for audience expectations. Today, we want our video on demand. Tomorrow, we will have a side order of metadata. As long as there are audiences hungry for both quantities and varieties of information, there will be industries to supply those needs. As television grows more Weblike, providing easy access to enormous amounts of digital information through digital hyperlinks, those charged with the preservation and access to television content will play a key role and perhaps in the process will finally win public recognition for their efforts.

Users

A measure of how the public uses digital assets is reflected in the coined term, "edutainment." The expression has caught on throughout the world and is used in several languages. Literally, it is the melding of the words "education" and "entertainment." Figuratively, it means "learning that is fun." What is often missing in academic discussions of electronic information is the "fun factor." Even tools for data retrieval, for example, are not only getting more attractive but also becoming easier to use.

The user base stretches beyond the general public: education professionals, researchers, the production community, and others have also embraced new technologies. All are benefiting from the use of television production assets created specifically for curriculum research, distance learning, and classroom reference. Moving-image collections have been developing Web sites for use by educators such as the WGBH New Television Workshop Project.

WGBH's National Center for Accessible Media (NCAM) makes public media accessible to disabled persons, minority language users, people with low literacy skills, and other underserved populations. For example, it offers closed captioning and descriptive video services (DVS) for those with special hearing and sight needs. NCAM researches and develops media access technologies and explores how existing technologies may benefit other populations. These access technologies create another set of production assets.

Implicatons for Long-Term Preservation

Storage

A distinction must be made between how we preserve broadcast materials and how we access them over time. Preserving data is crucial, but how readily available will these materials need to be? Offline storage takes the longest time to retrieve. It is usually boxed and stored on a shelf but is cataloged and available. Nearline storage provides intermediate access. Nearline storage is linked to the concept of the "jukebox" system—a collection of optical or tape drives that reside in a hardware device consisting of numerous slots, or "bays," and a robotic arm. The stored data are not instantly accessed, but instead are retrieved through various human or mechanical means. Online storage provides the most immediate access, typically spinning disk, possibly SAN (storage area network) or NAS (network attached storage), accessible through file systems and Internet/LANs (local area networks). In hardware terms, an *online* storage device is one that is perpetually available to authorized users. Digital storage will be so cheap in years to come that it will be possible to keep exact copies of our materials in several distinct locations at a relatively low cost. This "redundant" storage would help protect assets in times of disaster. On the other hand, limitless storage introduces new problems of access and management.

There are basically two approaches to storing digital video images. We can store whole programs and create databases that contain metadata. And we can store all of the clips that are included in the program as separate files and then rely on edit decision lists (EDLs) to serve as blueprints for our broadcasts. Both options rely on some form of stratification of the media. *Stratification* is a system of video annotation that uses time-codes to identify marking points within an audio or video object. Descriptions can be linked to these points by storing them with the time-code information. In the same way that video may contain many tracks, metadata may also have several layers, each with its own set of referenced time-codes. For example, a transcript may occupy one metadata layer, while captioning information may occupy another. Other layers may include DVS material, copyright, or image content description.

Even as storage space becomes limitless and more reliable, we still need to grapple with the problem of software obsolescence. Storing the same information in many different standard and proprietary formats may be one way to protect our assets, but this approach will require a great dependency on software tools to keep track of them. Broadcast materials are built upon a hierarchy: series, program, segment, clip, and even a single frame. Tools will have to be robust enough to manage these materials on all levels. As Howard Besser writes, those concerned with preservation need "to move away from an artifact-based approach [to preservation] and instead adopt an approach that focuses on stewardship of disembodied digital information" (Besser 2001, 4).

Proposed Solutions

In the archival communities, the debate over digital preservation has focused on three strategies: migration, emulation, and bundling.

Migration is the process of moving data from a digital format that is determined to be obsolete to a platform that is currently in use. As a preservation strategy, migration is prone to bad judgment calls. As a technical solution, migration may damage the essence of the material by dropping crucial data that could result in its loss of function or in its original look and feel.

Emulation approaches the problem through a kind of a virtual time machine. It aims to sustain a digital object's original look and feel by mimicing the application that created the object, the operating system upon which the application ran, and the hardware platform upon which the operating system was housed. This is not a one-time, fix-all strategy. Emulation software will have its own hardware and operating system dependencies. The virtual time machine itself may have to be emulated.

A problem with emulation specific to audio and image content is the possibility that the original playback application is limited as compared with later versions or other applications. In other words, the application that created the data file may not be the best application for playing it back. A digital media file often contains more information than may be displayed through its current application. For example, a moving-image file may be exported from a software application at a greater resolution than the application itself can display. Metadata fields may be hidden from the current application but available or reserved for future versions. In other words, the emulation time machine may need to know which version of an application best captures or extracts the data.

Bundling is the process of bonding metadata with content within the same file format. This bundling may include information about the provenance of a particular item. The Universal Preservation Format (UPF), which was proposed by WGBH, uses a data file mechanism that bundles metadata with the data representing the actual image, sound, or text. The metadata identify this data "essence" within a registry of standard data types and serve as the source code for mapping or translating binary composition into accessible or usable forms. The UPF is designed to be independent of the computer applications used to create content, of the operating system from which these applications originated, and of the physical medium upon which that content is stored. The UPF is characterized as "self-described" because it includes, within its metadata, all the technical specifications required to build and rebuild appropriate media browsers to access contained materials throughout time.

Other initiatives that use bundling or packaging include the Open Archival Information System (OAIS) and the Digital Rosetta Stone Model.

Longevity Problems

Howard Besser (2000, 156) outlines five longevity problems specific to preserving all digital records:

1. The *viewing* problem is the fact that electronic content is stored on physical devices that deteriorate and require proactive planning to migrate and assure longevity.
2. The *translation* problem focuses on understanding that "work translated into new delivery devices changes meaning" (Besser 2001, 3). A simple example is a motion picture resized for the television screen.
3. The *custodial* problem concerns determining who will be responsible for the long-term preservation and authentication of digital content. Will it be archivists, computer technologists, others, or a collaboration of many?
4. The *scrambling* problem for digital television is twofold and relates to the compromise of using compression techniques to satisfy limited storage and bandwidth transmission capabilities and encryption schemes to protect content, which make future access potentially a problem. Compression compromises the integrity of original content, and encryption adds another layer of complexity to a fragile digital object.

5. The *interrelational* problem concerns the complexity of related information to and within a digital object. Because boundaries of information sets or digital objects are not usually defined, this raises not only custodial concerns but also intellectual property concerns.

Unresolved Issues

Paul Messier (1996, 3) has suggested that an adequate digital video preservation plan should do the following:

- make a format accessible on standard equipment at various levels of access
- capture image at the highest-possible quality resolution rate using minimum or no compression
- develop guidelines for digital conversion that are based on the type of source material
- use formats and equipment that meet national and international standards
- ensure a data-migration path that is a hedge against format and machine obsolescence

Standards for cataloging moving-image materials are continually in evolution. The Library of Congress has set the most prevalent standard. Techniques for creating access to digital content on an international scale include the Dublin Core initiative and MPEG-7, to name a few. The Dublin Core, being developed by international cross-disciplinary groups, is a set of 15-plus basic information metadata fields for identifying content and access points. Working groups within the Dublin Core metadata initiative are proposing enhancements to this basic set of tags that address cataloging needs of specific industries or domains. These “application profiles” are being proposed for education, libraries, and bibliographic citations, among others. Some researchers have begun to lay the foundation for an application profile for static and moving-image and audio files.   MPEG-7 is the Multimedia Content Description Interface standard developed by the MPEG, whose goal is to provide a rich set of standardized metadata fields to describe multimedia content.

Ethical issues concern maintaining the integrity of original content and intent; this is particularly acute with digital morphing capabilities to change and manipulate images in ways that cannot be detected. Included in this dilemma is compression of files that can compromise original intent and artistic authenticity. For example, when moving-image materials are available only as low-resolution digital files or scanned from older analog formats, pixels might be filled in to give the illusion of a higher density resolution. Finally, there are the issues of adherence to copyright law, protection of privacy rights, and confidentiality.

In the not-too-distant future, the line between moving-image distribution and moving-image projection may fade completely. Already there have been experiments in which a motion picture was transmitted from a remote location and projected into a movie theater. The first such test occurred on June 6, 2000, when Cisco

Systems Inc. joined with Twentieth Century Fox to digitally transmit Titan A. E. from Burbank, California, to the Woodruff Arts Center in Atlanta, Georgia. The notion of an "artifact-free" method of distribution will have a great impact on preservation. Instead of moving digital information to tapes for distribution, data will simply consist of a file transfer to some temporary storage device, which might periodically be wiped clean. Failure to assign clear responsibility for preserving these broadcast materials may result in tremendous losses.

The issue of who is responsible for the preservation of digital content has not been satisfactorily resolved. Preservation of digital content must be a collaborative effort that involves the professional archivist, the technology expert, the user, and the creating and producing entity.

Inaction on the preservation front will ensure the continued loss of the nation's television heritage. As stated in the LC study, "all organizations having custody of American television and video materials, whether private or public bodies, should recognize their responsibilities for preserving a part of the historical and cultural heritage" (Library of Congress 1997, 123).

References

Besser, Howard. 2000. Digital Longevity. In *Handbook for Digital Projects: A Management Tool for Preservation and Access,* edited by Maxine Sitts. Andover, Mass.: Northeast Document Conservation Center.

Besser, Howard. 2001. Digital Preservation of Moving Image Material? Available at: www.gseis.ucla.edu/~howard/Papers/amia-longevity.html.

Council on Library and Information Resources. 2000. *Authenticity in a Digital Environment.* Washington, D.C.: Council on Library and Information Resources.

Gilliland-Swetland, Anne J., and Philip B. Eppard. 2000. Preserving the Authenticity of Contingent Digital Objects. *D-Lib Magazine* 6(7-8). Available at: www.dlib.org/dlib/july00/eppard/07eppard.html.

Gilliland-Swetland, Anne J. 1999. The Long-Term Preservation of Authentic Electronic Records: InterPARES. Speech presented at the Society of American Archivists Annual Meeting, Pittsburgh, Pa., August 28.

Granger, Stewart. 2000. Emulation as a Digital Preservation Strategy. *D-Lib Magazine* 6(10). Available at: www.dlib.org/dlib/october00/granger/10granger.html.

Hunter, Gregory S. 2000. *Preserving Digital Information: A How-To-Do-It Manual,* no 93. New York: Neal-Schuman Publishers, Inc.

Hunter, Jane. 1999. MPEG-7 Behind the Scenes. *D-Lib Magazine* 5(9). Available at: www.dlib.org/dlib/september99/hunter/09hunter.html.

Kernan, Alvin. 1990. *Death of Literature.* New Haven: Yale University Press.

Kula, Sam. 1990. Selected Guidelines for the Management of Records and Archives: A RAMP reader. PGI-90/WS/6. Paris: UNESCO. Available at: http://www.unesco.org/webworld/ramp/html/r9006e/r9006e00.htm#Contents

Library of Congress. 1997. T*elevision and Video Preservation: A Report of the Current State of American Television and Video Preservation.* 3 vols. Washington, D.C.: Library of Congress.

Lindner, Jim. 1998. *Digitization Reconsidered.* Available at: www.vidipax.com/articles/digirecon.html.

Liroff, David. 2001. Media Asset Management—The Long-Term View. Speech presented at the Sun Microsystems Digital Media Universe, Beverly Hills, Calif., August 21.

MacCarn, Dave. 2000. *Toward a Universal Data Format for the Preservation of Media.* Available at: http://info.wgbh.org/upf/papers/SMPTE_UPF_paper.html.

Messier, Paul. 1996. Criteria for Assessing Digital Video as a Preservation Medium. *Bay Area Video Coalition (BVAC) Playback 1996 [Conference] Report to the Field.* San Francisco: Bay Area Video Coalition.

National Research Council. 2001. *LC21: A Digital Strategy for the Library of Congress: Executive Summary.* Available at: http://stills.nap.edu/books/0309071445/html/.

OCLC/RLG Working Group on Preservation Metadata. 2001. Preservation Metadata for Digital Objects: A Review of the State of the Art. January 31.

Sadashige, Koichi, 2000. Data Storage Technology Assessment 2000. Available at: http://www.nta.org/Bibliography/techreports/part1.htm.

Su-Shing Chen. 2001. The Paradox of Digital Preservation. Computer 34(3): 24–28.

Wheeler, Jim. Video Q&A. *Newsletter of the Association of Moving Image Archivists.* 49;34.

WGBH New Television Workshop Project. Available at http://main.wgbh.org/wgbh/NTW.

Digital Video Archives: Managing Through Metadata

HOWARD D. WACTLAR AND MICHAEL G. CHRISTEL
Computer Science Department
Carnegie Mellon University

Executive Summary

As analog video collections are digitized and new video is created in digital form, computer users will have unprecedented access to video material—getting what they need, when they need it, wherever they happen to be. Such a vision assumes that video can be adequately stored and distributed with appropriate rights management, as well as indexed to facilitate effective information retrieval. The latter point is the focus of this paper: how can metadata be produced and associated with video archives to unlock their contents for end users?

Video that is "born digital" will have increasing amounts of descriptive information automatically created during the production process, e.g., digital cameras that record the time and place of each captured shot, and tagging video streams with terms and conditions of use. Such metadata could be augmented with higher-order descriptors, e.g., details about actions, topics, or events. These descriptors could be produced automatically through ex-post-facto analysis of the aural and visual contents in the video data stream. Likewise, video that was originally produced with little metadata beyond a title and producer could be automatically analyzed to fill out additional metadata fields to better support subsequent information retrieval from video archives.

As digital video archives grow, both through the increasing volume of new digital video productions and the conversion of the analog audiovisual record, the need for metadata similarly increases. Automatic analysis of video in support of content-based retrieval will become a necessary step in managing the archive; a recent editorial by the director of the European Broadcasting Union Technical Department notes that

"Efficient exploitation of broadcasters' archives will increasingly depend on accurate metadata" (Laven 2000). He offers the challenge of finding an aerial shot of the Sydney Harbour Bridge at sunset. Given a small collection of Sydney videos, such a task is perhaps tractable, but as the volume of video grows, so does the importance of better metadata and supporting indexing and content-based retrieval strategies.

Digital library research has produced some insights into automatic indexing and retrieval. For example, it has found that narrative can be extracted through speech recognition; that speech and image processing can complement each other; that metadata need not be precise to be useful; and that summarization strategies lead to faster identification of the relevant information. The purpose of this chapter is to discuss these findings. Particular emphasis is placed on the Informedia Project at Carnegie Mellon University and the new National Institute of Standards and Technology Text Retrieval Conference (NIST TREC) Video Retrieval Track, which is investigating content-based retrieval from digital video.

Introduction

We are faced with a great opportunity as analog video resources are digitized and new video is produced digitally from the outset. The video itself, once encoded as bits, can be copied without loss in quality and distributed cheaply and broadly over the ever-growing communication channels set up for facilitating transfer of computer data. The great opportunity is that these video bits can be described digitally as well, so that producers' identities and rights can be tracked and consumers' information needs can be efficiently, effectively addressed. The "bits about bits" (Negroponte 1995), referred to as "metadata" throughout this paper, allow digital video assets to be simultaneously protected and accessed. Without metadata, a thousand-hour digital video archive is reduced to a terabyte or greater jumble of bits; with metadata, those thousand hours can become a valuable information resource.

Metadata for video are crucial when one considers the huge volume of bits within digital video representations. When digitizing an analog signal for video, the signal needs to be sampled a number of times per second, and those samples quantized into numeric values that can then be represented as bits. Only with infinite sampling and quantization could the digital representation exactly reproduce the analog signal. However, human physiology provides some upper bounds on differences that can actually be distinguished. For example, the human eye can typically differentiate at most 16 million colors, and so representing color with 24 bits provides as much color resolution as is needed for the human viewer. Similar visual physiological factors on critical viewing distance and persistence of vision establish other guidelines on pixel resolution per image and images per second playback rate. For a given screen size and viewer distance, 640 pixels per line and 480 lines per image provide adequate resolution, with 30 images per second resulting in no visible flicker or break in motion. Digital video at these rates requires 640 x 480 x 30 x (24 bits per pixel) = 221 megabits per second, or 100 gigabytes per hour. The number of bits increases if higher resolution (such as high-density TV [HDTV] resolution of 1920 by 1080) is desired (for example, to allow for larger displays viewed at closer distances without distinguishing the individual pixels). Hence, even a single hour of video can result in 100 gigabytes of data. Associating metadata with the video makes these gigabytes of data more manageable.

Numerous strategies exist to reduce the number of bits required for digital video, from relaxed resolution requirements to lossy compression in which some information is sacrificed in order to reduce significantly the number of bits used to encode the video. Motion Picture Experts Group-1 (MPEG-1) and MPEG-2 are two such lossy compression formats; MPEG-2 allows higher resolution than MPEG-1 does. Because preservationists want to maintain the highest-quality representation of artifacts in their archives, they are predisposed against lossy compression. However, the only way to fit more than a few seconds of HDTV video onto a CD-ROM is through lossy compression. The introduction to scanning by the Preservation Resources Division of OCLC Online Computer Library Center, Inc., reflects this tension between quality and accessibility:

> *Although traditional preservation methods have ensured the longevity of endangered research materials, it has sometimes been at the cost of reduced access. With digital technology, images are used to reproduce rare items, allowing for virtually universal copying, distribution, and access. The technology also makes it possible to bring collections of disparate holdings together in digital form, making resource sharing more feasible (OCLC 1998).*

Hence, for long-term preservation, digital video presents a number of challenges. What should the sampling and quantization rates be? What compression strategies should be used—lossy or lossless? What media should be used to store the resulting digital files—optical (such as digital video disc [DVD]) or magnetic? What is the shelf life for such media, i.e., how often should the digital records be transferred to new media? What are the environmental factors for long-term media storage? What decompression software needs to exist for subsequent extraction of video recordings? These challenges are not discussed further here, as they warrant their own separate treatments. Regardless of how these challenges are addressed, digital video has huge size, but also huge potential, for facilitating access to video archive material.

Digital technology has the potential to improve access to research material, allowing access to precisely the content sought by an end user. This implies full content search and retrieval, so that users can get to precisely the page they are interested in for text, or precisely the sound or video clip for audio or video productions. Creating such metadata by hand is prohibitively expensive and inappropriate for digital video, where much of the metadata is a by-product of the way in which the artifact is generated. Current research will extend the automated techniques for contemporaneous metadata creation.

To realize this potential, video must be described so that its production attributes are preserved and so users can navigate to the content meeting their needs. Video has a temporal aspect, in which its contents are revealed over time, i.e., it is isochronal. Finding a nugget of information within an hour of video could take a user an hour of viewing time. Delivering this hour of video over the Internet, or perhaps over wireless networks to a personal digital assistant (PDA) user, would require the transfer of megabytes or gigabytes of data. Isochronal media are therefore expensive both in terms of network bandwidth as well as user attention. If, however, metadata enabled surrogates to be produced or extracted that either were nonisochronal or significantly shorter in duration, then both bandwidth and the user's attention could be used more efficiently. After checking the surrogate, the user could decide whether access to the video was really necessary. A surrogate can also pinpoint the region of interest within a large video file or video archive.

As video archives grow, metadata become increasingly important: "In spite of the fact that users have increasing access to these [digitized multimedia information] resources, identifying and managing them efficiently is becoming more difficult, because of the sheer volume" (Martinez 2001). The capability of metadata to enrich video archives has not been overlooked by research communities and industry. For example, a number of workshops addressed this topic as part of digital asset management (DAM) (USC 2000). Artesia Technologies (Artesia 2001) and Bulldog (Bulldog 2001) are two corporations offering DAM products. Digital asset management refers to the improved storage, tracking, and retrieval of digital assets in general. Our focus here is on digital video in particular, beginning with a discussion of relevant metadata standards and leading to the automatic creation of video metadata and implications for the future.

Metadata for Digital Video

As noted in a working group report on preservation metadata (OCLC 2001), metadata for digital information objects, including video, can be assigned to one of three categories (Wendler 1999):

1. *Descriptive:* facilitating resource identification and exploration
2. *Administrative:* supporting resource management within a collection
3. *Structural:* binding together the components of more complex information objects

The same working group report continues that of these categories, "descriptive metadata for electronic resources has received the most attention—most notably through the Dublin Core metadata initiative" (OCLC 2001, 2). This paper likewise will emphasize descriptive metadata, while acknowledging the importance of the other categories, as descriptive metadata can be automatically derived in the future for added value to the archive. Further details on administrative and structural metadata are available in the 2001 OCLC white paper and its references.

Various communities involved in the production, distribution, and use of video have addressed the need for metadata to supplement and describe video archives. Librarians are very concerned about interoperability and having standardized access to descriptors for archives. Producers and content rights owners are greatly interested in intellectual property rights (IPR) management and in compliance with regulations concerning content ratings and access controls. The World Wide Web Consortium (W3C) produces recommendations on XML, XPath, XML-Schema, and related efforts for metadata formatting and semantics. Special interest groups such as trainers and educators have specific needs within particular domains, e.g., tagging video by curriculum or grade level. This section outlines a few key standardization efforts affecting metadata for video.

Dublin Core

The Dublin Core Metadata Initiative provides a 15-element set for describing a wide range of resources. While the Dublin Core "favors document-like objects (because traditional text resources are fairly well understood)" (Hillman 2001), it has been tested against moving-image resources and found to be generally adequate (Green 1997). The Dublin Core is also extensible, and has been used as the basis for other metadata frameworks, such as an ongoing effort to develop interoperable metadata for learning, education, and training, which could then describe the resources available in libraries such as the Digital Library for Earth System Education (DLESE) (Ginger 2000). Hence, Dublin Core is an ideal candidate for a high-level (i.e., very general) metadata scheme for video archives. An outside library service, with likely support for Dublin Core, would then be able to make use of information drawn from video archives expressed in the Dublin Core element set.

Video Production Standardization Efforts

Professional video producers are interested in tagging data with IPR, production and talent credits, and other information commonly found in film or television credits. In addition, metadata descriptors from the basic Dublin Core set are too general to adequately describe the complexity of a video. For example, one of the Dublin Core elements is the instantiation date (Hillman 2001), but for a video, date can refer to copyright date, first broadcast date, last broadcast date, allowable broadcast period, date of production, or the setting date for the subject matter.

Producers are especially interested in defining metadata standards because video production is becoming a digital process, with new equipment such as digital cameras supporting the capture of metadata such as date, time, and location at recording time. The Society of Motion Picture and Television Engineers (SMPTE) has been working on a universal preservation format for videos, the SMPTE Metadata Dictionary (SMPTE 2000). For born-digital material, many of the metadata elements can be filled in during the media creation process.

The SMPTE Metadata Dictionary has slots for time and place, further resolved into elements such as time of production and time of setting, place of production and place setting, where place is described both in terms of country codes and place names as well as through latitude and longitude. The SMPTE effort is often cited by other video metadata efforts as a comprehensive complement to the minimalist Dublin Core element set.

In 1999, the European Broadcasting Union (EBU) launched a two-year project named "EBU Project P/Meta" designed to develop a common approach to standardizing and exchanging program-related information and embedded metadata throughout the production and distribution life cycle of audiovisual material. According to 1999 press releases, the project began by identifying and standardizing the information commonly exchanged between broadcasters and content providers, using the BBC's Standard Media Exchange Framework (SMEF) as the reference model. They then were to assess the feasibility of applying new SMPTE metadata standards within Europe to support the agreed exchange framework, and move toward implementation.

The TV Anytime Forum is an association of organizations that seeks to develop specifications to enable audiovisual and other services based on mass-market, high-volume digital storage.

MPEG-7 and MPEG-21

A number of professional industry and consortia standardization efforts are in progress to provide more detailed video descriptors. The new member of the MPEG family, Multimedia Content Description Interface, or MPEG-7, aims at providing standardized core technologies allowing description of audiovisual data content in multimedia environments. It will extend the limited capabilities of proprietary solutions in identifying content that exist today, notably by including more data types. An overview of MPEG-7 by Martinez (2001) acknowledges the diversity of standardization efforts and notes the purpose of MPEG-7:

> *MPEG-7 addresses many different applications in many different environments, which means that it needs to provide a flexible and extensible framework for describing audiovisual data. Therefore, MPEG-7 does not define a monolithic system for content description but rather a set of methods and tools for the different viewpoints of the description of audiovisual content. Having this in mind, MPEG-7 is designed to take into account all the viewpoints under consideration by other leading standards such as, among others, SMPTE Metadata Dictionary, Dublin Core, EBU P/Meta, and TV Anytime. These standardization activities are focused to more specific applications or application domains, whilst MPEG-7 tries to be as generic as possible. MPEG-7 uses also XML Schema as the language of choice for the textual representation of content description and for allowing extensibility of description tools. Considering the popularity of XML, usage of it will facilitate interoperability in the future.*

Because the descriptive features must be meaningful in the context of the application, they will be different for different user domains and different applications. This implies that the same material may be described using different types of features, tuned to the area of application. To take the example of visual material, a lower abstraction level would be a description of shape, size, texture, color, movement (trajectory), and position (where in the scene can the object be found?). For audio, a description at this level would include key, mood, tempo, tempo changes, and point of origin. The highest level would give semantic information, e.g., "This is a scene with a barking brown dog on the left and a blue ball that falls down on the right, with the sound of passing cars in the background." Intermediate levels of abstraction may also exist.

The level of abstraction is related to the way in which the features can be extracted: many low-level features can be extracted in fully automatic ways, whereas high-level features need human interaction.

Next to having a continuous description of the content, it is also required to include other types of information about the multimedia data. It is important to note that these metadata may also relate to the entire production, segments of it (e.g., as defined by time codes), or single frames. This enables granularity that can describe a single scene's action, limit that scene's redistribution because of its source, or classify that scene as inappropriate for child viewing because of its content.

- *Form:* An example of the form is the coding scheme used (e.g., Joint Photographic Experts Group [JPEG], MPEG-2), or the overall data size. This information helps in determining whether the material can be "read" by the user.
- *Conditions for accessing the material:* This includes links to a registry with IPR information, including such entries as owners, agents, permitted usage domains, distribution restrictions, and price.
- *Classification:* This includes parental rating and content classification into a number of predefined categories.
- *Links to other relevant material:* The information may help the user speed the search.
- *The context:* In the case of recorded nonfiction content, it is important to know the occasion of the recording (e.g., the final of 200-meter men's hurdles in the 1996 Olympic Games).

In many cases, it will be desirable to use textual information for the descriptions. Care will be taken, however, that the usefulness of the descriptions is as independent from the language area as is possible. A clear example where text comes in handy is in giving names of authors, films, and places.

Therefore, MPEG-7 description tools will allow a user to create, at will, descriptions (that is, a set of instantiated description schemes and their corresponding descriptors) of content that may include the following:

- information describing the creation and production processes of the content (director, title, short feature movie)
- information related to the usage of the content (copyright pointers, usage history, broadcast schedule)
- information about the storage features of the content (storage format, encoding)
- structural information on spatial, temporal, or spatio-temporal components of the content (scene cuts, segmentation in regions, region motion tracking)
- information about low-level features in the content (colors, textures, timbres, melody description)
- conceptual information of the reality captured by the content (objects and events, interactions among objects)
- information about how to browse the content in an efficient way (summaries, variations, spatial and frequency subbands)
- information about collections of objects
- information about the interaction of the user with the content (user preferences, usage history)

There is room for domain specialization within the metadata architectures, whether by audience and function (education vs. entertainment), genre (documentary, travelogue), or content (news vs. lecture), but there is also a risk of overspecificity. Because the technology continues to evolve, MPEG-7 is intended to be flexible.

The scope of MPEG-21 could be described as the integration of the critical technologies enabling transparent and augmented use of multimedia resources across a wide range of networks and devices to support functions such as content creation, content production, content distribution, content consumption and usage, content packaging, intellectual property management and protection, content identification and description, financial management, user privacy, terminals and network resource abstraction, content representation, and event reporting.

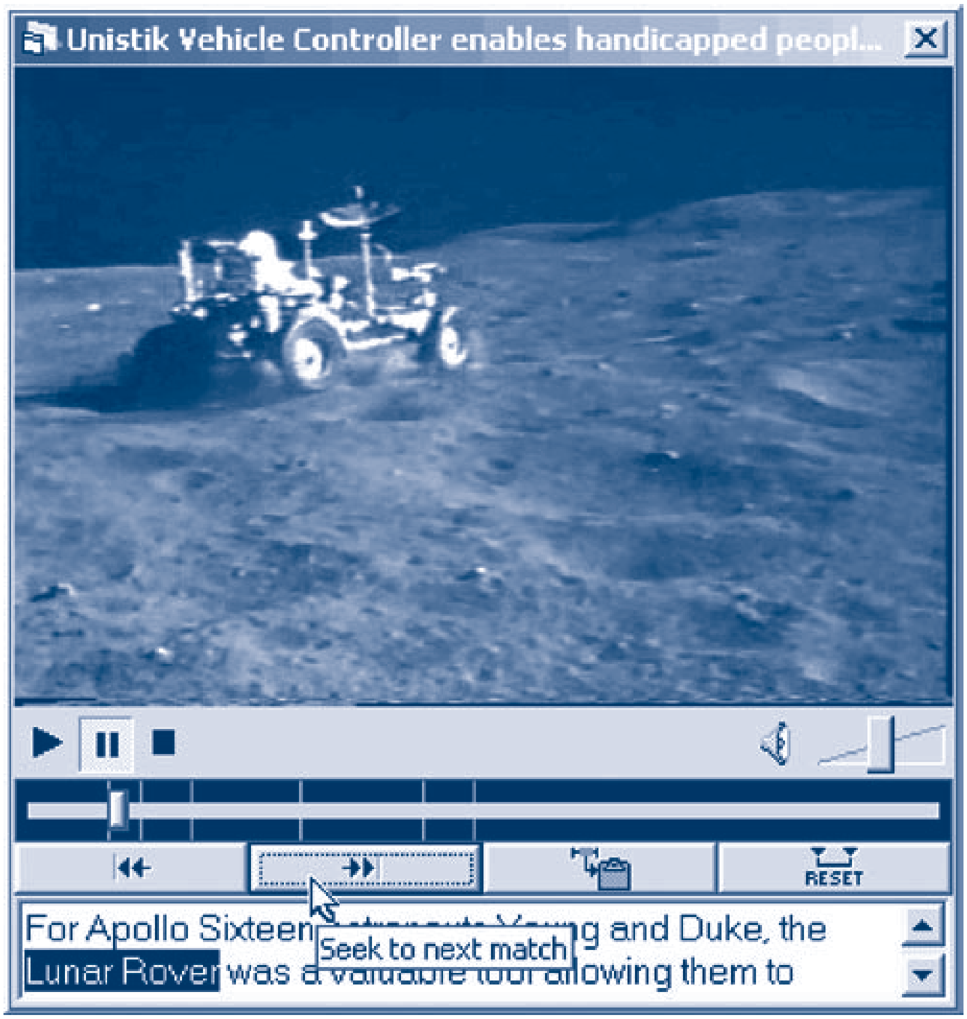
Standards for Web-Based Metadata Distribution

The W3C is a vendor-neutral forum of more than 500 member organizations from around the world set up to promote the World Wide Web's evolution and ensure its interoperability through common protocols. It develops specifications that must be formally approved by members via a W3C recommendation track. These specifications may be found on the W3C Web site.

A number of key W3C recommendations, published in 1999 and referenced below, enabled the separation of authoring from presentation in a standardized manner. For video archives, these recommendations allow the separation of video metadata from the library interface and from the underlying source material. This enables the interface to be customized for the particular application or audience (adult entertainment vs. secondary school education) and to the communication medium or device specifications (desktop PC vs. PDA), even though the same underlying data will be accessible to each use. The W3C recommendations useful for accessing, integrating, exploring, and transferring digital video metadata through the Web and Web browsers include the following:

- XML (Extensible Markup Language): the universal format for structured documents and data on the Web, W3C Recommendation February 1998 (http://www.w3.org/XML/)
- XML Schema: express shared vocabularies for defining the semantics of XML documents, W3C Recommendation as of May 2001 (http://www.w3.org/XML/Schema)
- XSLT (XSL Transformations): a language for transforming XML documents, W3C Recommendation November 1999 (http://www.w3.org/TR/xslt)
- XPath (XML Path Language): a language for addressing parts of an XML document, used by XSLT, W3C Recommendation November 1999 (http://www.w3.org/TR/xpath.html)

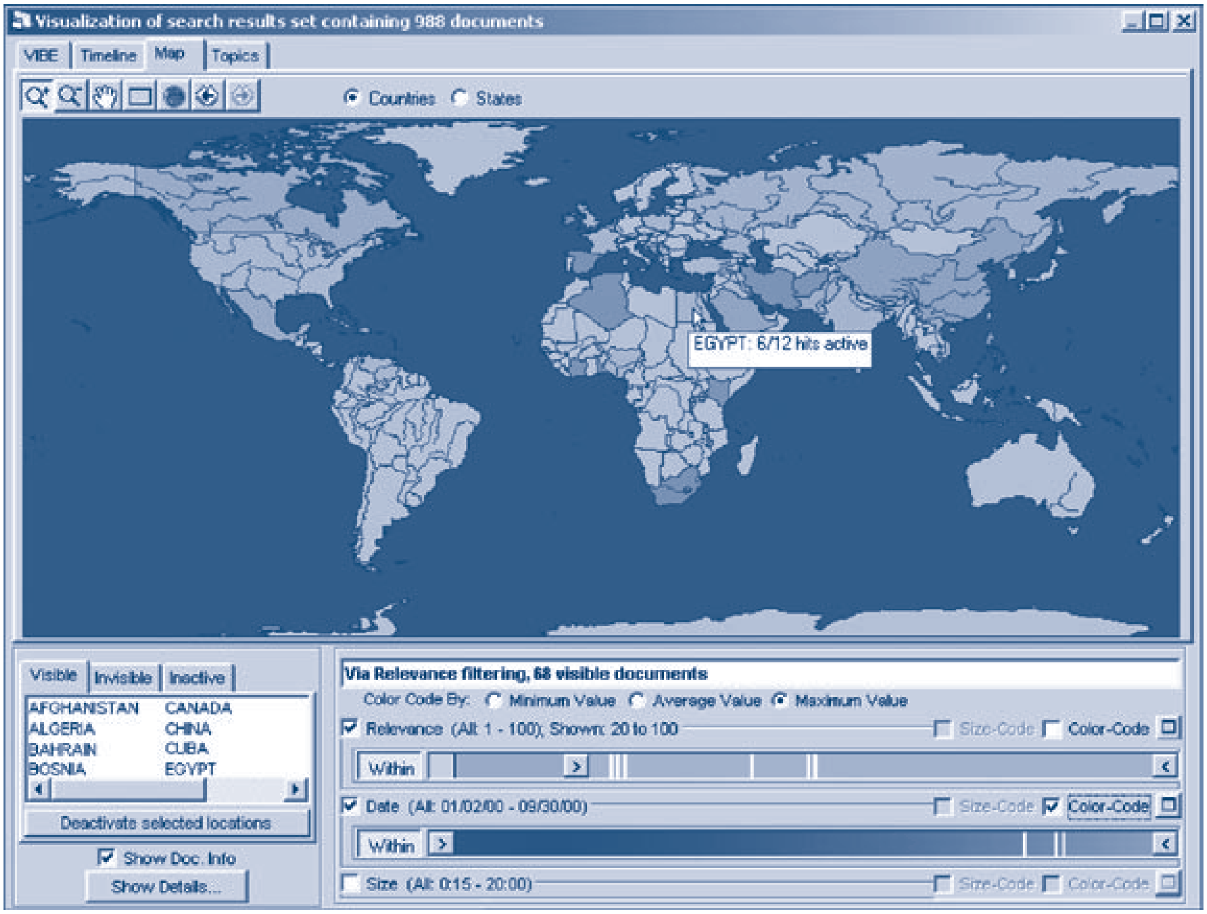
Figure 4. Effects of seeking directly to a match point on "Lunar Rover," courtesy of tight transcript to video alignment provided by automatic speech processing



Unistik Vehicle Controller enables handicapped peopl...

For Apollo Sixteen g and Duke, the
Lunar Rover was a owing them to

Figure 5. Map visualization for results of "air crash" query, with dynamic query sliders for control and feedback


Visualization of search results set containing 988 documents
VIBE
Timeline
Map
Topics
Visible
Invisible
Inactive
Deactivate selected locations
Show Details...

○ Countries ○ States

EGYPT: 6/12 hits active

AFGHANISTAN CANADA
ALGERIA CHINA
BAHRAIN CUBA
BOSNIA EGYPT

☑ Show Doc. Info

Via Relevance filtering, 68 visible documents

Color Code By: ○ Minimum Value ○ Average Value ● Maximum Value

☑ Relevance (All: 1 - 100); Shown: 20 to 100 □ Size-Code □ Color-Code

Within

☑ Date (All: 01/02/00 - 09/30/00) □ Size-Code ☑ Color-Code

Within

□ Size (All: 0:15 - 20:00) □ Size-Code □ Color-Code

Case Study: Informedia

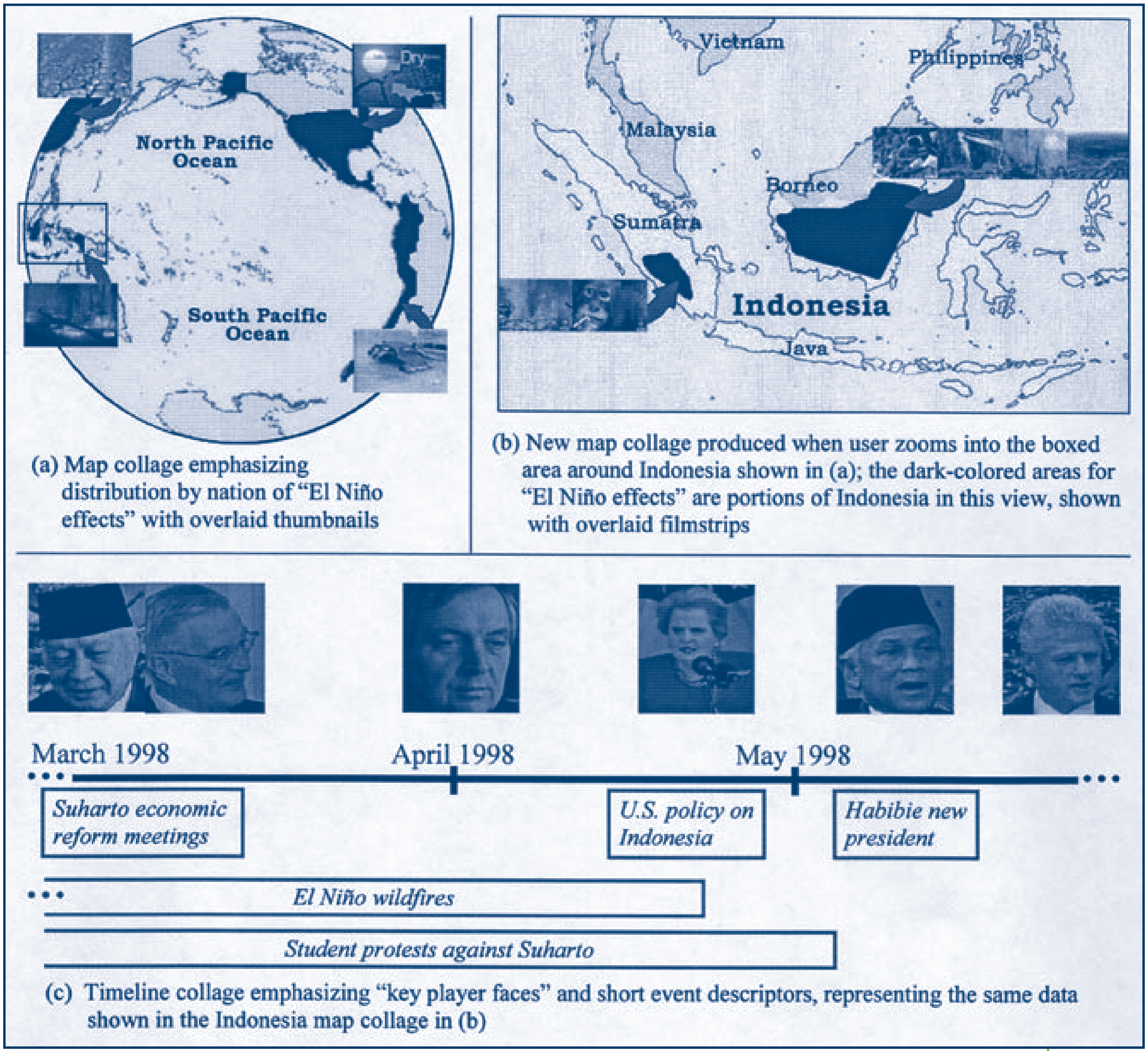
The Informedia Project at Carnegie Mellon University pioneered the use of speech recognition, image processing, and natural language understanding to automatically produce metadata for video libraries (Wactlar et al. 1999). The integration of these techniques provided for efficient navigation to points of interest within the video. For example, speech recognition and alignment allows the user to jump to points in the video where a specific term is mentioned, as illustrated in figure 4.

The benefit of automatic metadata generation is that it can perform a post-facto analysis for video archives that were produced in analog form and later digitized. Such archives will not have the benefit of a rich set of metadata captured from digital cameras and other sources during a digital production process. The speech, vision, and language processing are imperfect, so the drawback of automatic metadata generation, compared with hand-edited tagging of data, is the introduction of error in the descriptors. However, prior work has shown that even metadata with errors can be very useful for information retrieval, and that integration across modalities can mitigate errors produced during the metadata generation (Witbrock and Hauptmann 1997; Wactlar et al. 1999).

More complex analysis to extract named entities from transcripts and to use those entities to produce time and location metadata can lead to exploratory interfaces and allow users to directly manipulate visual filters and explore the archive dynamically, discovering patterns and identifying regions worth closer investigation. For example, using dynamic sliders on date and relevance following an “air crash” query shows that crashes in early 2000 occurred in the African region, with crash stories discussing Egypt occurring later in that year, as shown in figure 5.

The goal of the CMU Informedia-II Project is to automatically produce summaries derived from metadata across a number of relevant videos, i.e., an “autodocumentary” or “autocollage,” and thereby facilitate more efficient information access. This goal is illustrated in figure 6, where visual cues can be provided to allow navigation into “El Niño effects” and quick discovery that forest fires occurred in Indonesia and that such fires corresponded to a time of political upheaval. Such interfaces make

Figure 6. Prototype of Informedia-II collage summaries built from video metadata



(a) Map collage emphasizing distribution by nation of “El Niño effects” with overlaid thumbnails

(b) New map collage produced when user zooms into the boxed area around Indonesia shown in (a); the dark-colored areas for “El Niño effects” are portions of Indonesia in this view, shown with overlaid filmstrips

(c) Timeline collage emphasizing “key player faces” and short event descriptors, representing the same data shown in the Indonesia map collage in (b) use of metadata at various grain sizes. For example, descriptions of video stories can produce a story cluster of interest, with descriptions of shots within stories leading to identification of the best shots to represent a story cluster, and descriptions of individual images within shots leading to a selection of the best images to represent the cluster within collages such as those shown in figure 6.

Preserving Digital Data

Librarians and archivists have priorities that go beyond the agenda of content access, distribution, and payment systems for consumers and producers. Archivists and preservationists are vested with selecting a medium that will survive the longest and a system that will transcend the most generations of "player" hardware and software. Content that will be created digitally has both advantages and disadvantages over conventional analog film and video content. The National Film Preservation Board (NFPB) serves as a public advisory group to the Library of Congress (LC). Led by William J. Murphy, the LC produced a comprehensive report in 1997 that reviews the various facets of television and video preservation and surveys the various elements relevant to retention of all digitally produced content (LC 1997).

Media longevity problems exist both for analog and for digital content. Magnetic tapes will lose signal strength and stretch on stored reels. There are no standardized systems or methodologies for evaluating the physical or data-loss effects of tape aging. Digital video discs can delaminate, and many compact discs (CDs) with inadequate protective layers may be vulnerable to the effects of temperature, humidity variation, and pollution in less than five years. Such degradation can render digital data unreadable. On the positive side, digital media can be created with data redundancy, error-detection, and even error-correcting codes that detect and compensate for dropped bits. These techniques have long been used in digital communication and storage systems. Furthermore, digital content can be inexpensively recorded, or cloned, without generational loss, providing cheap and practical physical redundancy (there is no single master copy). Data that are kept online in disc-based systems can have data loss minimized by redundant array of inexpensive discs (RAID) storage systems. Such systems can also continuously or periodically refresh their data, thus sustaining their integrity.

Perhaps of greater concern is the rapid obsolescence of digital media formats and encoding schemes as advancing technology out-modes recording and playback devices in time frames much shorter than the media life. For example, two digital recording formats, D-1 and D-2, have been available to the industry since the late 1980s. Early generations of Sony's D-1 and D-2 equipment are already obsolete in production environments. The last few years have seen the introduction of numerous new video formats such as D-5 (for studio production), D-6 (for HDTV), DCT, Digital Betacam, DV, DVC, and Digital-S. Some new recording equipment also digitizes directly into digitally compressed formats, MPEG-1 (VHS quality) and MPEG-2 (studio-to-HDTV quality). The emerging standard for MPEG-7 will also allow for embedded metadata generated contemporaneously or following production. What is required is a format-independent cloning solution that will enable the digital content to be transparently interchanged, regardless of storage system, media type, encoding format, or transport mechanism, and without loss of data quality and fidelity.

DAM systems can separate the indexing and cataloging information that enable access from the underlying format of the medium. A database archive may be architecturally layered to render it medium-independent, thereby enabling access from one system to storage on another. This facilitates rapid and independent refreshing or conversion of the underlying data, data formats, and media. Modern systems should allow multiple types of archive storage media data banks to operate simultaneously through a common access interface. Thus, the lifetime of the metadata that index the content can far exceed that of the original media.

Conclusion

Content-based video retrieval is getting more attention as the volume of digital video grows dramatically. The Association for Computing Machinery (ACM) Multimedia Conference, started in 1994, has included a workshop dealing with multimedia information retrieval since 1999, and TREC started a new track on indexing and retrieval from digital video in 2001. TREC is an annual benchmarking exercise for information retrieval applications that has taken place at the National Institute for Standards and Technology for the last nine years (http://trec.nist.gov). TREC has been instrumental in fostering the development of effective information retrieval on large-scale corpus collections, and with the new digital video track signifies the emergence of digital video as an information resource.

These forums and others hosted by the Institute of Electrical and Electronics Engineers, Inc. (IEEE), the Audio Engineering Society, and other technical societies examine ways in which metadata can be generated for video through an automated analysis of the auditory and visual data streams. Evaluations are under way (for example, the TREC digital video track) to determine what metadata have value for identifying known items and exploring within a video archive. Metadata in the future should be more carefully tagged as to the confidence of the descriptor and producer to help the user direct the information search and exploration process. For an item known to be in the corpus, for example, the user might start by specifying that only metadata produced at the time the video was first recorded should be used. Another user exploring a topic may be willing to see all shots that might contain a face; an automated face detector returns a match in the shot but perhaps with low confidence. Through an appropriate interface, the user can quickly filter out those shots that truly contain faces from those that contain other images that only look like faces. Hence, along with an increased use of automatic metadata generators, these generators will also produce “metadata about the metadata,” including production credits and confidence metrics. MPEG-7 recognizes the value of metadata and provides intellectual property protection for the descriptors themselves as well as for the video content.

Digital video will remain an expensive medium, in terms of broadcast/download time and navigation/seeking time. Surrogates that can pinpoint the region of interest within a video will save the consumer time and make the archive more accessible and useful. Of even greater interest will be information-visualization schemes that collect metadata from numerous video clips and summarize those descriptors in a cohesive manner. The consumer can then view the summary, rather than play numerous clips with a high potential for redundant content and additional material not relevant to his or her specific information need. Metadata standards efforts discussed earlier can help with the implementation of such summaries across documents, allowing the semantics of the video metadata to be understood in support of comparing, contrasting, and organizing different video clips into one presentation.

Metadata will continue to document the rights of producers and access controls for consumers. Combined with electronic access, metadata enable remuneration for each viewing or performance down to the level of individual video segments or frames, rather than of distributions or broadcasts. Metadata can grow to include specific usage information; for example, which portions of the video are played, how often, and by what sorts of users in terms of age, sex, nationality, and other attributes. Of course, such usage data should respect a user's privacy and be controlled through optional inclusion and specific individual anonymity.

Metadata provide the window of access into a digital video archive. Without metadata, the archive could have the perfect storage strategy and would still be meaningless, because there would be no retrieval and hence no need to store the bits. With appropriate metadata, the archive becomes accessible. Furthermore, the window need not be fixed, i.e., the metadata should be capable of growing in richness through added descriptors for domain-specific needs of new user communities, unforeseen rights management strategies, or advances in automatic processing. By enhancing the metadata, the archive can remain fresh and current and accessible efficiently and effectively; there is no need to reformat or rehost the video contents to accommodate the metadata. Only the metadata are enhanced, which in turn enhances the value of the video archive.

References

Artesia Technologies. 2001. What Is Digital Asset Management (DAM)? Available at http://www.artesiatech.com/what_dam.html.

Bulldog. 2001. Welcome to Bulldog. Available at: http://www.bulldog.com/view.cfm.

Bormans, J., and K. Hill, eds. 2001. MPEG-21 *Overview.* ISO/IEC JTC1/SC29/WG11/N4318 (July). Available at: http://www.cselt.it/mpeg/standards/mpeg-21/mpeg-21.htm.

Ginger, K. Web page maintainer. 2000. DLESE Metadata Working Group Homepage. (November 6). Available at: http://www.dlese.org/Metadata/index.htm.

Green, D. 1997. Beyond Word and Image: Networking Moving Images: More Than Just the "Movies." *D-Lib Magazine* (July–Aug.). Available at: http://www.dlib.org/dlib/july97/07green.html.

Hillman, D. 2001. Using Dublin Core, DCMI Recommendation (April 12). Available at http://dublincore.org/documents/usageguide/.

Laven, P. 2000. Confused by Metadata? *EBU Technical Review,* No. 284 (September). Available at www.ebu.ch/trev_home.html.

Li, F., et al. 2000. Browsing Digital Video. *CHI Letters: Human Factors in Computing Systems,* CHI 2000 2(1) 169-176.

Library of Congress. 1997. *Television and Video Preservation: A Report of the Current State of American Television and Video Preservation.* vol. 1. Report of the Librarian of Congress (October). Edited by W. Murphy. Available at: http://lcweb.loc.gov/film/tvstudy.html.

Martinez, J. M., ed. 2001. Overview of the MPEG-7 Standard (Version 5.0). ISO/IEC JTC1/SC29/WG11 N4031 (March). Available at: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.

Negroponte, N. 1995. *Being Digital.* New York: Knopf.

National Institute of Standards and Technology Text Retrieval Conference. Video Retrieval Track. 2001. Available at: http://www-nlpir.nist.gov/projects/t01v/.

OCLC. 1998. Preservation Resources Digital Technology. Available at: http://www.oclc.org/oclc/presres/scanning.htm.

OCLC/RLG. 2001. Preservation Metadata Working Group Issues White Paper, *Preservation Metadata for Digital Objects: A Review of the State of the Art* (January 31). Available at: http://www.oclc.org/digitalpreservation/presmeta_wp.pdf.

Society of Motion Picture and Television Engineers. 2000. SMPTE Metadata Dictionary RP210a, Trial Publication Document, Version 1.0 (July). Available at: http://www.smpte-ra.org/mdd/Rp210a.pdf.

University of Southern California, Annenberg Center for Communication. Digital Asset Management Conferences I, II, and III, 1998–2000. Available at: http://dd.ec2.edu/.

Wactlar, H., et al. 1999. Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library. *IEEE Computer* 32(2): 66–73. See also: http://www.informedia.cs.cmu.edu/.

Wendler, R. 1999. LDI Update: Metadata in the Library. Library Notes, no. 1286 (July/August): 4–5.

Witbrock, M. J., and A. G. Hauptmann. 1997. Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents. In *Proceedings of the Association for Computing Machinery* DL '97. New York: Association for Computing Machinery.

Web sites noted:

World Wide Web Consortium. 2002. Available at http://www.w3.org

Informedia research at Carnegie Mellon University. 2002. Available at http://www.informedia.cs.cmu.edu

APPENDIX 3

Digital Preservation in the United States: Survey of Current Research, Practice, and Common Understandings

Digital Preservation in the United States: Survey of Current Research, Practice, and Common Understandings

DANIEL GREENSTEIN
Director, Digital Library Federation

ABBY SMITH
Director of Programs, Council on Library and Information Resources

Libraries and archives have long served to preserve significant portions of the published and unpublished record. They do this to ensure that the information in those records will be available to those who need it. Preservation has always been seen as a necessary condition for access. When information is recorded on paper and other analog media, the major challenges to preservation are posed by the fragility of the medium and by the costs of providing suitable storage, which are often quite high.

In the United States, preservation has traditionally been a distributed activity. Each library or archives is responsible for maintaining the accessibility of its own holdings, for its own users. Together, these individual collections constitute the national collection. The materials have traditionally been used on-site, although they may be loaned to other institutions through lending agreements that are designed, in part, to protect the artifact being lent. Sharing of resources occurs through reformatting (onto microforms, through preservation photocopying, and so forth). But in each case, the physical artifacts are assets that belong to the library or archives. The information contained in these artifacts may or may not belong to the institution; in fact, rarely are intellectual property rights given to the repository in which the materials are held. In the analog realm, fulfilling preservation responsibilities has entailed both meeting the information needs of (mostly on-site) users *and* protecting institutional assets. Preservation responsibilities are assumed upon the acquisition of a physical item and they continue through its life cycle.

These interests—preservation, physical possession or ownership, and access—are seldom as allied in the digital realm as they are in the world of analog media. The function of preservation for the purpose of providing physical or intellectual access does not fall automatically to an institution through the agency of physical ownership. The stakeholders in digital preservation often come from the same sectors as do stakeholders in the analog realm. They include creators, distributors or publishers, repositories or libraries and archives, and users. But these stakeholders may play very different roles in the digital realm than they do in the analog realm—roles that can put them in conflict with one another in areas where their interests once were parallel. Digital stakeholders can also create new alliances of interests.

One critical challenge to digital preservation in the near term is technical: the rapid rate at which hardware and software become obsolete means that information written in a specific code to run on specific hardware may be stranded by the adoption of newer, better code and hardware. This is the problem facing individuals who want to read an early version of a Lotus 1-2-3 spreadsheet that they have on a 5-1/4-inch disk they used to run on an IBM PC. The implication is that decisions about selection for preservation that can be deferred in the analog realm must be addressed early in the life cycle of digital files.

This paper summarizes activities under way in the United States that are designed to address the variety of preservation challenges—technical, legal, and social—and the changing roles and responsibilities of preservation stakeholders. It is divided into the following major sections:

- **Common understandings among stakeholders** describes the agreements that exist among those who take an interest in the long-term management of digital information.
- **Practical preservation activity** reports real archiving efforts and the circumstances under which they have emerged.
- **Experimental preservation activity** discusses significant practical experimentation in data archiving.
- **Preservation research** sets forth key areas for focused research and presents examples of projects in those areas.

Common Understandings Among Stakeholders

Limited but highly influential agreements about key issues exist among those who take an interest in the long-term management of digital information—interests that are intrinsically, if at times confusingly, interrelated. Those who create or publish such information, those who wish to use the information, and those who act as archival repositories for it all have a stake in maintaining digital assets over time. They often have different purposes in mind when they speak of making the information accessible in the future, but they share the conviction that such longevity is highly desirable.

The interests of the creator or distributor, user, and repository are interrelated because each group has a formative influence over whether, how, and at what cost digital information will be made accessible over the long term. The first decisive factor is how digital information is created and distributed. This may determine whether, how, and at what cost the information *can* be preserved and made accessible to users over time. The choice of some formats may make it more difficult to manage the digital object and ensure future, or even current, access. The selection of simple or standard formats (e.g., PDF files, TIFF images, or ASCII text) can simplify certain storage issues.

Another deciding influence is how, to whom, and under what terms or conditions archived digital information is to be distributed. This will determine how, by whom, and at what cost that information is created, distributed, and accessioned into an archive. Accordingly, preservation practice typically represents some continuing negotiation between creators or publishers, archives, and users. Each stakeholder makes choices that can influence the long-term accessibility of a digital asset. The Inter-university Consortium for Political and Social Research (ICPSR), for example, was designed to ensure long-term access to important social science research datasets. This membership organization states that "to ensure that data resources are available to future generations of scholars, ICPSR preserves data, migrating them to new storage media as changes in technology warrant" (ICPSR, no date). To support its activity, ICPSR has a sustainable, mission-driven business model, and it defines criteria for data entry, use, and preservation within the framework of that model. It has worked successfully for 40 years.

Stakeholders have reached a common understanding about what constitutes a trusted digital repository and what activities the repository must routinely undertake, even though the way in which some of the basic preservation functions will be undertaken remains uncertain. A viable digital archival repository must have a number of attributes. For example, it must be explicit about which digital information it preserves, why, and for whom. It also must be clear about the attributes of the archived information it intends to preserve. It must offer services that meet the minimum requirements of data creators and users. It must be prepared to negotiate and accept deposits of appropriate digital information from those who create or distribute that information, and the terms of those negotiations must be clear to all. The repository must also obtain sufficient control of deposited information so as to ensure its long-term preservation; this responsibility may include gaining access to data in order to check on their integrity while protecting those same data from access by unauthorized parties. The repository must make information available to users under conditions negotiated and agreed on with depositors.

Finally, given the rapidly changing technological environment in which the repository will take in and tend to digital information, it must actively seek new solutions as technology evolves.

Another area of common understanding is the emergence of the Open Archival Information System (OAIS) as the standard reference model. This model supplies a conceptual framework for discussing and describing archival practice. OAIS articulates the roles and interrelationships of the three groups that have a key stake in the digital process, i.e., creator or distributor, user, and repository. The reference model identifies preservation as a process that begins when digital information is created; this is a critical point of difference from the standard analog model, which considers preservation much later in the life cycle of an artifact. Finally, the OAIS model identifies the core functions and organizational features of a digital archival repository. This has influenced perceptions of what constitutes a trusted archives. OAIS is on the International Organization for Standardization (ISO) standards track and is the reference model of choice of those involved in digital preservation worldwide.

At present, there are four commonly understood technical approaches to digital preservation. These approaches are not mutually exclusive; indeed, there is an emerging consensus that all four approaches, and probably others not yet devised, will be deployed for the variety of digital object types and the demands for access to them.

Migration

In this approach, digital information is stored in software-independent formats. The information is reformatted as necessary so that it can be accessed using current hardware and software. Most digital archival repositories rely almost exclusively on data migration. It is doubtful that the strategy will work well with mixed media.

Technology Preservation

Under this approach, data are preserved along with the hardware and/or software on which they depend. Given the variety of hardware and software platforms and the rate at which they evolve, this strategy generally is not believed to be economically viable. Still, many data rescue efforts (see Digital Archaeology below) rely on the persistence of outmoded hardware and software.

Emulation

Often considered a form of technology preservation, emulation entails storing digital information alongside detailed information about how it looked, felt, and functioned in its original software/hardware environment. The look, feel, and functionality of the digital information are then “emulated” or re-created on successive generations of hardware/software. Emulation is particularly pertinent to mixed media. Individuals who are conducting research on the technical and economic viability of this approach include Jeff Rothenberg at the RAND Corporation and researchers at CAMiLEON. Emulation is in the exploratory phase; it has never been successfully used for preservation in a sustainable way.

Persistent Object Preservation

The opposite of migration, persistent object preservation (POP) entails explicitly declaring the properties (e.g., content, structure, context, presentation) of the original digital information that ensure its persistence. Of the strategies listed here, POP is the only one that starts with and remains focused on preserving the digital information from its inception. Other strategies attempt to counteract or overcome the generic technical problem of obsolescence.

Digital Archaeology (Data Mining)

Although not a preservation strategy as such, digital archaeology is worth mentioning. It enables digital information to be rescued or recovered from disks, tapes, and other storage media that are no longer readable as a result of physical deterioration, neglect, obsolescence, or similar reasons.

To remain viable over the long term, appropriate documentation or metadata must accompany digital information. Key preservation metadata initiatives are reviewed in a white paper by the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG).[1]

Practical Preservation Activity

There are several practical preservation efforts under way that demonstrate the range of experience and expertise around the country.

Active preservation programs are under way in archives where preservation is often legally mandated. For example, the archives of national and state governments are legally bound to preserve selected records of government, including electronic records, in perpetuity. Business archives, such as those at financial, pharmaceutical, chemical, and other companies, may maintain records for legal and other reasons. Statutes of limitations often govern these mandates; consequently, such archives do not typically keep data in perpetuity as do government archives. These systems can be said to be more analogous to records management than to archiving; nonetheless, managing digital records even for seven years can provide technical challenges. Archives are also established at not-for-profit institutions, such as universities, that maintain records (including electronic records) for legal, business, and cultural reasons.[2]

Preservation is also under way in organizations in which data creators and producers perceive the long-term commercial value of digital information. Publishers such as Elsevier Science preserve the electronic scholarly journals they produce. The entertainment industry, most notably music and film companies, have large investments in digital assets that they wish to reuse over time, and they have developed digital asset management systems tailored for their specific needs.

Preservation programs also are active in organizations that perceive a noncommercial value of digital information for use and reuse. Libraries, archives, and museums that digitize objects in their collections for online presentation, for example, may seek to maintain those objects over time rather than to rescan them as they become obsolete.

In places where data archives and systems vendors see commercial possibilities in the provision, supply, and support of long-term data storage facilities, preservation has become vital to commercial development. Data warehousing is a cottage industry with numerous related trade associations, exhibitions, and certification procedures. Data archives are beginning to emerge in the library community; for example, both OCLC and RLG are considering offering data archiving facilities on a cost-recovery basis.

Specific research communities, where data creators are also data users and where both groups recognize the importance of being able to reuse research data, undertake large-scale preservation of their intellectual assets. Both the ICPSR and the Roper Center preserve social science and government statistical data.

There are also major preservation activities in communities where data creators and data users recognize their interdependence and the value of the digital information in which they maintain a common interest. Through PubMed Central, the National Library of Medicine acts as a digital archival repository for medical publications and other medical information.

Finally, archival repositories may be developed as a by-product of a commercial process. The Internet Archive is an archive of "snapshots" taken of selected Web pages by Alexa. An information company can use information gained from those snapshots for commercial purposes. Alexa assesses the visibility of Web pages by seeing who links into a site.

Experimental Preservation Activity

The InterPARES (International Research on Permanent Authentic Records in Electronic Systems) Project is a major international research initiative involving archival scholars, computer engineering scholars, and representatives of national archival institutions and private industry. Its goal is "to develop the theoretical and methodological knowledge essential for the permanent preservation of records generated electronically, and, on the basis of this knowledge, to formulate model policies, strategies, and standards capable of ensuring their preservation." The InterPARES Project is investigating numerous issues in digital preservation, including the authenticity of electronic records.

The National Archives and Records Administration is developing a strategic and technical framework within which it may preserve in perpetuity selected electronic records of the federal government. It is closely involved with the InterPARES Project, the OAIS reference standard, the National Partnership for Advanced Computational Infrastructure led by the San Diego Supercomputer Center, and others. It is an international leader in research in selected areas, including requirements and processes for the preservation and reproduction of authentic records, development of the persistent archives method, application of advanced computing tools to records-management processes, and integration of digital preservation technologies with infrastructure technologies for e-government and e-business.

Under the auspices of the Andrew W. Mellon Foundation's e-journal archiving program, seven major libraries (the New York Public Library and the university libraries of Cornell, Harvard, Massachusetts Institute of Technology [MIT], Pennsylvania, Stanford, and Yale) are engaged in planning digital archival repositories for different kinds of scholarly journals. Yale, Harvard, and Pennsylvania have worked with commercial publishers on archiving the full range of their electronic journals; Cornell and the New York Public Library have worked on archiving journals in specific disciplines. MIT's project involves archiving "dynamic" e-journals (i.e., those that change frequently), and Stanford is investigating the development of archiving software tools under the auspices of its LOCKSS (Lots of Copies Keep Stuff Safe) program.

RLG and OCLC are jointly conducting preservation research. At present, their work focuses on the attributes of a digital archival repository and on preservation metadata.

The Andrew W. Mellon Foundation has invested in an investigation of emulation as a viable preservation strategy. Jeff Rothenberg at the RAND Corporation is conducting this research.

The IBM Almaden Research Center is investigating the possibility of using a universal virtual machine for digital preservation

The University of Pennsylvania is conducting work on data provenance.

Preservation Research

There are currently nine areas of significant research into preserving digital files. They are:

1. **Architecture and performance of archival repositories.** Key research is under way at the San Diego Supercomputer Center, Stanford University, the National Archives and Records Administration, the Library of Congress National Audio-Visual Conservation Center in Culpeper, Va., Cornell University, Yale University, MIT, and Harvard University.
2. **Persistent identification of and naming for archived information** (e.g., International Digital Object Identifier [DOI], Persistent Uniform Resource Locator [PURL]).
3. **Methods for recording and ensuring authenticity of archived information** (digital signatures, watermarking, mechanisms for recording information about provenance). Determining the authenticity of a digital object is likely to require the use of techniques whose reliability is still being debated. Techniques appropriate to digital images may include digital signatures and watermarking. Checksums and other technical routines that produce message digests are appropriate for objects in virtually all formats. They help determine authenticity by analyzing the object's structure and composition and whether it has been changed in any way since a particular benchmark point.

   Information may be found at:

   - *Authenticity in a Digital Environment* (CLIR 2000). Report of a group of experts convened by CLIR to address the question: What is an authentic digital object? *http://www.clir.org/pubs/reports/pub92/contents.html*

- The importance of verifying the authenticity of an information object is well described in *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections* (CLIR 2001) *http://www.clir.org/activities/details/artifact-docs.html*
- MD5 unofficial home page *http://userpages.umbc.edu/~mabzug1/cs/md5/md5.html*
- On checksum, see *http://www.checksum.org/*
- On digital signatures, see *http://www.w3.org/DSig/* and information from the Electronic Privacy Information Center
- On digital watermarking, see The Information Hiding Homepage. Steganography and Digital Watermarking. Available at: *http://www.cl.cam.ac.uk/~fapp2/steganography/*

4. **Degradation and testing of magnetic and other media used to store digital information** (work being conducted at the National Institute of Standards and Technology).

5. **Attributes of preservable digital information.** These efforts focus on specific kinds of digital information. For example, research communities interested in social science and in space data have defined standards for formatting and describing information in their respective fields.

6. **Attributes of trusted digital archival repositories.** This work centers on specific kinds of digital information and on the organizations that arise to preserve it. Participants in the Mellon e-journals archiving program, for example, are looking at the organizational, business, and rights issues that surround archives that are established to preserve scholarly e-journals.

7. **Development of standards** (including standards for data and metadata formats, digital storage media, and data management practice). Formal standardization takes place through bodies such as the ISO, the World Wide Web Consortium (W3C), the National Information Standards Organization (NISO), and the Internet Engineering Task Force (IETF) and reflects the emerging consensus of stakeholder communities. It is important to distinguish between the standards themselves and the understandings that need to be reached among stakeholders about how the standards are to be applied in certain instances (see item 5).

8. **Automatic copying and distribution of digital information** (LOCKSS).

9. **Policies and implementation mechanisms for the preservation, risk management, and assessment of Web-accessible content** (Project Prism at Cornell University).

If preservation activity in the near future bears any resemblance to that activity in the past 18 months or so, there will be further significant and unpredictable changes in this dynamic field.

References

ICPSR. No date. "About ICPSR." Available from *http://www.icpsr.umich.edu/ORG/about.html.*

Web Sites Noted in Text

Alexa. *http://info.alexa.com*

CAMiLEON. *www.si.umich.edu/CAMILEON/index.htm*

Cornell University. *www.library.cornell.edu/preservation/digital.html*
http://rmc-www.library.cornell.edu/online/studentrecords/

Electronic Privacy Information Center. *www.epic.org/*

Elsevier e-journal archiving. *www.elsevier.nl*
www.elsevier.nl/homepage/about/resproj/tulip.shtml
www.diglib.org/preserve/yale0206.htm

Harvard University. *www.news.harvard.edu/gazette/1999/03.25/diglibrary.html*

IBM Almaden Project. *www.almaden.ibm.com*

Internet Engineering Task Force. *www.ietf.org*

International Digital Object Identifier (DOI). *www.doi.org*

International Organization for Standardization (ISO). *www.iso.org*

Internet Archive. *www.archive.org/about*

InterPARES Project. *www.interpares.org*

Inter-university Consortium for Political and Social Research (ICPSR). *www.icpsr.umich.edu*

Library of Congress National Audio-Visual Conservation Center in Culpeper. *http://lcweb.loc.gov/rr/mopic/avprot/avprhome.html*

Lots of Copies Keep Stuff Safe (LOCKSS). *http://lockss.stanford.edu*

Massachusetts Institute of Technology (MIT). *http://web.mit.edu/newsoffice/nr/2000/libraries.html*

Mellon e-journal archiving. *www.diglib.org/preserve/ejp.htm*

National Partnership for Advanced Computational Infrastructure (NPACI). *www.npaci.edu/online/v6.2/perm.html*

National Archives and Records Administration. *www.nara.gov*
www.nara.gov/nara/vision/eap/eapspec.html
www.nara.gov/nara/electronic

National Information Standards Organization (NISO). *www.niso.org*

National Institute of Standards and Technology (NIST). *www.nist.gov; www.itl.nist.gov/div895*

Online Computer Library Center (OCLC). *www.oclc.org/research/pmwg/*

Open Archival Information System (OAIS) standard “reference model.” *http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html*

Persistent Uniform Resource Locator (PURL). *www.purl.org*

Project Prism. *http://prism.cornell.edu/PrismWeb/AboutPrism.htm*

PubMed Central. *www.pubmedcentral.nih.gov*

Research Libraries Group (RLG). *www.rlg.org/longterm/index.html; www.rlg.org/pr/pr2000-oclc.html*

Roper Center. *www.ropercenter.uconn.edu/catalog40/StartQuery.html*

Rothenberg, Jeff (RAND Corporation). *www.rand.org/methodology/isg/archives.html*

San Diego Supercomputer Center. *www.sdsc.edu/DigitalLibraries.html*

Stanford University. *www.sul.stanford.edu/depts/spc/findaids.html*

University of Pennsylvania (work on data provenance). *http://db.cis.upenn.edu/Research/provenance.html*

World Wide Web Consortium (W3C). *www.w3.org*

Yale University. *www.yale.edu/opa/newsr/01-02-23-02.all.html*

Footnotes

[1] See *http://www.rlg.org/longterm/index.html.*

[2] The National Archives and Records Administration’s Center for Electronic Records is perhaps the largest government archive for electronic records *(http://www.nara.gov/nara/electronic/).*

APPENDIX 4

Council on Library and Information Resources Survey on Digital Archiving

Council on Library and Information Resources Survey on Digital Archiving

DALE FLECKER
Associate Director of the University Library for Planning and Systems, Harvard University

At the request of the Library of Congress, the Council on Library and Information Resources distributed a survey to the 24 nonfederal government research libraries of the Digital Library Federation concerning their plans for digital archiving in February 2002. This paper summarizes the 14 answers received.

The three questions asked in the survey were:

1. What types of born-digital information resources do you expect your library to take preservation responsibility for?
2. Would you be interested in working in partnership with the Library of Congress by including the materials you are intending to preserve in the national plan?
3. What are your institutional priorities for preserving born-digital material?

Priorities and Assumption of Responsibility

A number of the respondents conflated answers to the first and third questions, while others gave specific instances in answers to question 1 for the general priorities they listed in question 3. These two questions will be treated together here. While there were a large number of different specific instances listed in response to the first question, and varying language used in defining priorities, most of the responses can be summarized in the following four categories (given in overall priority order):

Materials Created Within the Institution

Thirteen of the 14 responses listed these as their first priority in archiving. These libraries expect to take responsibility for the digital intellectual output of their institutions. Specific instances of resources in this category mentioned include:

- Institutional records (mentioned explicitly by seven institutions); digital materials received as part of heterogeneous archival collections (mentioned by five institutions).
- Locally hosted e-journals (mentioned five times).
- Materials created or collected by local faculty (working papers, databases, converted textual documents, etc.; listed by eight institutions).
- Miscellaneous other local materials mentioned include dissertations, local Web sites, student portfolios, and learning or classroom objects.

External Resources under a Coordinated National Program

Five institutions explicitly expressed a willingness to assume responsibility for the preservation of categories of external research resources as their share of a coordinated national digital archiving program.

Topically Relevant Collections

Four institutions said they expect to continue the sort of topical collection building in the digital era that has characterized research libraries in the past. Specific instances of such collections listed include political and social documentation, avant-garde literature and art, Latin American resources, and foreign legal materials. One assumes that these would generally represent collections including both traditional and digital materials.

Locally Hosted Materials

Several types of materials are frequently hosted locally on campuses or in libraries rather than being accessed remotely over the Internet. Instances of such materials mentioned include numeric and survey datasets, visual materials, audio materials, and geographic information. At least three institutions indicated that once such materials were brought on campus for use, they would expect to assume preservation responsibility.

Partnership with the Library of Congress

Ten institutions expressed willingness (and frequently eagerness) to participate in the Library of Congress national archiving program. Sometimes implicit and sometimes explicit in the comments was an expectation that this cooperation involved a formal division of labor in the coverage of materials archived. Other areas of partnership mentioned included work on archiving models and on technology. Concern about the inclusion of other key players (OCLC and RLG were explicitly mentioned) was expressed, as was a concern that the partnership be one of equals, rather than being dictated by the Library of Congress (one can easily think of other instances, particularly in the bibliographic realm, where the Library acts as the host and convener of a cooperative, but where all participants have a role in policy and in setting direction).

Observations

Several other aspects of these responses merit note:

Early Thinking

We are early in the development of digital preservation programs, and it is easy to discern in the responses the tentative nature of current plans in this area. Few if any of the respondents had existing formal digital preservation policies in place on which to base their responses. In many cases this survey was probably the first time priorities were publicly enunciated. One suspects that further discussion and the opportunity to review plans at other institutions would result in many changes and refinements in the responses.

Reborn Digital

While the survey was very specifically only concerned about born-digital materials, a number of responses (and a number of respondents in subsequent discussions) could not help but mention materials digitized from existing collections. One suspects many respondents had a hard time not emphasizing that they will also give high priority to the preservation of a large amount of digital materials converted from existing collections. While not specifically relevant to the Library of Congress program, it does imply the availability of local digital preservation infrastructure and expertise as discussed below.

Infrastructure

The intent to accept responsibility for a wide range of digital materials seen in the replies means that these institutions either have created or intend to create significant infrastructures for archiving and preserving digital materials. Among the infrastructure components one might expect to be required to accept these responsibilities are a robust digital repository, collection ingestion and quality control systems, access management and persistent identifier facilities, expertise in specific digital formats, and a technology monitoring and format migration function.

Range of formats

One striking thing about the specific examples of materials these libraries intend to preserve is the wide range of technical formats represented: audio files, humanities texts, geospatial resources, survey datasets, Web sites, e-journals, video, pictorial materials, etc. Preserving digital materials is primarily a format-by-format question and requires both expertise in and appropriate tools for maintaining the vitality of each specific format. Is it realistic to think that each institution will have the necessary infrastructure to preserve all of these disparate materials? The need to support a wide and diverse range of technical formats suggests one natural area for national partnership: the creation of centers of format expertise that could be drawn upon by all partners in carrying out their local responsibilities.

Remote Resources

These libraries seem to readily accept that they should take responsibility for the preservation of materials stored in their individual institutions. This is a natural extension of the preservation patterns for traditional collections. The proliferation of remote resources distributed across the Internet, however, poses a particular challenge when thinking about digital preservation. Such materials would seem to be the general responsibility of everyone, but not the natural responsibility of anyone in particular. Yet these resources are some of the most important of today's digital materials and in general now represent a majority of what is used day-to-day on university campuses. Addressing these "common good" remote resources is one of the key areas for a coordinated national plan.

Relevant Activity

The number of activities relevant to the Library of Congress plan mentioned by respondents is striking:

- Four institutions (Cornell, Harvard, Stanford, and Yale) are involved in the e-journal archiving program of the Andrew W. Mellon Foundation.
- Repository systems oriented toward digital preservation are available or under development at the California Digital Library, Cornell, Harvard, and Stanford, among other institutions.
- Stanford is pursuing the construction of a "Dark Cave" service to provide long-term archiving for external depositors.
- Cornell is engaged in the Prism project, which is investigating ways of evaluating preservation risks for Web-based resources.

Overall, the survey seems to demonstrate that there is a set of research libraries that would be natural partners to the Library of Congress in the creation of a national cooperative plan for digital preservation. These institutions have already identified sets of digital resources for which they expect to take responsibility, are creating the infrastructures to support digital preservation activities, and have expressed a willingness to work with the Library in this domain.

APPENDIX 5

National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and Related International Activity

National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and Related International Activity

NEIL BEAGRIE
Beagrie Consulting

Executive Summary

Aims, Scope, and Methodology

This report aims to provide an overview of selected key national and multinational initiatives in digital preservation occurring outside North America. The report has examined current digital preservation initiatives in four countries: Australia, France, the Netherlands and the United Kingdom, as well as related multinational initiatives. The programs in these four countries and the multinational initiatives were chosen in consultation with the Library of Congress (LC) and the Council on Library and Information Resources (CLIR) as being of particular relevance and interest to the National Digital Information Infrastructure and Preservation Program (NDIIPP).

This study aims to put these initiatives into their national and international context and to outline the major developments that are in progress. It is intended to provide a high-level survey. As such, it does not aim to be exhaustive or detailed in terms of practice and procedures. This report presents the key findings from the survey and details of the main initiatives in each country.

The survey has been undertaken primarily from desktop research and information supplied by the national libraries in each country between January and March 2002. A detailed questionnaire was developed in consultation with the Council on Library and Information Resources and the Library of Congress. This covered specific questions on national libraries' initiatives. The questionnaire also allowed the respondents to provide information that discussed their respective national contexts and to identify related initiatives

©Neal Beagrie 2002. The author has asserted his moral rights in this work.

Staff in the national libraries were interviewed as part of the consultancy. Supplementary visits to learn more about these specific initiatives were made to the British Broadcasting Corporation, which is the lead partner in the PRESTO (Preservation Technology for European Broadcast) project, and to the Digital Longevity digital preservation test bed in the Netherlands. Prior to the interviews, desktop research was completed on Web sites of relevant organizations and their staff publications; pertinent information was entered into the draft questionnaire. On completion of the interviews, the completed draft questionnaire and draft report section for that country were sent to the interviewees for any comments, additions or corrections.

Key Observations and Recommendations

This section of the executive summary provides the author's observations on principal trends and lessons. This is followed by individual observations and recommendations from each national library on the lessons from their work for the NDIIPP. Their views on opportunities for future international collaboration in digital preservation are presented separately in each national overview in the main body of the report.

Author's Observations and Recommendations

Introductory Remarks It should be noted there are substantial differences in the scale and scope of collections among the national libraries surveyed. Although all the libraries have responsibilities for the print and literary heritage of their respective countries, their responsibilities for audiovisual materials vary substantially. Each country may also have adopted a slightly different focus in terms of digital publications or will be at different stages of development in terms of progress with developing digital collections and digital preservation.

These differences in the scope and scale of collections and individual national circumstances need to be borne in mind when considering the implications and lessons of this survey for the Library of Congress or the NDIIPP.

This report and the national surveys it contains are a snapshot of the current position as of March 2002. As such, it should be noted that further changes and initiatives will need to be taken into account as time progresses from completion of the interviews and surveys.

Despite these caveats, the author believes that there are significant lessons and opportunities for both the Library of Congress and the NDIIPP highlighted within this report.

National Initiatives and Funding A starting point must be that there are no single national initiatives for digital preservation in the countries surveyed. In practice, there are many institutional missions that are being extended into the digital domain, including those of national institutions such as the national archives and national libraries.

There are an emerging number of efforts to provide national or international coordination and collaboration between such initiatives. One national example is the Digital Preservation Coalition in the United Kingdom. International examples focusing on exchange of information are ERPANET and PADI.

Internationally, digital preservation is poorly funded in relation to the scale of the challenges faced. There has been limited or no additional core funding made available to institutions to address digital preservation. As a result, institutions have relied on short-term external project funding or made difficult and sometimes painful reallocation of internal resources. However, there are clearly limits to what can be achieved by such means, particularly in larger institutions or in national programs.

In providing a funded and coordinated national program for digital preservation, the U.S. National Digital Information Infrastructure and Preservation Program (NDIIPP) is seen internationally as a world-leading initiative.

It was noted that it remains far easier to obtain funding for digitization for access than for digital preservation itself. The long-term benefits and requirements of preservation seem often to be overshadowed by the immediate benefits of current access initiatives. There is increasing emphasis on short-term "challenge funding" in many countries and reluctance to increase the core funding of institutions. Increases in core funding will be necessary to make the longer-term commitments needed for preservation of large digital collections.

Digital preservation relies substantially on the collaboration of key stakeholders outside the memory institutions and the professional sectors they represent. An important part of digital preservation activity as a public good is funded either from public funds by government or through private benefactors. However, awareness of digital preservation issues among the public, government and other key stakeholders remains low. I would recommend significant effort be placed in targeted outreach to key individuals and audiences as part of the development of the NDIIPP to ensure it has effective support and engagement with key communities. The public relations campaign and launch by the Digital Preservation Coalition in the United Kingdom is seen by its member organizations to have been highly successful and may provide useful parallels for part of any outreach program in the United States.

Underlying Trends The digital domain is changing the nature of institutional missions and existing relationships with other organizations. These changes can be summarized as:

- *Changing patterns of distribution.* Increasingly institutions do not hold physical copies of digital works but licence access to them. The responsibility for archiving or the level of trust in archiving arrangements is currently uncertain.
- *Changing timescales for preservation.* Digital media are fragile and access to them dependent on rapidly evolving and quickly obsolete hardware and software. Preservation of digital materials will, therefore, not happen by accident and requires early action, often at the point of creation, to be successful. In the digital

environment, there is a need to have a much closer relationship with creators and distributors of digital materials. It is necessary to take preservation actions earlier than may be the case with traditional materials. Selection decisions can be harder, as they may have to be taken earlier in the life of the material and without the benefit of several decades having passed and the historical importance of different trends and material being clear.

- *Changes in IPR and archiving rights.* No country in the survey currently has comprehensive legal provisions for archiving digital publications. The term of copyright has been increasing and the investment in and economic value of IPR have also increased dramatically. The commercial need to protect IPR can overshadow other considerations. The needs of memory institutions for legal exceptions to undertake archiving are often overlooked or not widely understood.
- *Globalization.* Activities increasingly take place on a global scale and outside of the traditional national frameworks for digital preservation. With the development of international publishers who can deliver their digital publications from anywhere in the world, the role of archiving in a national context is less clear. Similarly, the growth of the Web and the international activity it empowers transcend national boundaries.
- *Globalization also applies to developments in hardware and software.* The fact that information technology companies and market trends operate on a global scale and apply to many different sectors means that there is more substantial common ground among institutions internationally and across sectors. There are, therefore, greater potential and benefits from international collaboration in this field.
- *The information explosion.* The volume and range of information produced is expanding dramatically. At present, digital publications in many countries are a supplement to, rather than a replacement for, traditional publication. This increase in both traditional and digital information is placing a strain on the national institutions, particularly the national libraries that have a tradition of comprehensive collection in specified areas. At the same time, many of the traditional filtering and editing roles of publishers are disappearing as the Web opens up publishing to individuals and organizations. This places greater demands on the libraries in terms of selection of material.

  This exponential increase in information is not solely confined to publishing, but applies to an even greater degree to data in the academic and research sectors, particularly in the sciences.
- *Publications and records.* It is no longer necessary in the digital environment to generate many copies to publish material. A single copy can be networked and accessible to anyone worldwide with a PC and an Internet connection. The boundaries between what is a “publication,” a “manuscript” or an “archival record” have become blurred. The respective roles of libraries and archives may have a greater degree of overlap in the digital environment.

- *The cultural record.* Publications are now only one aspect of popular culture and the cultural record. An increasing part of our culture is defined by film, television, and the World Wide Web. Mechanisms to consider new areas of collection development and future research needs may be required as part of any national scheme.
- *The role of the private collector.* The role of private individuals has frequently been vital in preserving collections of material, particularly ephemera that have not been in areas of contemporary collection by curators. It is possible to point to contemporary examples of key initiatives started by private individuals (perhaps the sharing of early computer games and emulators by private enthusiasts). However it seems likely that digital preservation challenges and copyright protection mechanisms will make such efforts harder in future decades. This could entail greater reliance on the selection decisions made by institutions and/or developing new tools to support personal archiving by individuals.

DIGITAL PRESERVATION Institutions such as archives and libraries have evolved over many centuries as custodians of the "collective memory." They are custodians over very long periods of time. Other institutions and sectors may be focused on much shorter time horizons and rarely have this chronological perspective. It is not surprising, therefore, that memory institutions have been first to identify the challenges associated with digital preservation.

However, the challenges identified by these institutions will in time affect a wider range of institutions and may have a profound effect on the individuals and wider society in which they operate. Digital preservation is therefore not solely a cultural heritage issue. In the longer term, it will affect the nature of the "information society" that many governments worldwide are seeking to develop. There is a surprising lack of discussion or research into these deeper trends and the implications behind digital preservation issues.

Digital preservation is still a relatively new field. Most initiatives have focused on selection and acquisition and storage and maintenance of digital collections. Actions needed for long-term preservation are only now being identified and addressed.

Because digital preservation is a new field, it is important to make a start and identify discrete areas of work to move forward within each institution. The most successful initiatives noted in the survey had been institutions that have been working on practical implementations and policy over a number of years.

COLLABORATION AND PARTNERSHIP Collaboration between institutions occurs on many different levels.

The existence of external funding has encouraged collaboration on research. In some cases, collaboration with other institutions has been a requirement of such research funding. Research collaboration has also occurred without this external incentive although it is often on a more informal basis or more constrained level of resources.

Collaboration and coordination of collection policies have been harder to put into effect. The PANDORA archive in Australia is the only real example of this in the survey, and this initiative has evolved over many years. Coordination and distribution of responsibility between institutions are also seen as important requirements in the United Kingdom, but there is still some way to go to put appropriate arrangements into effect.

Partnerships seem to work best when the institutions have their own initiatives and experience, and both parties have something to offer and to gain. It is important to develop in-house expertise as well as utilize experience available externally.

Working with key stakeholders. This is essential and was emphasized by all the libraries in the survey. There are many examples of successful approaches included in a report. Agreement between publishers or publishers' trade bodies and national libraries are noted in the survey. Very successful outreach publications targeted at data creators have also been produced by the Arts and Humanities Data Service in the United Kingdom (the *Guides to Good Practice* series) and the National Library of Australia (information leaflets on *Safeguarding Australia's Web Resources*).

Digital libraries are a relatively small sector, and there are clear benefits both to libraries themselves in working together but also in being aware of trends and potential partnerships in other sectors.

A clear example of this is the Open Archival Information System (OAIS) reference model, which is emerging as the first international standard in digital preservation. This model was first developed within the communities engaged in earth observation for their requirements. However, it has much wider applicability and has also been widely adopted in the library community. The library community, in turn, has heavily influenced the development of the draft reference model.

It seems likely that the digital preservation and related issues such as mass storage and automation of metadata will be an important element of the "research grids" being developed to support collaborative science and the scientific research infrastructure. I would recommend opportunities for synergies with these developments be explored and encouraged wherever possible in the NDIIPP.

Governments worldwide are encouraging developments in "e-government" and "information society" that are having a major impact on the provision of digital access and development of digital work processes and procedures. Electronic records management often features in such programs, but it is very rare for longer-term issues to be considered. The Digital Longevity program in the Netherlands is a rare example of digital preservation being included in such programs. I would recommend a close engagement with and awareness-raising targeted at such initiatives in the United States.

Many of the current certainties of publication and archiving are in flux as we move into the digital environment. It seems likely that fewer institutions will, in practice, be involved in digital preservation directly. Rather, many institutions are likely to be involved in providing access services that may rely on such archiving activities for long-term access. Funding and institutional models for this set of relationships remain to be defined. However, there are a number of interesting developments within the survey. The future development/recognition of some archives as official archives by international publishers for all their published output is one such development (see the Koninklijke Bibliotheek and Elsevier in the Netherlands). Another is the potential development of collaborative archiving arrangements for consortia linked to national deposit libraries or academic research libraries (see COUPERIN in France or JISC in the United Kingdom).

In terms of future international collaboration, there are several suggestions that recur and are prominent in responses from national libraries.

A requirement for effective long-term preservation identified in the survey is the need to develop a preservation technology watch for file formats and new technologies, emulators and migration routines, and information on and repositories for obsolete software. National libraries felt there is significant scope for international collaboration and potential cost benefits in developing these services on a shared basis.

It also seems likely within larger national programs that there may be scope to develop some shared services and central support for digital preservation in a distributed network of digital archives.

In the academic sector, the Open Archives Initiative and exploration of new methods of scholarly communication are growing rapidly. The focus of these initiatives is on improving current access, and there is at present less consideration of long-term requirements for preservation. The position of such repositories, the materials they hold and any long-term preservation requirements should be considered further in any national collaborative scheme or partnerships.

There is also a need to foster further research on long-term preservation and develop standards and good practice. This would be an obvious area for international effort and for developing closer partnerships with national research funding bodies and academic research institutes and departments.

Staff Training and Development Staff training and development issues were raised by most institutions in the survey. Digital preservation may require some new posts in terms of individuals with a crossover set of skills and an overview to coordinate and direct activities. However, the majority of effort will be drawn from existing staff and requires the encouragement of teamwork across different departments and skill sets within the institutions.

Audiovisual Materials The audiovisual preservation community is in many senses unique within the survey. As the media and technologies used in their industries have been impermanent and cannot be preserved long-term, digitization is widely accepted as their preferred method of preservation. The film, audio, and video archiving communities, therefore, have a direct stake in resolving digital preservation challenges over the next decade.

Although not a primary focus of this study, audiovisual materials have been included to some degree. Three observations stand out:

- The audiovisual community has undertaken more research and evaluation of the archival qualities of storage media such as CD-R than any other. This work is not widely known in the library community, is highly relevant to a wider audience and deserves to be better known.
- Their storage requirements are very large. The desired process of moving from offline storage, such as CDs, to mass storage systems will require very large-scale storage systems.
- The PRESTO project is one of the few examples encountered of an attempt to both identify the scale of preservation requirement for a group of institutions and construct an effective business case for further investment. I would recommend that the survey questionnaire, the survey outcomes and the technologies being developed be closely examined by the Library of Congress.

RESEARCH AND DEVELOPMENT Web archiving, either in terms of selective gathering of specific Web sites or whole domain capture for specific national territories, has been highlighted as a key area by the national libraries. I would recommend that work within the United Kingdom, France, Australia, and Scandinavia be followed closely and that opportunities for a joint development of tools and practice are explored.

The Open Archival Information System (OAIS) reference model has become widely accepted as a key standard in digital preservation. I would recommend it be utilized within the NDIIPP and support be given to further key developments. These include production of accessible guides to key concepts behind the standard, dissemination and sharing of experience between implementers of the standard, and efforts to develop supporting guidelines and standards in areas such as identifiers, ingest, and certification.

A key issue highlighted in the survey is the need for persistent identifiers. This is an issue not only for online publications and the Web but also for linking and citation of primary research and datasets. I would recommend further work on persistent identifiers as part of the NDIIPP and close liaison with international developments in this area.

A number of “ingest” (acquisition and processing of digital objects into collections) activities need further development. Within the space science community, efforts are focusing on space mission data. Within the library community, there is a need to focus on metadata from publishers. A number of promising preservation metadata schema have emerged and an international framework is being developed by a Research Libraries Group/OCLC Online Computer Library Center Preservation Metadata Working Group. The logical next step is to examine implementation issues. Many of the publishers are international and would respond positively to international standards and coordination of requirements. Links to metadata standards being developed within the publishing community such as ONIX would also be desirable. The needs of a range of publishers from small to medium scale to the largest operations will also need to be considered. In this respect, current work to develop SIMONE, a tool for automating production of ONIX metadata by publishers, being funded by the British Library at Book Industry Communication (BIC), may be of wider interest.

With the development of new roles and potentially new interdependencies among different organizations, certification of digital archives will be of increasing importance. I would recommend further consideration and support be given to efforts to define appropriate benchmarks or institutional standards for digital preservation and certification models.

There has been relatively little major research in digital preservation, and there is a need to invest in this activity in terms of further research and development.

I would recommend that the outcomes of the research projects noted within the survey be considered carefully within the NDIIPP. Consideration should be given to building on these projects and/or developing a digital preservation test bed(s) to evaluate the scalability, strengths and limitations, and costs of promising approaches.

Dissemination There are pronounced differences among institutions in terms of the levels of dissemination of their work on digital preservation, for example, the information concerning digital preservation placed on their Web sites.

There is a clear need to ensure that information, tools, and experience are shared effectively within the international community. There are a number of international and national efforts detailed in the survey that have this objective. I recommend that the NDIIPP consider carefully how information on U.S. initiatives is disseminated, how such dissemination of information is given effective support, and how the initiatives relate and participate with similar activities internationally. A number of exemplars and suggestions are provided in the survey, including the need to ensure such efforts are specifically resourced. There is also potential to coordinate such work with that being undertaken by the Digital Preservation Coalition in the United Kingdom, ERPANET in Europe, and the National Library of Australia.

The levels to which different institutions are exposed to, are aware of, and respond to international developments are also pronounced. Such exposure seems highly beneficial and is apparent in many of the most successful initiatives included in the survey. Such international exposure should be encouraged within the NDIIPP.

National Library of Australia Observations and Recommendations for the NDIIPP

Know who your critical stakeholders are and work with them. For example, for the National Library of Australia (NLA), it has been essential to build goodwill with publishers, particularly in the online environment. The NLA would put a strong stress on this and the necessity of building these relationships.

Collaboration takes a lot of effort and leadership, and has its own limitations. The investment in relationship building and the diplomatic skills needed should not be underestimated. A lot of give and take is needed and results accumulate over time.

We would emphasize the importance of making a start and letting experience, practice, and policy evolve and inform each other. It is important to recognize that one cannot solve all the problems at once. Starting small on defined areas and building in feedback mechanisms for continuous learning are critical to progress.

Integrate digital preservation into the institution and do not rely solely on time-limited external or project funding to achieve your aims.

Build on the existing people and expertise. The NLA has developed its internal staff and established teams working across departments to bring together relevant skills. Look for internal synergies to support the activity.

It is initially hard to calculate costs. Cost models are dependent on a very large number of variables, and we are in an experimental phase of development. However, cost recognition and management can be improved over time.

Recognize the major challenges are not only conceptual but also practical. They need development of both policy and experimentation with strategies and procedures.

The NLA is convinced of the value of the selective approach to archiving online resources. It is one of many approaches, but for research use, the intervention of the librarian is important and cannot be replaced.

Bibliothèque Nationale de France Observations and Recommendations for the NDIIPP

How to address the deposit and preservation of online materials is a key issue. The Bibliothèque nationale de France (BnF) stressed the value of sharing research and jointly developing approaches to Web-archiving among institutions.

There is a need to do more research on collecting and preserving database-driven Web sites.

There is concern that if libraries have difficulty with CDs and other materials in proprietary standards today, it will be even more difficult for these resources to be accessed tomorrow. Influencing what is being produced by publishers is therefore a critical issue.

The BnF itself is considering undertaking more initiatives with publishers and believes early contacts with producers of electronic materials should be seen as critical.

Awareness-raising on digital preservation within institutions remains as major an issue as influencing others externally.

Koninklijke Bibliotheek (KB) Observations and Recommendations for the NDIIPP

Recognize the difference between the publishers' value-added service environment and the underlying content. Take the publications out of the service environment and into the archiving environment of the library.

Use standards such as the Open Archival Information System standard where they exist.

Work together with other organizations to encourage the development of commercial market solutions and systems for digital preservation.

From our experience in the NEDLIB Project, begin by identifying commonalities with potential partners rather than the differences. Use this to focus and scope what will be done together.

A major part of successful digital preservation initiatives is getting staff involved across the institution. There are significant change management issues that need to be addressed.

Collaboration takes time and needs a sense of community. Face-to-face contact and knowledge of partners is required.

Try and keep membership of project teams stable and avoid unnecessary changes. Continuity can be essential to maintaining progress and the relationships built up with partners.

There is a need to communicate more between institutions and share the lessons learned. All institutions agree with this in principle but it needs a staffing commitment from them to make it happen, and in practice this is rarely done because of other time commitments. Specific funding to allow institutions that develop and practice digital preservation to communicate their work may be needed.

Appoint project leaders who make things work and have a positive attitude to finding solutions to problems.

British Library Observations and Recommendations for the NDIIPP

It is important to have leadership from the front on this issue and to have strong commitment from senior management.

It is important to communicate the urgency of the problem. There is a digital time bomb with the potential for total loss.

The Digital Library Store is seen as one of the library's major initiatives. Progress has not been easy and the scoping of the project has been difficult. The British Library (BL) has learned that the requirements for access in such a large implementation proved to be very complex. There is a need for a modular approach focusing on the store, and access and integration via other systems.

There is a need for an overarching e-strategy, particularly in very large libraries with complex systems. It is important to keep all digital developments in step and to consider the interface between systems.

Much can be learned from parallel work in other institutions. Collaboration with the National Library in the Netherlands has been particularly useful for the BL. However differences in scale are an important issue when looking at the different national libraries and transferable lessons.

Strategically, it is important to do more in partnership if digital preservation is to be addressed successfully. However, collaboration can complicate things and does have costs as well as benefits.

Does the institution have the people? This is a frequently underestimated issue. The pool of specialists/generalists in digital preservation is very small. In its recruitment for a digital preservation coordinator, the British Library recruited from Australia. Digital preservation also cuts across a wide range of activities and departments. There is a need to build up awareness and capacity internally so that a wide range of staff can contribute as part of their day-to-day activities.

There are many competing initiatives in the field. It helps to try and focus and get behind one initiative, e.g., the Digital Preservation Coalition.

For collaborative activity, it can help to have the work focused at one remove from any one partner but with heavy involvement from each of the key players.

For audiovisual materials, the National Sound Archive saw the key lesson as being not to seek the "ultimate preservation solution." Many have sought it, but it is yet to be found. We need to recognize that all the challenges will not be solved instantly and a combination of approaches is likely to be appropriate at this time. We must use professional skills and harness technology now to maintain holdings in our generation and to ensure that we can plan to migrate them for future generations.

Recommendations for Further Detailed Investigation by the PricewaterhouseCoopers Consultancy

I would recommend that the following be considered in greater detail by PwC in the technical consultancy that it is undertaking on behalf of the Library of Congress:

Australia

- The PANDORA distributed national online collection and the software used to support this collaborative archiving effort;
- The NLA Digital Objects Management system and proposals for developing its capacity to manage long-term preservation;
- The NLA digital preservation work program;
- Proposals for ADRI, the national Australian Digital Resource Identifier scheme.

France

- The Web harvesting tools and approaches being developed by BnF;

- The preservation technologies being developed by INA as part of the PRESTO project.

Netherlands

- The DNEP digital deposit system for electronic publications being developed by the Koninklijke Bibliotheek (KB) and IBM-Netherlands;
- The outcomes of the Long-Term Preservation Study being conducted by the KB and IBM-Netherlands and its implications for development of a long-term preservation module as part of the DNEP;
- The digital preservation test bed being conducted by the Dutch Ministry of the Interior and outcomes that emerge from its experiments.

United Kingdom

- The digital library store in development by the British Library;
- The e-preservation strategy and systems being developed by the Public Record Office;
- The outcomes of the Cedars research project;
- The outcomes of the CAMILEON research project;
- The audiovisual and new-media preservation technologies and projects being developed by the British Broadcasting Corporation (BBC).

Other Projects and Initiatives

- The audiovisual preservation technologies being developed by RAI as part of the PRESTO Project;
- The research outcomes and tools from the NEDLIB Project;
- The Open Archival Information System reference model and implementations noted in the report.

Aims, Scope, and Methodology

This report aims to provide an overview of selected key national and multinational initiatives in digital preservation occurring outside North America. The report has examined current digital preservation initiatives in four countries: Australia, France, the Netherlands, and the United Kingdom, as well as related international initiatives. These countries and initiatives were chosen in consultation with the Library of Congress (LC) and the Council on Library and Information Resources (CLIR) as being of particular relevance and interest to the NDIIPP.

The report aims to put these initiatives into their national and international context and outline the major developments that are in progress. It is intended to provide a high-level survey. As such, it does not aim to be exhaustive or detailed in terms of practice and procedures. This report presents the key findings from the survey and details of the main initiatives in each country. It will be used by the Library of Congress to provide an international perspective on current initiatives to inform the development of national policies and programs in the United States.

The survey is written for senior administrators and for policymakers—that is, for people who are not specialists in digital preservation or access to networked information. In this respect, it is a high-level synthesis rather than a detailed document—one that teases out significant issues and main lines of development as well as their implications. The report focuses on national digital preservation initiatives in libraries, but can also include reference to relevant significant developments in other sectors. The preservation of audiovisual resources has been included as an aspect of the national reporting, e.g., the National Sound Archive as part of the British Library. Preservation of audiovisual material has not, however, been a main focus of the report.

The survey has been undertaken primarily from desktop research and information supplied by the national libraries in each country. A detailed questionnaire was developed in consultation with the Council on Library and Information Resources (CLIR) and the Library of Congress (LC). This covered specific questions on the national libraries' initiatives. The questionnaire also allowed the national context to be explored and to identify related initiatives.

Staff in the national libraries were interviewed and each national library has been visited. Supplementary visits were made to the British Broadcasting Corporation, which is the lead partner in the PRESTO project, and the Digital Longevity digital preservation test bed in the Netherlands to learn more about these specific initiatives. Prior to the interview, desktop research was completed on Web sites of relevant organizations and their staff publications, and information was entered onto the draft questionnaire.

On completion of the interviews, the completed draft questionnaire and draft report section for their country was sent to the interviewees for any comments, additions or corrections.

The following individuals were interviewed during the consultancy:

Bibliothèque nationale de France
Catherine Lupovici, Julien Masanès

British Broadcasting Corporation (BBC)
Richard Wright

British Library
Helen Shenton, Crispin Jewitt

Digital Preservation Test Bed, Netherlands
Jacqueline Slats, Maureen Potter, Tamara van Zwol, Remco Verdegem, Bill Roberts, Ingmar Evers, David Bowen

Koninklijke Bibliotheek (KB)
Hans Jansen, Titia van der Werf, Johan Steenbakkers

National Library of Australia
Colin Webb, Margaret Phillips, Pam Gatenby

The research has been coordinated where appropriate with members of the PricewaterhouseCoopers consultancy, which has been commissioned by the Library of Congress to evaluate appropriate preservation technologies and system architectures.

National Surveys

Australia

National Context

The Commonwealth of Australia has a system of federal and individual state governments. This is mirrored in its library system with a National Library in Canberra and libraries in the states and territories supported by local government. There is, therefore, a distributed system of national and regional library collections.

There is no explicit national legal deposit for electronic publications in Australia at this time, although it is anticipated that this will be introduced in due course. However, there is some provision in legislation in some states.

The National Library of Australia (NLA) has a national responsibility for preservation of the national print-based and oral documentary heritage under the National Library Act, but ScreenSound Australia is responsible for film, sound, and broadcast. ScreenSound and NLA are working jointly on proposals to extend legal deposit legislation to electronic materials and audiovisual materials in physical formats.

There is a strong digital online culture. Internationally, Australia has one of the highest levels of Internet connections among its population (surpassed only by the United States and Singapore). In part, this reflects the relative distances among population centers across the continent and the need for organizations to reach many of their audiences online. For a relatively small population, Australia has many internationally leading-edge online projects across all sectors. Archiving these online materials has become a significant area of effort for Australia’s memory institutions, and both the national library and the national archive activities and guidelines are frequently cited internationally as exemplars in this area.

There is an absence of large international Australian publishers. Most commercial Australian publishing is currently focused on print publications. Online publishing has tended to be from new entrants to the market and noncommercial sources. There are some 85 commercial publications within the national online collection (PANDORA), but this is a small part of the collection as a whole.

The National Office for the Information Economy in the Federal Government set a target for all federal services to be online by 2001. There has been a major push to rapid access to information and services from government departments.

There is a very active electronic records management/archive sector in Australia. Work at Monash University, the Public Record Office of Victoria, and the National Archive of Australia has an international profile.

There is a national bibliographic database (KINETICA), and Australian libraries collaborate in its development for resource sharing purposes. There has been a tradition of collaboration in developing the national catalog and this has provided a foundation for collaboration in other fields.

The National Library of Australia

The NLA has a staff of 492 full-time equivalents and is a statutory authority within the portfolio of the Department of Communications, Information Technology, and the Arts. Its budget in 2001 was 206.7 million $AU—$45 million of this is for operational expenses.

Its remit covers Australia's published and documentary heritage, and its sound holdings include oral and folk history.

Development of Digital Systems in NLA In 1999 the NLA prepared a tender specification for a Digital Collection Management System and issued a request for information to potential suppliers. It was unable to identify a supplier meeting all NLA requirements and has proceeded with a mixture of in-house development and external procurement in three areas:

- *The Digital Object Storage System (DOSS).* This is an external procurement built from a number of subcomponents. It was installed and accepted in June 2001.
- *The Digital Objects Management System (DOMS).* This is being built in-house for the management of both archived electronic publications and digitized objects in the NLA collections. It is a phased development, and future releases will incorporate digital sound and long-term preservation management.
- *Digital Archive System.* This software is being developed in house to support the national distributed archiving system for online publications (PANDORA). There is Web access to all functions to facilitate involvement and use by partner organizations. Version 1 has been implemented and is highly regarded by NLA partners such as the State Library of Victoria. It has substantially reduced staff time needed to archive online titles. Version 2, for release about June 2002, will also support distributed storage for any partner that requires it.

FUNDING All digital preservation activities at NLA have involved reallocation of internal resources to these activities rather than new funding. As a result there has perhaps been greater emphasis on mainstreaming these activities within the library than might have occurred if this work had been externally financed through project funding. This reallocation has been difficult, but there is now a core commitment to these activities in NLA.

Given limited funding, NLA has invested heavily in staff time and infrastructure to support collaborative archiving and seeking to develop distributed responsibility for these activities.

NLA Digital Preservation Policy and Action Plan The NLA has developed a Digital Preservation Policy that indicates the future directions the library intends to take in preserving its own electronic information resources and in collaborating with others to maximize the effectiveness of digital preservation activities. The policy is available on the NLA Web site. The NLA is six months into the two-year action plan to implement it. The NLA is interested in developing a wider national action plan with partners.

Digital preservation technologies under evaluation include:

- File format migration testing for PANDORA collection of html v4.01 migration.
- Emulation test bed for obsolete DOS systems. Test bed ongoing.
- Australian Web domain harvesting feasibility study in 2001. On hold.
- Data recovery. Work on recovery and transfer from floppy disk and CDs documented in NLA staff papers.
- Viewers for obsolete formats. The TRIM software from Tower systems has been purchased for the library's records management needs. The functionality this provides for viewing obsolete word-processing formats is being evaluated.
- CD-R and mass storage system's extensive evaluation of CD-R as an archival medium (see staff papers).
- Some evaluation of concepts for a software repository and of technology watch for file formats.

National and Institutional Initiatives

In Australia, there are a number of different national or institutional initiatives led by national bodies in different areas. National initiatives do not exist in all sectors (even institutional initiatives may be absent). The NLA has led on national collaborative initiatives for published materials, e.g., PANDORA.

NLA initiatives are coordinated either through model agreements with trade bodies, or through formal or informal bilateral arrangements with individual organizations.

Formal arrangements may be distributed by institutional mission on a geographical basis (e.g., national or state) or subject matter (archival records, publications, film and broadcast or audio). The load may be distributed unevenly depending on the resources and mission of partners in such arrangements.

The partnership that is building the National Collection of online Australian Publications (PANDORA) is based on a formal exchange of letters and entails each institution taking responsibility to varying degrees for selecting, archiving, cataloging, preserving, and providing access to selected Australian online publications, according to agreed criteria and processes. PANDORA has been in operation since 1996, and the partnership has gradually been extended over this period to other organizations, including ScreenSound Australia, the State Libraries, and one Territory library. The State Library of Tasmania has developed its own procedures and policy for its institutional initiative, Our Digital Island, and liaises closely with the National Library toward joint goals. It is strongly considering the use of the PANDORA Digital Archiving System with the option of storage of files on its own server. The diversity of approaches has been beneficial in enabling the NLA and the State Library of Tasmania to share lessons learned and to coordinate initiatives, such as developing a scheme for a national persistent identifier.

There are some areas of the national collection that remain to be covered, e.g., the evolving pre-print archives.

Incentives for participation vary from sector to sector. For publishers, deposit involves inclusion in the National bibliography, greater exposure for their publications, and ongoing access to their publication without the cost of maintaining it. The NLA agrees to restrictions on access for commercial material so that commercial interests are not threatened by deposit.

For other libraries or institutions, collaboration may secure:

- access to shared infrastructure or policy that would be expensive to procure individually, e.g., PANDORA selection guidelines,
- stronger advocacy, e.g., NLA and ScreenSound Australia's joint representation on legal deposit, and
- access to and sharing of expertise and project learning internationally, e.g., involvement in RLG/OCLC working groups.

For all entities a degree of empathy is implicit for securing the cultural heritage of Australia and therefore support for the national mission of the NLA in achieving this.

National initiatives include:

PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) The National Collection of Australian Online Publications (also known as the PANDORA Archive) is maintained collaboratively by NLA and partners. This has been in operation since 1996 and is internationally recognized as a key initiative in the selective archiving of online materials. Over time the collaboration has extended to include all state libraries (and one territory library), and ScreenSound Australia. Material for inclusion in PANDORA is selected either by the NLA itself or its partners. There is central storage of material at NLA and facilities for distributed selection, gathering and deposit through the archiving software developed in-house by the NLA. Version 2 of this software will allow distributed storage and accommodate the specific local development within Tasmania.

A template for shared selection guidelines for the National Collection of Australian Online Publications has been developed by NLA in consultation with the Council of

Australian State Librarians (CASL). This provides a consistent basis for developing a distributed national collection of online materials, while allowing for institutional collection approaches to be incorporated.

Code of Practice for Providing Long-Term Access to Australian Online Publications The NLA has developed this draft in consultation with the Australian Publishers Association (APA) to cover archiving, preservation, and access to commercial publications produced in Australia. Given the small size of the Australian commercial publishing industry, the code will not have extensive application outside of Australia for some time. However, it has been invaluable in developing awareness among the commercial publishers and preparing the ground for discussion on the legal deposit of electronic materials.

Australian Digital Resource Identifier (ADRI) The NLA is developing a national persistent identification scheme for electronic information resources in collaboration with the State Library of Tasmania and on behalf of CASL. The scheme, to be known as the Australian Digital Resource Identifier (ADRI), will provide a guide for organizations to name their resources in a way that will ensure continued access to these resources in the future. CASL endorsed in principle a draft schema for ADRI in November 2001.

Other Australian projects and initiatives include:

Our Digital Island (Tasmanian State Library) A selective Web archiving initiative for online publications in the state of Tasmania developed by the state library.

Higher Education Sector To date there has been relatively little digital preservation work in the Australian higher education sector, although a major conference, Digital Continuity, was convened in November 2001 to consider the state-of-the-art and how Australian universities should engage with the issues. There is a national digital theses program with distributed archiving by institutions but a central interface for access. Two university libraries are establishing e-print archives.

Sound Archives Australia has an active sound archiving community, which has been using digital formats for archiving for some years.

The NLA and ScreenSound Australia have instituted many evaluation and life-testing trials on CD-R and DAT tapes. The expanding capabilities of mass storage systems now make them viable for the storage and preservation of large amounts of audio data, and the NLA is progressively migrating its audio holdings from CD-R to mass storage.

The radio network of the Australian Broadcasting Corporation has implemented a computer-based digital on-air system. As a result, they do not generate analog copies of new material and now archive on recordable compact disc (CD-R).

Picture Australia and Music Australia There are a number of significant and innovative national resource discovery initiatives to access outcomes from digitization projects involving the NLA and other partners. These include Picture Australia and Music Australia.

International Initiatives

The NLA feels international collaboration at many levels is essential in digital preservation. It is keen to develop collaboration both with the Library of Congress and other international agencies.

Current collaborative international activities include the following:

Preserving Access to Digital Information (PADI) PADI is a digital preservation gateway maintained by NLA and individual/institutional partners (Australian and international). Initiated as a collaborative voluntary initiative among a number of Australian organizations, it was found necessary to create a single institution to give the program sufficient resourcing for it to develop fully. The NLA has led in development of PADI and provides staff and systems support. In 2001, functionality of PADI was extended to allow registered individuals outside the NLA to input directly into the PADI database.

There is an international advisory group for PADI and the NLA has sought to develop collaboration in maintaining PADI internationally. Individuals have been able to register as contributors and input directly since 2001. More recently, the NLA and the Digital Preservation Coalition have agreed to a memorandum of understanding on collaborative activity. This will include DPC input to PADI and a series of links and joint activity. This arrangement could be mirrored in future with other organizations worldwide.

Safekeeping Initiative This was established with set-up funding from the Council on Library and Information Resources (CLIR) in the United States. It aims to identify key digital preservation resources recorded in PADI and to secure agreements for their long-term preservation. This initiative is currently being evaluated by NLA.

Conference of Directors of National Libraries (CDNL) The Director-General of NLA is chair of the CDNL. CDNL has set up a digital issues group, which has an action plan in place that concentrates on legal deposit, persistent identification, and digital archiving and preservation research needs. This group was instrumental in submitting a digital preservation resolution to UNESCO. The digital issues group is chaired by the KB.

International Research Projects NLA staff contribute to both the OCLC/RLG Preservation Metadata Working Group and the OCLC/RLG Attributes of Trusted Digital Repositories Working Group. Review comments from NLA have had a significant input to the OAIS reference model, Networked European Deposit Library (NEDLIB), and to other international projects in digital preservation, including the development of *Preservation Management of Digital Materials: A Handbook.*

Future International Collaboration

The following were seen as potentially important areas for future international collaboration by NLA:

- work on persistent identifiers;
- exploring how national collections can be linked for wider access;
- developing a global or distributed software archive;
- documenting and sharing information on preservation dependencies in publications;
- technology watch for file and media formats;
- sharing and discussing research and evaluations of specific implementations;
- implementing preservation metadata with international publishers;
- archive certification models arising out of the OCLC/RLG Attributes of Trusted Digital Repositories Working Group;
- fail-safe mechanisms globally for collections (it was recognized this is more difficult and sensitive than some of the above suggestions and might be a lower or long-term priority).

However, it was noted that international collaboration is often easier to achieve (there is substantial goodwill) but harder to make progress on absent dedicated resources. There needs to be rigorous discussion of what is useful for both parties and identification of the resources that need to be committed.

France

National Context

The national legal deposit legislation covers publications of all types produced or distributed in France. The current legal deposit legislation was passed in 1992 and was implemented in 1993. The legislation does not specifically mention electronic publications, but implementation of the act has been applied to offline electronic publications such as CD-ROMs that have been produced in France. Under the legislation, responsibilities are divided among the following institutions:

- Bibliothèque nationale de France (BnF)—has responsibility for all published documents, videos and multimedia works;
- Le Centre national de la cinematographie (CNC)—has responsibility for film;
- L'Institut national de l'audiovisuel (INA)—has responsibility for radio and television broadcasts.

There is regional deposit for printed material and 19 regional libraries. However, there is no regional deposit for electronic publications. Two copies of handheld electronic publications must be deposited with BnF.

It is estimated that there are more than 300,000 Web sites in France (excluding hosted sites). A recommendation was made in July 2000 that the legal deposit legislation should be extended to cover electronic materials on the Web. There is currently a legal process in place to achieve this.

When this becomes law there will be an obligation for producers to deposit their Web sites if the producers are based in France. This obligation can be fulfilled by the producers themselves depositing directly by ftp or on a physical carrier, or through arrangements for harvesting by the library. The law will not specify whether Web archiving is to be selective or exhaustive, and selection decisions would be at the discretion of the library. A lot of discussion with producers is expected over implementation of any new law. Any new legislation is unlikely to be declared before 2003 or 2004.

All librarians in French research libraries, which include the university libraries and the national library, are civil servants employed by the Ministry of Education. For this reason, there is a regular movement of staff between the national library and the provinces. There is a single national school for training librarians.

There is substantial government investment in scientific research and Institut National de Recherche en Informatique et Automatique (INRIA) is one of the three centers worldwide for the Worldwide Web Consortium (W3C).

The French Archives Law sets out rules for managing public archives and for protecting private archives, which applies to all local and national public organizations. Although they are under central direction through the Archives of France (a directorate of the French Ministry of Culture), French archives are highly decentralized, with the National Archives, for example, consisting of five separate centers.

There is growing interest in the issue of long-term preservation of digital information across many sectors in France. This is reflected in a number of international conferences arranged there on the issue in the last year.

The French National Library—Bibliothèque nationale de France (BnF)

The Bibliothèque nationale de France (BnF) is funded through the French Ministry of Culture and has a staff of 2,800. It has an annual budget of 1 billion French francs for its running costs. This excludes salaries, which are controlled by and paid for separately by the Ministry of Education (the librarians) or the Ministry of Culture (other staff). Six hundred staff are on short-term contracts funded by the running costs. The library does not have a lending or document supply role. It is solely a library of last resort with onsite access to any material in copyright that has been deposited.

Digitization of collections started in 1992 and includes material both in the national library and associated library collections. There has been a strong focus on digitizing public domain print collections, which have been made available through the Gallica Web site. The digitized collection consists of homogenous documented formats and has already been migrated once. A large program for digitizing video has just started.

Digital Systems The library has 100 Unix servers, 150 NT servers, and 3000 workstations. There is a 150 Mb ATM network internally and a 150Mb connection externally via the research network in Paris. There is distributed computing power and 24/7 service capability across the library. A central archival store is, however, considered a necessary future development. The current main approach to long-term preservation is to develop a preservation metadata database to inform migration and preservation decisions across these distributed storage systems.

Funding Digital preservation initiatives are funded through existing running costs rather than additional funding. Current experimentation with Web archiving is achieved by reallocation from other budgets. However, additional funding is being sought to continue this work next year.

Digital Preservation Policy and Actions There is a separate workflow for electronic legal deposit publications within BnF. The audiovisual department takes all electronic deposit materials, as it already has equipment for accessing recorded CDs and digital tapes. Of the two deposit copies, one is retained within the audiovisual department and another copy is sent to a BnF conservation building outside Paris.

The library has just started a working group to develop digital preservation across all its departments. This has representatives from the Digital Library Project team, and the Audio-Visual, Information Technology, Conservation and Collections departments. The working group will meet every month and gather information on the scale of work needed across the library, what is being done, and what is being considered. It will adopt the OAIS model and apply it within BnF. Julien Masanès is acting as coordinator for the group and as project leader on evaluation of Web site archiving.

In December 2000, BnF launched a set of experiments concentrating on archiving the national Web domain. No access is given to materials in this experimental Web archive. Future public access onsite at the BnF will be dependent on arrangements in any revised legal deposit legislation. The goal of these experiments is to evaluate costs and to define procedures for selection, transfer, and preservation that can be applied for any new legal deposit law extension to online materials.

The library is working with the Institut National de Recherche en Informatique et Automatique (INRIA) to test its XYLEME software as a tool for Web archiving. The project leader is working with collections staff to see if the automated weighting provided by this software can be used to assist in the selection of Web sites for archiving.

National and Institutional Initiatives

The division of responsibility for legal deposit is set out in legislation and the load divided according to the type of material as noted above. There is a scientific committee that oversees implementation of the legislation.

There has also been coordinated research funded through the Ministry of Culture for research on technology. This has paid for research on producing archival quality CDs.

The national space center (CNES) has led the development of the Open Archival Information System reference model standard within France and has coordinated development of an informal group (PIN) working on this and other standards and guidelines.

Perennisation des Informations Numeriques (PIN) The Networked European Deposit Library (NEDLIB) project made extensive use of the draft Open Archival Information System (OAIS) reference model standard. This led to initial contact from BnF with staff at the national space center CNES who had been working as part of the international earth observation and space data community on developing the standard. A meeting was held of interested organizations in June 2000 to discuss the OAIS model. PIN was then established as an informal forum and a discussion list was administered by CNES. The purpose of the forum is to contribute to work on developing the OAIS standard and standards and practices for its implementation, and to share information among different organizations. Participation is voluntary and relies on contribution of in-kind effort by the individuals and organizations that attend. Meetings are hosted in rotation by members. Participants include:

- Archives of France (Archives de France);
- Archive-17;
- Bibliothèque nationale de France;
- Contemporary Archives, one of the five centers within the National Archives (Centre des Archives Contemporaines);
- CNES, the National Space Center (Centre National d'Etudes Spatiales);
- CEA, the Atomic Energy Commissariat (Commissariat à l'Energie Atomique);
- Groupe Mederic;
- Institut national de l'audiovisuel (INA);
- Institut Pasteur.

Public Records The Archives of France is developing guidelines for electronic archives. The Archives of France control the National Archives, the regional, departmental, and municipal archive agencies, as well as the archive agencies of those organizations that are authorized, by way of derogation, to manage their permanent archives.

The Archives of France is exploring working jointly on archiving government Web sites with the BnF. As French archives are very decentralized, central information technology support is limited, and there are significant technical benefits to such a collaboration. It is anticipated they will want to process Web sites differently, given archival interests in the hierarchy and administrative context of the documents.

L'Institut national de l'audiovisuel (INA) INA is responsible for the national cultural audiovisual heritage. Under legal deposit legislation it is responsible for deposits from the six national TV channels (public and commercial) and five public radio channels. Under the French Communications Law it is also has responsibility for maintaining the archives for public radio and TV.

It is one of the three major partners in the PRESTO project and is making heavy use of digitization for preservation as well as increasingly taking material in born-digital form.

INA is looking to extend its mission to the French Web and is developing a harvester with the Ecole nationale Superiore.

Academic Sector The university libraries are starting a scheme for submitting university theses in electronic formats. The scheme provides style sheets in Word and reformats submissions into XML. The project is based at Lyon University and is just beginning to consider long-term preservation. The theses will be archived by the institutions and not deposited with the BnF.

COUPERIN, the main purchasing consortia for university libraries, is concerned about the archiving and future access to journals for which they subscribe. There is reluctance to rely solely on publishers for these long-term arrangements. It has begun discussing possible arrangements for archiving electronic journals that fall outside of legal deposit with the BnF. The BnF wants to seek payment for this, but feels it will be first necessary to know more about costs. Information on the costs of digital preservation are currently too uncertain for the BnF to make contractual commits to third parties, and this will need further investigation before arrangements could be taken forward.

International Initiatives

Networked European Deposit Library (NEDLIB) The BnF was a partner in the NEDLIB project and led work on defining preservation metadata. Catherine Lupovici and Julien Masanès co-authored the NEDLIB metadata report (Lupovici and Masanes 2000).

PRESTO INA is one of the three lead partners in the PRESTO project.

Open Archival Information System (OAIS) CNES has a major involvement in the OAIS standard. It is currently leading work within the archiving group on ingest methodologies.

Future International Initiatives

The BnF is keen to follow and participate in international activities, but there are time pressures that can make it hard to participate in a meaningful way or follow everything that is happening or disseminated through e-mail lists or digital preservation gateways. The BnF highlighted the following as priority areas for future international collaboration:

- The library would like to see joint research in technical areas such as harvesting of the Web or reformatting databases behind database-driven Web sites into XML. The BnF believes this area would also be of interest to the Library of Congress.
- The Networked European Deposit Library (NEDLIB) project was highly regarded by the BnF, and it would like to see some practical extension of this activity among national libraries.

Netherlands

National Context

There is no legal deposit legislation in the Netherlands for either print or electronic publications, so the national library, the Koninklijke Bibliotheek (KB), has developed work in this area on its own initiative and as a natural extension of its national mission to collect the published heritage.

The library has developed voluntary agreements on deposit of electronic publications with publishers, starting first with bilateral agreements with Elsevier and Kluwer, then a few years later with a general agreement with the Dutch Publishers Association, signed in June 1999.

The KB still wishes to see a statutory right to archive publications, perhaps through the national implementation of exceptions in the European Union Copyright Directive. Voluntary agreements have limitations in that publishers do not always have appropriate rights in third-party materials. These difficulties could only be resolved by statutory provisions.

Dutch publishing output is dominated by two international publishers, Elsevier and Kluwer. These two publishers provide the majority of electronic journal titles accessioned by the KB.

The Dutch government aims to process 25 percent of transactions between government and citizens digitally by 2002. Because of this, there is significant investment in a program to develop strategies, methods, techniques, and tools to support e-government and information society initiatives. Concerns over business continuity and electronic records led to establishment of a Digital Longevity program within these initiatives. There are five projects within this, including a digital preservation research test bed and a task force to support awareness-raising and communication across different government agencies.

Activity in the academic sector to date has concentrated principally on establishing e-print and digital archives concerned with access and new models for electronic publishing.

There is a national plan for preservation (the Delta Plan), which has been in operation since 1991 and has assessed the preservation needs of print and manuscript materials. In 1997, a national program for the preservation of library materials (Metamorfoze) was launched. This is coordinated by and grants are distributed through the National Preservation Office of the Netherlands, which is organized by and housed in the KB. The focus of the program is on reformatting paper to microfilm, deacidification and some assessment of digitization as a preservation surrogate.

All publications deposited with the KB are cataloged into the national bibliography. The cataloging is done using a joint system of the KB and all the research libraries in the Netherlands. The joint cataloging system is technically maintained by Pica/OCLC. From the resulting bibliographic database, the national union catalog is produced and used for resource sharing.

The National Library of the Netherlands

The KB has a staff of 350 (about 260 full-time equivalents) and is funded through the Ministry of Education, Culture and Science. Its annual budget in 2002 is 80.4 million guilders (36.5 million euros). The library is funded through the science section of the ministry.

The KB collects the published and literary heritage of the Netherlands. Its collections are primarily focused on book and serial publications. This can include multimedia publications, but it does not collect any audiovisual, film, broadcast, databases, software or games. Databases may, however, become a future collecting area. It is also interested in future selective archiving of parts of the Dutch Internet domain.

KB initiatives include development of the digital archive store project (DNEP), a national agreement on voluntary deposit with publishers, a long-term digital preservation study with IBM and many digitization projects, including the Memory of the Netherlands and Treasures of the National Library. The latter are focusing on improving access and interoperability with other collections.

Development of Digital Systems There has been investment in developing access systems and particularly in Web access to the catalogs.

Development of the new deposit system for electronic publications has occurred in a number of distinct phases.

The KB was the lead partner in the European Union-funded Networked European Deposit Library (NEDLIB) project and helped develop its guidelines for electronic deposit systems. These guidelines propose creation of a controlled environment for storage and maintenance of electronic publications (the deposit system) and development of transfer procedures for electronic publications to the deposit system. NEDLIB also employed Jeff Rothenberg from the RAND Corporation to investigate the feasibility of emulation as a long-term solution for digital preservation.

In 1999, the KB investigated the feasibility of obtaining an operational deposit system for its electronic publications from commercial integrated circuit technology suppliers. The KB concluded that the storage and management functions could be obtained from existing vendors. However, for long-term preservation and access, it was clear that there were no off-the-shelf solutions available, so it would be necessary to commission specific research to develop the required functionality.

In September 2000, the KB contracted with IBM-Netherlands to build the new deposit system. The Deposit of Netherlands Electronic Publications-implementation (DNEP-i) contract also includes applied research from IBM to develop new functionality for long-term preservation and access. A major requirement of the KB was that the system should be compliant with the Open Archival Information System (OAIS) standard. The KB required the design of the system and the long-term preservation (LTP) study to be strictly linked together. IBM is developing the data model so that in the near future an operational LTP-module can be fitted into the system.

At the end of 2001, the first module for Delivery and Capture was made available. The system will be completed in October 2002.

It is intended that the DNEP-i project will result in an OAIS-compliant operational deposit system, as well as test and demonstrate requirements for the future development of a Long-Term Preservation Module, which must be added to the deposit system.

This long-term preservation module will be needed to:

- Identify digital objects in danger of becoming inaccessible due to technology changes;
- Implement preservation strategies to address these dangers, i.e., migration and emulation;
- Supply the technical metadata necessary to generate and validate the required viewing environments for digital objects during delivery.

Preservation Actions The long-term preservation study (LTP) will involve six months of work elapsed over one year and cost 300,000 guilders (136,134 euros). Its objective is to investigate the functionality required for the long-term (hundreds of years) preservation of the digital information stored in the DNEP. The study began in November 2000, with the start of the DNEP-i project. It aims to cover the following issues:

- *Implementation of Long-Term Preservation.* The initial DNEP system has only a limited functionality for maintaining the technical data (hardware and software components) needed to render the stored digital objects. One of the main responsibilities of the LTP Study is to define the functional requirements of the Preservation subsystem not considered in the initial DNEP release. In the end, the Preservation subsystem should maintain all the relevant technical metadata needed to render the digital objects.
- *UVC Proof of Concept.* The preservation approach advocated by Raymond Lorie at the IBM Almaden Research Center, based on the use of a Universal Virtual Computer (UVC), is being refined and validated in the specific context of the KB.
- *Large Media Migration.* Due to the high volumes involved, electronic deposit applications face specific problems while migrating information from one media to another.
- *Authenticity.* A workable framework to define authenticity of digital objects is needed to evaluate the success of the preservation activities of any specific electronic deposit.

Five LTP Study reports are being produced (on these four issues above plus a general synthesis) jointly by the KB and IBM and will be published in summer 2002.

The KB is participating as a test site in the final year extension of the Cedars Project (see below).

The KB has undertaken an experiment with the NEDLIB Web harvester to investigate the Dutch Web domain. It found only 20 percent of sites were of interest to the KB and a significant number of these were database-driven.

The KB has undertaken research on workflows for electronic journals, which are being implemented in the new system. It is also developing a new workflow for CD-ROMs to be integrated into the new deposit system.

Funding The KB has a national role in the public interest and therefore bases its activities on public funding from government. It believes its public service role is paramount and would not wish to adopt a commercial model. Services are billed on a cost-recovery basis.

Between 1998 and 2001, the KB has received 3.2 million guilders (1.45 million euros) over four years plus some research funding to prepare the development of the new deposit system. Structural funding of 2.5 million guilders (1.14 million euros) per annum for ongoing support of this activity will be available in 2003 and beyond.

The KB currently holds the Dutch imprint of the Elsevier group under a voluntary deposit arrangement and has been archiving a subset of its electronic journals for some years. It has agreed with Elsevier to archive a copy of all its electronic titles. This extension of its activities in the end might need additional funding depending on the range and nature of the services to be delivered. Economic models to support this are under investigation. The KB would not wish to charge users other than for cost recovery of specific services. It is interested in funding models in which such services are free to the user but paid for by the producer, who recovers this cost in its product pricing (examples of this funding model are the Digital Object Identifier [DOI] used in publishing or barcodes).

Most collaborative initiatives are not funded but rely on matching in-kind contributions of staff and other resources from the partners.

National and Institutional Initiatives

DIGITALE DUURZAAMHEID (DIGITAL LONGEVITY) There are five projects within the government Digital Longevity program, including a digital preservation research test bed and a task force to support awareness raising and communication across different government agencies. Other projects concern central government databases, record-keeping systems, and quality of records. The program is run by ICTU, an agency established to oversee the e-government program.

The KB is a member of the task force for the Dutch government Digital Longevity program. As part of the program, the National Archives has been discussing renting part of the storage space on the KB platform to provide interim storage for electronic records transferred from government departments.

Digital Preservation Test Bed (Test Bed Digitale Bewaring) The Ministry of the Interior and the Ministry of Education, Culture and Science (the National Archives) established a three-year digital preservation "test bed" as part of this program. The project began in October 2000 and will conclude in September 2003. The test bed was preceded by a research study by Jeff Rothenburg (Rothenburg and Bikson 1999). The Digital Preservation Test Bed is carrying out experiments according to predefined research questions. It is researching three different approaches to long-term digital preservation: migration, emulation, and XML, and is experimenting with text documents, spreadsheets, e-mail messages, and databases of different size, format, complexity, and nature. The effectiveness of each approach for different material is being evaluated, together with their limitations, costs, and application potential.

The following outcomes are expected:

- advice on approaches for current digital records in government departments;
- recommendations for the best preservation approaches applying in specific circumstances;
- functional system requirements for preservation;
- cost models for different preservation approaches;
- preservation approach decision trees;
- recommendations for new legislation.

To date, the project has produced the following public outputs: a research base (list of relevant projects is available online) and a white paper on migration.

The project is collaborating with the Public Record Office (PRO) in the United Kingdom and the National Archives and Records Administration (NARA) in the United States and has informal links to ERPANET and Interpares through staff at the Dutch National Archives.

Public Records Dutch archives are funded through the Ministry of Education, Culture and Science. Historically, there have been a federal government archive and 12 state archives with some local archives for specific municipalities or polders. The national structure is currently being reorganized to create a federal government archive with a national archive service of regional archive centers. The 1995 archives legislation covers electronic public records and requires that they be transferred to the archives after 20 years. Regulations introduced in 2000 specify the formats and metadata in which the records must be presented.

Netherlands Institute for Scientific Information Services (NIWI) NIWI is an institute of the Royal Netherlands Academy of Arts and Sciences. It curates and provides access to primary research data and research information, in the fields of biomedicine, social sciences, history and Dutch language and literature. In addition, it supplies information about research and researchers in the Netherlands, in all scientific fields. Its current projects include one in the field of digital preservation. Archiving Digital Academic heritage (ADA) is a pilot project to explore the feasibility of setting up digital archiving services for scientific or scholarly research material in the Dutch academic sector. In the pilot, the research data files of the Meertens Institute are being archived. Marketing research will also be undertaken to establish the level of the demand for archiving services in the academic sector.

Roquade Three Dutch university libraries were partners in the Roquade project: Utrecht University Library, Delft University of Technology Library, and the Netherlands Institute for Scientific Information Services. The project researched development of electronic archives in the academic sector to enhance scientific communication. The costs of metadata assignment, administration, and quality control and technical infrastructure for an electronic archive accepting 5,000 items per year was estimated by the project to be 29 euros per information item.

The Academic Research in the Netherlands Online (ARNO) This project is developing university document servers to make available the scientific output of participating universities. Project participants are the University of Amsterdam, Tilburg University, and the University of Twente. The project is building on earlier Dutch electronic publishing projects and the Open Archives Initiative.

International Initiatives

NEDLIB (Networked European Deposit Library) The KB chaired the NEDLIB project funded by the European Union between 1998 and 2000. NEDLIB was a collaborative project of national libraries and other partners researching the basic infrastructure upon which a networked European deposit library could be built.

CONFERENCE OF EUROPEAN NATIONAL LIBRARIANS (CENL) The KB participates in the Conference of European National Librarians (CENL) and occupies the CENL chair. This is an independent association of the chief executives of the national libraries in member states of the Council of Europe.

COBRA+ FORUM The KB also participates in the COBRA+ Forum. COBRA+ is a standing committee of CENL. COBRA was the key forum for developing proposals for European projects such as NEDLIB or TEL (The European Library).

KB/BRITISH LIBRARY MOU The KB has had a memorandum of understanding since 1995 with the British Library covering collaboration on digitization. In December 2000, this agreement was updated to include collaboration on digital preservation. The BL has observer status on the KB/IBM Long-Term Preservation study, and there is joint review of documents and other items as they both develop their deposit systems.

CEDARS The KB is participating as a test site in the final-year extension of the Cedars Project. They have participated in the discussion to defining significant properties of publications and use of the Cedars namespace in the demonstrator project to look at allocating and cross-referencing persistent identifiers.

CONFERENCE OF DIRECTORS OF NATIONAL LIBRARIES (CDNL) CDNL has set up a group on digital issues, which is chaired by the KB. This has an action plan that concentrates on deposit agreements, persistent identifiers, and digital preservation research needs. This group was instrumental in getting a digital preservation resolution adopted by the UNESCO General Conference. The Dutch national government submitted this UNESCO resolution and the KB played an active role in formulating the text. The KB occupies the CDNL vice chair.

INTERNATIONAL RESEARCH PROJECTS KB staff contribute to the OCLC/RLG Preservation Metadata Working Group. Review comments from the KB have had a significant input into the OAIS reference model. Staff from the KB also regularly present papers at relevant international conferences.

Future International Initiatives

The following were seen as potentially important areas for future international collaboration by the KB:

- For long-term preservation activities, there will be a need to develop registries of file formats, migration tools and emulators, and technology libraries with obsolete software and their documentation. Although such registries could be developed individually by libraries, there are obvious cost benefits in collaboration. Such services could easily be networked and shared on an international basis.

- There is a need for more research on long-term preservation.
- National libraries are developing new workflows and skills to handle digital materials. Experience and emerging practices should be shared internationally.
- There should be more discussion and collaboration internationally on selection and who takes responsibility for long-term preservation. National libraries will always have a responsibility for their own cultural heritage. However, increasingly, electronic publishing and businesses are global rather than national in scope, and national imprints are less easy to define. Alongside the national collections we may see the development of archives for international publishers. There is an issue of how these international collections can be funded and fitted within national frameworks and institutions. Potentially some national libraries may undertake a wider international role where new funding models can support this activity.
- The KB has undertaken some pilot activity in Web archiving but recognizes that some of the other national libraries now have substantial experience in this field. It believes Web archiving is an area with substantial scope for collaboration and sharing of experience and tools among the national libraries.
- The perception from the KB is that there is little real research on digital preservation in memory institutions. In part at least this is due to reliance on external funding. Funding bodies are currently focusing on a lot of low-risk activity: workshops, reports, etc. To counteract this, there would be a strong case for raising funds among institutions that could be targeted at digital preservation technologies research. Such research need not be expensive if the cost is shared by several institutions.

United Kingdom

National Context

There is a network of copyright deposit libraries in the United Kingdom consisting of the British Library (the national library for the United Kingdom), the National Library of Wales, the National Library of Scotland, the Bodleian Library Oxford, Cambridge University Library and Trinity College Library Dublin. A Standing Committee of Legal Deposit Libraries (SCOLD) provides a forum for joint discussion and activities.

There is currently no legal deposit legislation for electronic materials, although forthcoming legislation is anticipated. The British Library has a Joint Committee on Voluntary Deposit (JCVD), which is a forum for discussion with publishers and the other United Kingdom copyright libraries on progress with voluntary deposit of electronic publications and future legislation. It is anticipated that some degree of distributed archiving will be adopted, although the substantial part will probably be at the British Library.

There is a significant government-led move toward devolving powers to the regions, and there are few institutions with an absolute or United Kingdom-wide mission.

There are very large and long-established publishing and music industries in the United Kingdom, and this is reflected in the size of the British Library's and other library holdings and collections. A significant number of commercial and noncommercial publishers based in the United Kingdom are now producing digital works. These publishers include a large number of small and medium-size as well as large international publishers. The national mapping agency is now entirely based on digital surveys and deposits snapshots of its national topographical database (a Geographical Information System) with the British Library.

The United Kingdom government has a significant drive toward electronic delivery of all local and central government services as set out in *Modernising Government.* The target date is 2004, and this will have significant impact on Web delivery and electronic record-keeping. This is reinforced by progressive implementation across all public sectors of a Freedom of Information Act.

The United Kingdom has a diversified range of cultural institutions with digital preservation initiatives arising from their institutional missions (many of which extend beyond the institution concerned). These initiatives have had a high profile internationally.

There is significant centralized digital research and development funding and programs for the higher education and further education sectors in the United Kingdom through the Joint Information Systems Committee of the Higher and Further Education Councils (JISC). The digital focus of JISC and its central funding and direction mean that the higher education sector has played a major part nationally and internationally in digital preservation initiatives.

Across all sectors, memory institutions face significant funding constraints and have static or declining core budgets in real terms. They are required to balance the demands of traditional and electronic materials, and demands in both areas continue to grow.

Across the United Kingdom (perhaps with the exception of the data centers provided for primary research data), the focus of digital preservation to date has been on pilot projects, research and development of guidelines. Although much has been achieved, there is a growing desire to move from projects to services. This is difficult to achieve when "new" funding is often of relatively short duration and project-oriented.

The limited funding available to institutions individually and the scale of challenges involved have prompted partnership and collaboration among institutions and serious discussion of the issue of whether responsibilities can be identified and shared among them. Discussions are at an early stage and arrangements are likely to take some time to evolve.

Although there is now a reasonable degree of awareness of digital preservation among curators and academic sectors in the United Kingdom, there is extremely low awareness publicly and across key stakeholders, including senior civil servants, members of Parliament, funding bodies, and publishers. This is seen as a major impediment to growing funding for digital preservation activities and in engaging with major stakeholders.

The British Library

The British Library has a staff of 2,400 and is a nondepartmental government body funded through the Department of Culture, Media, and Sport. Its budget in 2000-2001 was 110.26 million pounds, of which 82.27 million pounds was government grant in aid and the remainder other income of 28 million pounds (principally from document supply services).

It is the national library of the United Kingdom and has major international collections. Its sound holdings include published music, drama and literature, international music, wildlife sounds, and oral history.

There have been a number of significant digitization projects for enhancing access to the collections made possible by external project funding from organizations such as the Mellon Foundation and New Opportunities Fund (a distributor of United Kingdom lottery funds). Within the next four years, it is anticipated there will be more than 1 million digitized images.

Development of Digital Systems A recent procurement exercise for a long-term preservation facility, the Digital Library Store (DLS), led to a 10-year contract with IBM, and design of the system is now in progress. There is significant collaboration with the national library of the Netherlands on this development.

Digital storage for large-scale items such as digital master (TIFF) images is currently provided through a contract with the University of London Computer Centre.

Funding All activities had been funded from existing government grant-in aid funding. There has been no increase for digital preservation activities. As other demands are also increasing, this has meant cutting back in some areas to fund new developments. The Digital Library Store budget is commercial in confidence but this is a multimillion-pound investment by the BL.

Collaboration in digital preservation activities has been on the basis of joint in-kind contributions of staff time and resources. For the Digital Preservation Coalition, the BL also contributes 10,000 pounds per annum as a full member.

There have been bids to the government from the six copyright libraries to develop a secure network among deposit libraries. This would allow shared access to a single deposit for electronic materials and scope for distributing the archiving responsibility.

So far, these bids have not been successful. The libraries are now proceeding with a small demonstration project, with project funding contributed jointly among them.

The library has made a bid to government for 600,000 pounds annually starting in 2004 to begin selective Web site archiving combined with regular snapshots of the United Kingdom Web domain.

Digital Preservation Policy and Actions The BL has issued *Strategic Directions,* a future strategy for the BL, which emphasizes development of electronic collections, digital preservation, and partnerships with other organizations. It suggests the collection policy will increasingly focus on the United Kingdom's published and literary heritage, and there will be more focused acquisition of overseas publications. A public consultation on the proposed strategy is being undertaken and responses are being evaluated. There is a digital preservation policy used as a working document internally.

It is intended that legal deposit of electronic publications will cover all physical format and online publications in the United Kingdom but possibly with special arrangements for commercial databases. Development work on the Digital Library Store is a key preparatory action prior to the introduction of any extension to legal deposit. The Domain UK project has harvested 100 United Kingdom Web sites for a selected range of subject areas with permission from rights holders, and this experience is helping shape future selection guidelines.

Voluntary deposit of electronic publications was introduced in January 2000. It has been concentrated on physical formats published in the United Kingdom, but some publishers have also chosen to deposit online materials. Since January 2001, 3,000 electronic publication titles have been received under voluntary deposit from publishers. The voluntary deposit has been focused on the British Library initially but may extend to other copyright libraries in due course. However, it is anticipated only one copy would be deposited and access would be shared over a secure network.

Significant issues that have emerged from operation of the voluntary deposit scheme include the treatment of very high-value commercial databases and the metadata that can be supplied by publishers (particularly small and medium-size publishers) to accompany the deposit.

To address the metadata issue, the British Library is partially funding development of a software package to assist generation of metadata to the ONIX standard being developed by publishers. The purpose of the Simple ONIX Editing Tool (SIMONE) is:

- training ONIX users, especially those involved in ONIX record entry and record maintenance;
- record entry, maintenance, and export (for delivery as ONIX messages) by small ONIX users.

A “small ONIX user” is typically expected to be a small publisher needing to create fewer than 100 records per year and in total maintain fewer than 1,000 records, including front and back product lists.

By contributing to the development of the SIMONE software, which encourages the use of a common standard, the British Library is hoping to create future efficiencies and simplify the data input to its digital library systems.

In 2001, the BL appointed a digital preservation coordinator based in the Preservation Department to coordinate activities across the library and with external agencies and to provide a focus for advice and guidance to staff.

A range of pilot digital preservation projects are under way, ranging from the voluntary deposit of electronic materials with publishers (to test procedures and policy in advance of any legal deposit provisions); the Web site archiving pilot (Domain UK); to e-manuscripts and e-correspondence. The BL has also undertaken earlier pilot activities such as the CD-ROM demonstrator, which explored ingest procedures and costs for CD-ROM accessions.

The library contains the National Sound Archive, which has a longstanding voluntary deposit scheme with the music industry. The Sound Archive has undertaken research on the archival quality of CDs and will gradually move from offline CD storage toward mass storage as the Digital Library Store is completed.

Staff training has been seen as a significant issue within the BL, and it has organized a number of internal “e-fairs” to demonstrate current projects to all staff, lectures, seminars, and “learning circles” across departments.

The library has been used as a test bed for a number of research projects, including Cedars, *The Preservation Management of Digital Materials Handbook*, and LOCKSS.

The library is one of the founding members of the Digital Preservation Coalition and its chief executive is its current chair.

National and Institutional Initiatives

The Digital Preservation Coalition Establishing a Digital Preservation Coalition was the primary recommendation of a United Kingdom digital preservation workshop convened at Warwick in 1999. The Coalition was established in July 2001 with the aim of pursuing a United Kingdom digital preservation agenda within an international context. It is a membership organization and has a structure of full members, associate members, and allied organizations. Over a period of eight months, its membership has grown to 19 organizations. It is cross-sectoral and includes all the significant institutions in the United Kingdom library and archive sectors as well as publisher organizations, research institutes, government agencies, and service providers.

Initial support for developing the Coalition has come from JISC through part-time involvement of a JISC-funded program director and funds built up by membership contributions. The coalition is a limited company. It has been a grassroots development with limited initial funding. It has focused initial activity on advocacy, including a successful public relations campaign to raise public awareness of digital preservation through the national press, and a launch at the House of Commons. To date it has held two members' forums, the first on digital curation (particularly the Open Archival Information System [OAIS]) standard and the United Kingdom e-science program) and the second on Web archiving. Information on the Coalition is disseminated through its Web pages and the digital-preservation list on the JISCmail listserv.

The Coalition has the following long-term goals:

- producing, providing and disseminating information on current research and practice and building expertise among its members to accelerate their learning and generally widen the pool of professionals skilled in digital preservation;
- instituting a concerted and coordinated effort to get digital preservation on the agenda of key stakeholders in terms that they will understand and find persuasive;
- acting in concert to make arguments for appropriate and adequate funding to secure the nation's investment in digital resources and ensure an enduring global digital memory;
- providing a common forum for the development and coordination of digital preservation strategies in the United Kingdom and placing them within an international context;
- promoting and developing services, technology, and standards for digital preservation;
- forging strategic alliances with relevant agencies nationally and internationally and working collaboratively together and with industry and research organizations to address shared challenges in digital preservation;
- attracting funding to the Coalition to support achievement of its goals and programs.

The United Kingdom focus of the Coalition was adopted on the pragmatic basis that the initiative would involve considerable time in building relationships and membership and realistically should therefore be confined to the United Kingdom. At the same time, the founding members recognized the global nature of the challenges and the need to be linked to and foster international activity. The early work of the Coalition has therefore already focused on this international context with involvement in Open Archival Information System standard workshops and development of a collaboration agreement with the National Library of Australia. It also has a number of international members with United Kingdom interests.

Public Records In the United Kingdom, public records are those of the central government rather than local government, which are covered by separate legislation. The Public Records Act is seen as covering electronic records, although it is likely that further legislation will be required. The Public Record Office has an electronic records program and is providing guidance and tool kits for government departments. It is currently procuring a new storage system as part of its new e-preservation strategy and developing new initiatives to support this. Preservation of large-scale government datasets has been undertaken via a seven-year service contract with the University of London Computer Centre. Responsibility for public records in Northern Ireland and Scotland lies with the Public Record Office of Northern Ireland and the National Archives of Scotland respectively.

JISC Digital Preservation Focus The Joint Information Systems Committee of the Higher and Further Education Councils (JISC) is an institution unique to the United Kingdom and is funded through a "top-slice" from public funding distributed via the councils to universities and colleges. It has had a significant involvement in United Kingdom digital preservation initiatives. In June 2000, it established the JISC Digital Preservation Focus to provide further coordination to these initiatives, to develop strategy and guidelines, and to establish a Digital Preservation Coalition with partners.

The JISC Interim Preservation Strategy is shortly due to be revised, and the initial three-year program of activities is drawing to a close. The JISC is therefore developing a new future program. This will include further digital research initiatives and establishing a number of new programs and services for digital preservation in the HE/FE sector. A major area of concern is scholarly publishing and archiving arrangements for the large number of e-publications used in United Kingdom research and teaching that will fall outside any likely extension to United Kingdom legal deposit legislation. Other significant areas are likely to be institutional e-print archives and electronic records, project Web sites, structures to support digital preservation research, and e-science.

Primary Research Data The United Kingdom has a Data Archive for the Social Sciences, which was established in the early 1970s, and a series of Data Centres for data funded by the Natural Environmental Research Council. The national laboratories and the Sanger Centre also hold significant collections. Large-scale effort and funding are now being directed to developing "e-science" and a research grid in the United Kingdom. There are close links to similar developments in the United States, Europe, and elsewhere worldwide. There are significant digital curatorial issues within the grid, and digital preservation is seen as an important issue for scientific data that will be generated over the next decade. Linkages with digital library and preservation research are being explored and could lead to significant investment in collaborative research.

Since 1996, the Arts and Humanities Data Service (AHDS) has been developed to provide data and preservation services in the arts and humanities. Established in 1996 by JISC as a three-year project to collect and preserve primary digital materials for research in the arts and humanities in the United Kingdom, the AHDS has subsequently moved to being a jointly funded service of JISC and the Arts and Humanities

Research Board. The service was established on a distributed model with a central executive and five subject-based service providers.

The AHDS has produced a number of highly regarded guides to good practice and a distinctive digital collections policy. This formed the initial basis for the research study called *A Strategic Framework for Creating and Preserving Digital Collections* (Beagrie and Greenstein 1998). Currently the AHDS is undertaking a digital preservation audit of its holdings and will utilize this to inform and revise its preservation guidance.

CEDARS Cedars is a four-year research project to examine preservation of electronic publications funded by JISC and undertaken by the Consortium of University Research Libraries. Initially funded for three years, it was extended another year to fully document and disseminate its findings and extend involvement in its work to new institutions. Five reports are being produced from the final year of the project. A national invitational workshop involving publishers and libraries was held in February 2002. Further information and documents are available on the Web site.

The project has provided important conceptual advances in preservation metadata and influential ideas on significant properties, representation networks, and distributed archiving. It has also raised awareness of digital preservation issues among the research library and publishing communities in the United Kingdom.

The project concludes in March 2002. JISC is now undertaking a consultancy on archiving e-publications for United Kingdom higher and further education. It is seeking to develop and move forward outcomes from Cedars through the future programs of JISC, the Digital Preservation Coalition, and work in other archiving programs such as the National Libraries.

THE NATIONAL PRESERVATION OFFICE The National Preservation Office for the United Kingdom and Ireland is based at the British Library. It coordinated development of a series of seven JISC/NPO digital preservation research studies and has published other studies in this field. It is an allied organization of the Digital Preservation Coalition, and the two organizations have agreed to a memorandum of understanding on their respective roles and joint collaborative activities.

PRESERVATION MANAGEMENT OF DIGITAL MATERIALS: A HANDBOOK Development of this handbook was undertaken by AHDS and the JISC Digital Preservation Focus. The research aimed to provide overviews of the key issues, decision trees and checklists and to select significant research and exemplars worldwide. It was published by the British Library in October 2001 (Jones and Beagrie 2001). A Web version will be made available and maintained by the Digital Preservation Coalition during 2002. The *Handbook* is being linked to the Preserving Access to Digital Information Gateway through a collaboration agreement between the Digital Preservation Coalition and the National Library of Australia.

British Broadcasting Corporation (BBC) Preservation Program The BBC is one of the largest and oldest public broadcasters of television and radio programs and has a significant corporate archive. Responsibility for archiving this content rests with the BBC under its charter (but there is no separate funding stream for this). The BBC archive is a corporate archive, and the overwhelming majority of use is focused on servicing internal users. It is investing significantly in digital content both online and through digital delivery of programs. Although about 5 percent of TV holdings are digital (less than 5 percent for radio), most new programming is now digital, and digitization is seen as a key preservation method for analog holdings. There is a 60 million-pound preservation program over 10 years for its TV and radio archives. BBC Online is one of the most popular Web sites in Europe. A new-media archivist has been appointed to develop records management and archiving of this and other digital content. The BBC is also one of the leading players in the European Union-funded PRESTO program examining preservation of broadcast archives.

International Initiatives

A feature of the United Kingdom is that most of its digital preservation projects and initiatives involve international participation, and this is often on a significant scale. The United Kingdom also has a major role in the development of international standards and working groups, including development of the ISO Open Archival Information System standard, Interpares, and the RLG/OCLC working groups on preservation metadata and attributes of trusted digital repositories.

Significant emphasis is placed (particularly within the higher education sector) on dissemination and current awareness through the Web and e-mail discussion lists as well as through printed publications. There is therefore extensive international access to information on current digital preservation work in the United Kingdom or work internationally that is seen as significant from the United Kingdom perspective.

The Humanities Advanced Technologies and Information Institute (HATII) is one of the four partners in the recently established ERPANET project funded by the European Union.

The Digital Preservation Coalition has a memorandum of understanding with the National Library of Australia and directly supports the Preserving Access to Digital Information (PADI) Gateway through input of United Kingdom material.

There are also specifically international projects:

Camileon A three-year research project on digital preservation strategies, particularly emulation, jointly funded by JISC and the U.S. National Science Foundation, based at Leeds (technical research) and Michigan (user evaluation). The project has looked at both emulation and migration as preservation strategies using now obsolete operating systems, programs, and data in its test materials. Funding for the United Kingdom research will conclude in September and all United Kingdom deliverables are expected by December 2002. The technical approaches advocated for both migration on demand and emulation are of considerable interest and deserve wider discussion and testing.

National Library of the Netherlands (KB) and the British Library There has been a memorandum of understanding for sometime between the two organizations and this has recently been extended to cover digital preservation activities. There is collaboration on deposit systems implementation, and also the BL is an observer on the KB long-term preservation research study.

European Sound Archives A formal network of European national audiovisual archives is being established. There is a draft statement of intent on cooperation between them that is expected to be finalized and published after a meeting in Denmark later this year.

Future International Initiatives

The following are seen by the British Library as potentially important areas for future international collaboration:

- The British Library is already working closely with the national library of the Netherlands and would welcome including the Library of Congress in future research on digital preservation. There is an opportunity to look at research on metadata and working with producers, particularly publishers who are operating internationally.
- Within Europe, the European Union-funded European Library (TEL) project is undertaking a feasibility study into shared access to digital collections in the European national libraries. This could also provide opportunities for collaboration on digital preservation.
- There is potential for future services to support digital preservation in institutions to develop on an international basis. This could include software repositories and other tools.
- There is scope for greater international participation in and collaboration with the Digital Preservation Coalition, particularly as it develops services.
- A number of national libraries and other institutions internationally are now using the OAIS standard as a reference model for development of their digital archives. As initiatives develop, there is opportunity to share experience of implementations and issues that arise. For example, the British Library has recently had to consider whether records of items should never be deleted (as suggested in the OAIS standard) or whether in some exceptional cases, this may be required.
- Overall the library would like to see more research and involvement with computer science research departments both in the United Kingdom and internationally on digital preservation issues.

- It envisages the further international collaboration could occur on many different levels, from pure “blue sky” research to research and development projects on specific problems with sister institutions.

Related Multinational Initiatives

Electronic Resource Preservation and Access NETwork (ERPANET)

The European Commission-funded ERPANET Project was launched in November 2001 and will run initially for 36 months. Nine hundred thousand euros of this 1.2 million euro project comes from the European Commission. The project is managed by four partners:

- The Humanities Technology and Information Institute (HATII), University of Glasgow
- Rijksarchiefdienst, the Netherlands
- Institute for Archival and Library Science, Università degli studi di Urbino, Italy
- Schweizerisches Bundesarchiv

It is a new initiative and is just establishing itself. The following information is taken from its Web site *(www.erpanet.org)*.

The ERPANET project aims to establish an expandable and self-sustaining European Initiative, which will serve as a virtual clearinghouse and knowledge base in the area of preservation of cultural heritage and scientific digital objects.

The dominant feature of ERPANET will be the exchanging of knowledge on state-of-the-art developments in digital preservation and the transfer of expertise among individuals and institutions. More specifically, ERPANET will deliver a range of services (e.g., content creation, advisory service, training, and thematic workshops and forums), both to information creation and user communities. It will make accessible tools, knowledge, and experience. ERPANET will not directly carry out new research to develop such tools, but it will create a coherent platform for proactive cooperation, collaboration, exchange and dissemination of research results, and experience in the preservation of digital objects. It will bring together research institutions, memory organizations, and the integrated circuit technology, entertainment and creative (e.g., broadcasting) industries and provide effective, multidisciplinary knowledge and resource-sharing infrastructure.

ERPANET will enhance the preservation of cultural heritage and scientific objects through nine core objectives. It will:

- identify and raise awareness of information about the preservation of digital objects;
- appraise and evaluate information sources and developments in digital preservation and make available results of research including ongoing EU supported projects;

- provide an inquiry and advisory service on preservation issues, practice, and technology;
- implement six development workshops to bring together experts to tackle key preservation issues;
- hold a suite of eight training seminars based on best practice reflecting the needs of the community;
- develop a suite of tools, guidelines, templates, and 60 case studies;
- stimulate research and encourage the development of standards in the areas of digitization and digital preservation from within existing European Union-supported projects and within Europe;
- build an online community; and
- stimulate awareness among software producers of the preservation needs of the user community.

Networked European Deposit Library (NEDLIB)

This three-year project was launched on January 1,1998, with funding from the European Commission and ended on January 31, 2001. The project was established to explore the technical and managerial issues involved in developing digital deposit libraries for electronic publications.

The project partners were eight national libraries, a national archive, two information technology organizations, and three publishers. The Koninklijke Bibliotheek (KB), the National Library of the Netherlands, led the project with Johan Steenbakkers as the project director.

Outcomes

Deliverables from the project included:

- the addition to the OAIS standard of a function for long-term preservation planning;
- a model for a deposit system supporting the capture, storage, access, and long-term preservation of electronic publications;
- guidelines to best practices, technical standards and solutions, methods and procedures for practical implementation;
- small-scale development and testing of software tools used to build deposit systems;
- a proof-of-concept demonstrator of a deposit system for electronic publications.

A series of seven reports were produced as follows:

- *An Experiment in Using Emulation to Preserve Digital Publications* (Rothenberg 2000).

- *Metadata for Long Term Preservation* (Lupovici and Masanès 2000).
- *Standards for Electronic Publishing: An Overview* (Bide & Associates 2000).
- *Standards for a DSEP: Standards for the Implementation of a Deposit System for Electronic Publications (DSEP)* (Feenstra 2000).
- *The NEDLIB Guidelines: Setting up a Deposit System for Electronic Publications* (Steenbakkers 2000).
- *A Process Model: The Deposit System for Electronic Publications* (van de Werf 2000).
- *List of NEDLIB Terms* (Clavel-Merrin 2000).

The NEDLIB work has been taken forward in implementation of the KB's new Deposit system, the DNEP, by IBM-Netherlands. The preservation metadata have also been adopted for use within the BnF and in its planning for a database of preservation metadata. A report of the local situation in each national library partner was published in July 2000 (Borbinha and Cardoso 2000).

NEDLIB also provided a small-scale development and testing of software tools used to build deposit systems including:

- *The NEDLIB Harvester.* A freeware application for harvesting and archiving Web resources. The application is maintained jointly by Helsinki University Library and the Center for Scientific Computing. The harvester, its pilot use within NEDLIB and its subsequent operational use by the national libraries of Iceland and Finland is described by Juha Hakala (Hakala 2001). Further collaborative development of access tools for Web archives is being undertaken by the Nordic Web Archive (NWA).
- *MMB-System for Multimedia Access.* MMB is an integrated client-server environment to support the workflow for electronic publications. Since October 1999, the MMB system has been in use at Die Deutsche Bibliothek, Frankfurt, Leipzig and Berlin.

BENEFITS The benefits of NEDLIB were described by the project partners as follows: It provides a forum for the exchange of best practices in developing digital deposit systems. It serves the purposes of consensus building and spreading research costs. It acts at an intermediary level between global initiatives in the field of digital preservation and local efforts from project participants. It directs those efforts toward converging solutions and thereby contributes to an emerging infrastructure for digital deposit libraries. For national libraries worldwide, NEDLIB delivers guidelines and a toolbox for local implementation of deposit systems.

Open Archival Information System (OAIS) Standard

In 1995, Panel 2 of the Consultative Committee on Space Data (CCSDS) was asked by the International Standards Organization (ISO) to coordinate the development of standards to support the long-term preservation of digital information obtained from observations of the terrestrial and space environments. CCSDS began by developing a "Reference Model" to establish common terms and concepts for long-term digital preservation. Although rooted in the space and earth observation communities, from a very early stage other communities, including the National Archives and Records Administration (NARA) in the United States, became involved in the development of this model. This involvement has grown as other initiatives became aware of the draft standard and contributed to its development. In 2001, the draft Reference model (CCSDS 2001) was submitted for adoption as a formal ISO standard and is expected to be formally adopted in 2002.

The reference model sets out to:

- provide a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access;
- provide the concepts needed by nonarchival organizations to be effective participants in the preservation process;
- provide a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives;
- provide a framework for describing and comparing different long-term preservation strategies and techniques;
- provide a basis for comparing the data models of digital information preserved by archives and for discussing how data models and the underlying information may change over time;
- provide a foundation that may be expanded by other efforts to cover long-term preservation of information that is *not* in digital form (e.g., physical media and physical samples);
- expand consensus on the elements and processes for long-term digital information preservation and access, and promote a larger market that vendors can support;
- guide the identification and production of OAIS-related standards.

The model has been developed in a series of international workshops, augmented with e-mail exchanges and occasional teleconferences. National workshops in the United Kingdom, the United States, and France have taken place between the international meetings. The national workshops have been focused on developing national positions and input for the international efforts. The development of the reference model can be seen by surveying the reports and papers from past U.S., French, British, and international workshops.

Adoption and Implementation of the OAIS Reference Model

Development of the draft OAIS reference model has been an open process, with drafts available online. Despite the process of standards development and approval being a protracted one, this openness has allowed the draft model to be reviewed, critiqued, and adapted by a wide range of organizations. It now has wide acceptance and influence. Sectors and initiatives that have adopted the model as a basis for their digital preservation efforts include:

- deposit libraries, e.g., the British Library and the KB—the Dutch National Library; are specifying conformance with OAIS in their system development;
- national archives, e.g., the National Archives and Records Administration;
- scientific data centers, e.g., the U.S. National Space Science Data Center;
- commercial organizations, e.g., the U.S. Aerospace Industries Association;
- NEDLIB project;
- CEDARS research project in the United Kingdom;
- SIPAD: the French space agency plasma physics archive;
- RLG/OCLC Preservation Metadata Working Group;
- RLG/OCLC Working Group on Attributes of Trusted Digital Repositories.

Future Developments

With the growing maturity and acceptance of the draft OAIS standard, attention has turned to identifying and starting additional archival standardization efforts. This is reflected in the Digital Archive Directions (DADs) Workshop held in 1998 and the Archival Workshop on Ingest, Identification, and Certification Standards (AWIICS) held in 1999.

The DADS workshop identified the three most urgent areas requiring additional work as being Ingest, Identification and Certification of archives. The October 1999 Archival Workshop on Identification, Ingest and Certification (AWIICS) explored these three areas in greater detail. Further work is ongoing within CCSDS panel 2 on Ingest led by CNES in France and Archive Certification led by NARA in the United States.

There is also increasing interest among implementers of the standard in sharing experiences of implementation. In this context, it is interesting to note that the Research Libraries Group (RLG) is implementing an Open Archival Information Systems (OAIS) Resources Web Site and mailing list as part of the RLG Long-term Retention Initiative.

Achievements and Constraints

Considerable intellectual effort has gone into development of the reference model over the past seven years. It has also been an open process that has benefited from input from many sectors. It provides a common language and concepts for different professional groups involved in digital preservation and developing archiving systems. The outcome has been a reference model that has won widespread acceptance as a basis for digital preservation effort in all sectors that have reviewed it.

It is a good example of both the advantages of a formal standards process in terms of the intellectual rigor, developing consensus and utilizing a wide range expertise and experience and the disadvantages in terms of time to reach widespread consensus and delays before it becomes an official standard. The language of a formal standard can be very offputting for the uninitiated, and there can be a need for "vernacular" and accessible versions for a wider audience.

The reference model is a high-level model for describing digital archives. It does not mandate any implementation of the model. As such, the model has to be supplemented with additional standards and guidelines to achieve any implementation of the concepts. However, the OAIS reference model has already proved itself to be a critical foundation internationally for digital preservation efforts and seems likely to be the starting point for most, if not all, future initiatives in the field.

PRESTO (Preservation Technology for European Broadcast Archives)

PRESTO is a 21-month, 4.8 million–euro European Union project to develop broadcast archive preservation technology. The project is led by the BBC (British Broadcasting Corporation), and the two additional partners are INA (Institut national de l'audiovisuel) in France, and RAI (Radiotelevisione Italiana) in Italy. Each partner leads with technology partners on a specific area of audiovisual material in the work packages: RAI for audio, INA for video, and the BBC for film.

Although not focused on digital preservation specifically (it is primarily concerned with the preservation of analog material), the issues being addressed are relevant to the study. Audiovisual material is one of the few areas currently where digitization is considered to be the main option for preservation because the originals are unstable and/or locked into obsolete technology. Resolving digital preservation issues therefore does have a major bearing on the long-term preservation of these materials.

As noted by the project, broadcasting technology was never developed as a mechanism to create and hold permanent audiovisual history. The content of European public service broadcast archives is the social and cultural history of 20th century Europe, and a major part of this material is now at risk.

PRESTO consists of two major components: a survey of broadcast archives and developing new technology for reducing preservation project costs.

The Survey

A detailed survey was conducted of the archives of the three partners and other national broadcast archives in the user group. The purposes of the survey were to establish the scale of the problem, identify the solutions required, and also help individual archives construct a business case for investment in preservation.

Key findings from this survey (Wright 2001) were as follows:

- Some 75 percent of the holdings surveyed are now at risk or inaccessible.
- Collections are growing at roughly four times the rate of current preservation work.

- An estimated 10 million hours of broadcast material of national and European significance are at risk.
- The cost of preserving such material is between 100 euro per hour for audio and videotapes and 2,000 euro per hour for film.
- The total cost of preserving this material using present methods and technology is well over 1 billion euros.
- Unless new, more cost-effective preservation methods and technology can be found, the preservation price may simply be too high and significant portions of the audiovisual memory of the 20th century will be lost.
- Digitization and mass storage is about 50 percent more expensive than copying to other formats, but is expected to double the usage of an asset.
- The aim of preservation work is to retain for the future, as cost effectively as possible, that portion of existing broadcast archives that will contribute most to future usage.
- The conclusion from current overall archive usage figures is that the value of an item must be more than four times the preservation cost in order to be financially justified on a commercial basis.
- For most broadcast archive material, this condition can easily be met because one minute of sold or reused archive material will pay for preservation of one hour of archive material.
- For material that cannot pass the “commercial economics” criterion as outlined above, there should be a safety net of assessment for cultural-historical value and a separate funding mechanism.

Preservation Technology

The final phases of the project consist of a program of technology development to assist mass digitization and preservation activities in the archives. This starts with surveying and documenting current methods of preservation work; documenting the factors of time, cost, and quality, and identifying key areas of high cost or time, and areas of low quality; second, surveying the opportunities offered by new technology (e.g., digital mass storage). The same factors of time, cost, and quality are to be specified—but also the new business opportunities and their potential costs and benefits are being documented. Based on the above analysis, the project is then selecting key technology gaps regarding archive preservation and specifying the precise, detailed requirements of the technology. The overall objective of the development phase is to produce new links in the preservation workflow that substantially reduce the cost of archive preservation.

Benefits

The survey has been completed and already has demonstrated its value in quantifying the scale of the challenges faced by the broadcast archives, identifying cost elements of preservation, and potential benefits of investment. The collection of the information was laborious, but the sharing of information on costs and potential savings is seen as immensely valuable.

The technology development is aimed at establishing "preservation factories" with throughput on a massive scale. Any bottlenecks that slow down throughput are being identified and opportunities for automation and developing new tools explored. It is too early to say how successful this part of the program will be at this stage as the work is still in progress.

It should be noted that audiovisual archives with very heterogeneous collections may have less scope for mass preservation processes, but it is believed this approach will be essential for broadcast archives. It was also noted that cost models are a major and complex issue. Accounting practices may be critical to the process used. If you have few technical staff, it may be easier to fit preservation work into small-scale activity as part of existing programs and absorb costs into ongoing staff budgets rather than establishing specific preservation programs. Where activity-costed accounting practices are applied, this will not be the case.

References

Beagrie, N. and Greenstein, D., 1998. *A Strategic Policy Framework for Creating and Preserving Digital Collections.* Version 4.0 (Final Draft). ELib Supporting Study p3. Library Information and Technology Centre, South Bank University, London.

Bide, M. & Associates, 2000. *Standards for Electronic Publishing: an overview.* NEDLIB Report series 3. Koninklijke Bibliotheek, The Hague. ISBN 90-62-59-1477

Borbinha, J.L. and Cardoso, F., 2000. *NEDLIB Local Situations.* Available from: *http://www.kb.nl/coop/nedlib/results/local_situations_v2.htm*

Clavel-Merrin, G., 2000. *List of NEDLIB Terms.* NEDLIB Report series 7. Koninklijke Bibliotheek, The Hague. ISBN 90-62-59-1515

Consultative Committee for Space Data Systems (CCSDS), 2001. *Reference Model for an Open Archival Information System (OAIS).* Red Book. Issue 2. June 2001. Available from: *http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html*

Feenstra, B., 2000. *Standards for the Implementation of a Deposit System for Electronic Publications (DSEP).* NEDLIB Report series 4. Koninklijke Bibliotheek, The Hague. ISBN 90-62-59-1485

Hakala, J., 2001. "Collecting and Preserving the Web: Developing and Testing the NEDLIB Harvester", *RLG Diginews* 15 April 2001 Volume 5 Issue 2 available from: *http://www.rlg.org/preserv/diginews/diginews5-2.html#feature2*

Jones, M. and Beagrie, N., 2001. *Preservation Management of Digital Materials: A Handbook.* British Library, London. ISBN 0-7123-0886-5

Lupovici, C. and Masanès, J., 2000. *Metadata for Long Term Preservation.* NEDLIB Report series 2. Koninklijke Bibliotheek, The Hague. ISBN 90-62-59-1469

Rothenburg, J. and Bikson, T., 1999. *Digital Preservation: Carrying Authentic, Understandable and Usable Digital Records Through Time.* Report to the Dutch National Archives and Ministry of Interior by Rand-Europe, The Hague.

Rothenberg, J., 2000. *An Experiment in Using Emulation to Preserve Digital Publications.* NEDLIB Report series 1. Koninklijke Bibliotheek, The Hague. ISBN 90-62-59-1442

Steenbakkers, J., 2000. *The Nedlib Guidelines: Setting up a Deposit System for Electronic Publications.* NEDLIB Report series 5. Koninklijke Bibliotheek, The Hague. ISBN 90-62-59-1493

van de Werf, T., 2000. *A Process Model : The Deposit System for Electronic Publications.* NEDLIB Report series 6. Koninklijke Bibliotheek, The Hague. ISBN 90-62-59-1507

Wright, R., 2001. *Broadcast Archives: Preserving the Future.* Available from: *http://presto.joanneum.ac.at/projects.asp#Pres*

Web Sites Consulted for the Study

Australia

National Archives of Australia
http://www.naa.gov.au

Council of Australian University Libraries
http://www.anu.edu.au/caul/

Digital Continuity Conference November 2001
http://www.swin.edu.au/lib/DigCon2001.htm

National Library of Australia
http://www.nla.gov.au/

Our Digital Island (Tasmania State Library)
http://odi.statelibrary.tas.gov.au/

PADI–Preserving Access to Digital Information
http://www.nla.gov.au/padi/

PANDORA
http://pandora.nla.gov.au/

ScreenSound Australia
http://www.screensound.gov.au/

France

Archives de France
http://www.archivesdefrance.culture.gouv.fr/

Bibliothèque nationale de France
http://www.bnf.fr

Couperin
http://buweb.univ-angers.fr/COUPERIN.html

Institut national de l'audiovisuel (INA)
http://www./ina.fr

PIN
http://sads.cnes.fr:8010/pin/welcome.html

Netherlands

ARNO
http://www.uba.uva.nl/en/projects/arno/

Digitale Duurzaamheid
http://www.digitaleduurzaamheid.nl

Koninklijke Bibliotheek
http://www.kb.nl
http://www.kb.nl/kb/resources/frameset_kenniscentrum.html

NIWI
www.niwi.knaw.nl

Rijksarchiefdienst
http://www.archief.nl

Roquade
http://www.roquade.nl

Test Bed Digitale Bewaring
http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=181&categorie=2

United Kingdom

Arts and Humanities Data Service
http://ahds.ac.uk/

British Library
http://www.bl.uk

CAMiLEON Project
http://www.si.umich.edu/CAMILEON

CEDARS Project
http://www.leeds.ac.uk/cedars/

Digital Preservation Coalition
http://www.jisc.ac.uk/dner/preservation/prescoalition.html

JISC Digital Preservation Focus
http://www.jisc.ac.uk/dner/preservation/

National Preservation Office
http://www.bl.uk/services/preservation/national.html

Public Records Office
http://www.pro.gov.uk

UK Research Councils–e-Science Programme Website
http://www.research-councils.ac.uk/escience/membership.shtml

Multinational Initiatives
ERPANET: Electronic Resource Preservation and Access Network
http://www.erpanet.org/

NEDLIB
http://www.kb.nl/coop/nedlib/homeflash.html

Nordic Web Archive
http://nwa.nb.no/

OAIS
Consultative Committee for Space Data Systems
http://www.ccsds.org/

Digital Curation: digital archives, libraries, and e-science
http://www.jisc.ac.uk/dner/preservation/digitalarchives.html

ISO Archiving Standards Overview
http://ssdoo.gsfc.nasa.gov/nost/isoas/

PRESTO
http://presto.joanneum.ac.at/index.asp

All URLs cited were correct as of March 31, 2002, when last accessed.

APPENDIX 6

Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment

Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment

JUNE M. BESEK
Director of Studies, Kernochan Center for Law, Media and the Arts
Columbia Law School

1.0 Introduction

Collection and long term preservation of digital content pose challenges to the intellectual property regime within which libraries and archives are accustomed to working. How to achieve an appropriate balance between copyright owners and users is a topic of ongoing debate in legal and policy circles. This paper describes copyright rights and exceptions and highlights issues potentially involved in the creation of a nonprofit digital archive.[1] It is necessarily very general, since many decisions concerning the proposed archive's scope and operation have not yet been made. The purpose of an archive (e.g., to ensure preservation, or to provide an easy and convenient means of access), its subject matter, the manner in which it will acquire copies, and who will have access to the archive, from where, and under what conditions, are all factors critical to determining the copyright implications for works to be included.[2] The goal of this paper is to provide basic information about the copyright law for those developing such an archive so that they will be able to recognize areas in which it could impinge on copyright rights, and plan accordingly. When further decisions have been made, a more detailed and refined analysis will be possible. As noted below, there are a number of areas that would benefit from further research. Such research may not yield definitive legal answers, but could narrow the issues and suggest strategies for proceeding.

Copyright © 2002 June M. Besek, Kernochan Center for Law, Media and the Arts, Columbia Law School.

2.0 Copyright Subject Matter

"Copyright" exists in any original work of authorship fixed in a tangible medium.[3] That medium can be almost anything, including paper, computer disk, clay, canvas, and so on. For a work to be "original," it must meet two qualifications: it cannot be copied from another work, and it must exhibit at least a small amount of creativity. Copyright lasts for the life of the author and seventy years thereafter.[4]

3.0 Copyright Rights

A copyright provides not just a single right but a bundle of rights that can be exploited or licensed separately or together. The economic rights embraced within a copyright include:

- The reproduction right (the right to make copies). For purposes of the reproduction right, a "copy" of a work can be any form in which the work is fixed and from which it can be perceived, reproduced or communicated, either directly or with the aid of a machine.[5] Courts have held that even the reproduction created in the short-term memory (RAM) of a computer when a program is loaded for use qualifies as a copy.[6]
- The right to create adaptations, or derivative works. A "derivative work" is a work that is based on a copyrighted work, but contains new material that is "original" in the copyright sense. For example, the movie "Gone With the Wind" is a derivative work of the book by Margaret Mitchell. "Version" is not a term of art in copyright law. If a new version consists merely of the same work in a new form—such as when a book or photograph is scanned to create a digital version—then it is a reproduction of the work. However, if new copyrightable authorship is added, then it is a derivative work. For example, Windows 2000 is a derivative work based on Windows 98.
- The right to distribute copies of the work to the public. The distribution right is limited by the "first sale doctrine," which provides that the owner of a particular copy of a copyrighted work may sell or transfer that copy. In other words, the copyright owner, after the first sale of a copy, cannot control the subsequent disposition of that copy.[7] Making copies of a work available for public downloading over an electronic network qualifies as a public distribution.[8] However, so far neither the courts nor the Copyright Office have endorsed a "digital first sale doctrine" to allow users to retransmit digital copies over the Internet.[9]
- The right to perform the work publicly. To perform a work means to recite, render, play, dance or act it, with or without the aid of a machine.[10] Thus, a live concert is a performance of a musical composition, and so too is playing a CD on which the composition is recorded.
- The right to display the work publicly.

"Publicly" is a broad concept. To perform or display a work publicly means to perform or display it anywhere that is open to the public or anywhere that a "substantial number of persons outside of a normal circle of a family and its social acquaintances is gathered."[11] Transmitting the performance or display to such a place also makes it public. It does not matter if members of the public receive the performance at the same time or different times, at the same place or different places. Making a work available to be received or viewed by the public over an electronic network is a public performance or display of the work.[12]

There is a distinction between ownership of a copy of a work (even the original copy, if there is only one) and ownership of the copyright rights. A museum does not, by acquiring a painting, automatically acquire the right to reproduce it. Libraries and archives commonly receive donations of manuscripts or letters, but they generally own only the physical copies and not the copyright rights.[13]

Not all rights attach to all works. For example, some works, such as sculpture, are not capable of being performed. Other works—notably musical compositions and sound recordings of musical compositions—have rights that are limited in certain respects. For example, reproduction of musical compositions in copies of sound recordings[14] is governed by a compulsory license which sets the rate at which the copyright owner must be paid.[15] Sound recordings, for historical reasons, long had no right of public performance, and now enjoy only a limited performance right in the case of digital audio transmissions.[16]

Even though works can be converted into mere 1's and 0's when digitized, they generally retain their fundamental character. In other words, if the digitized work is a computer program, it is subject to the privilege the law provides to owners of copies of computer programs to make archival copies. If it is an unpublished work, it retains the level of protection that attaches to unpublished works (discussed in sections 4.0 and 8.0, below).

4.0 Relevant Copyright Exceptions

Copyright rights are not absolute, and are subject to a number of limiting principles and exceptions. Those most relevant to the creation of a digital archive are:

(1) The exception for certain archival and other copying by libraries and archives in section 108 of the Copyright Act. Libraries and archives are permitted to make up to three copies of an unpublished copyrighted work "solely for purposes of preservation and security or for deposit for research use in another library or archives."[17] The work must be currently in the collections of the library or archives, and any copy made in digital format may not be made available to the public in that format outside the library premises.

Libraries and archives may also make up to three copies of a published work to replace a work in their collections that is damaged, deteriorating, lost or whose format has become obsolete, if the library determines that an unused replacement cannot be obtained at a fair price. As with copies of unpublished works, copies in digital format may not be made available to the public outside the library premises.[18]

Even if copying a work is not expressly allowed by section 108, it may still be permitted under the fair use doctrine. However, the privileges under section 108 do not supersede any contractual obligations a library may have with respect to a work that it wishes to copy.[19]

(2) Fair use is the copyright exception with which people are often most familiar. Whether a use is fair depends on the facts of a particular case. There are four factors that must be evaluated. The first is the purpose and character of the use. Among the considerations is whether the use is commercial or for nonprofit educational purposes. Works that transform the original by adding new creative authorship are more likely to be considered fair use, but a use can be fair even if it is merely a reproduction. The second factor is the nature of the copyrighted work. The scope of fair use is generally broader for fact-based works than it is for fanciful works, and broader for published works than for unpublished ones.[20] The third fair use factor is the amount and substantiality of the portion used. Generally the more that is taken, the less likely it is to be fair use, but there are situations in which making complete copies is considered fair.[21] The fourth factor is the effect on the potential market for or value of the copyrighted work. A use that usurps the market for the original is unlikely to qualify as fair use.

Certain uses are favored in the statute: criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship and research. A nonprofit digital archive for scholarly or research use would be the kind of use favored by the law. However, favored uses are not automatically deemed fair, nor are other uses automatically deemed unfair. The factors discussed above must be applied and evaluated in each case. It can be a source of frustration to some users that there is no magic formula to determine whether a use is fair. However, the same flexibility that sometimes makes it difficult to predict whether a use will be considered fair also allows the statute to evolve through case law with new circumstances and new types of uses. A statute that provides greater certainty would inevitably be more rigid.

(3) Section 117 allows the owner of a copy of a computer program to make an archival copy of that program.[22] However, section 117 does not apply to all works in digital form, but only to computer programs.[23]

(4) The first sale doctrine. As discussed above in section 3.0, the first sale doctrine prevents the copyright owner from controlling the disposition of a particular copy of a work after the initial sale or transfer of that copy. The first sale doctrine enables, for example, library lending and markets in used books.

5.0 Copyright Requirements

Two processes tend to be confused by non-specialists: registration of copyright and mandatory deposit of copyright-protected works (discussed in the next section). A copyright owner is not required to register her copyright or to use a copyright notice in order to establish or maintain copyright in a work. This fact is often misunderstood, particularly in the Internet context where people sometimes assume that if there is no copyright notice, a work is in the public domain. A copyright owner is required to register her copyright before filing an infringement suit, if the work is of U.S. origin. There are incentives in the law to motivate copyright owners to file a timely registration. However, many copyright owners choose not to register for a variety of reasons, and it is a mistake to assume that the Copyright Office has a record of all copyright-protected works.

6.0 Mandatory Deposit

Copyright owners are required to deposit two copies of the "best edition" of any work published in the United States, within three months of publication, with the Copyright Office for the benefit of the Library of Congress ("LC").[24] Even if the copyright owner does not register the copyright in her work, she must comply with the deposit requirement. Failure to do so does not affect the status of the copyright, but it can result in fines.[25] LC may also demand copies of specific "transmission programs," even though they are technically unpublished, or make a copy itself from the transmission.[26] A transmission program is "a body of material that, as an aggregate, has been produced for the sole purpose of transmission to the public in sequence and as a unit."[27]

LC is entitled to keep the deposit copies of published works for its collections, or use them "for exchange or transfer to any other library."[28] LC may also keep the deposit copies of unpublished works for its collections, or may transfer them to the National Archives or a federal records center.[29] The rights that LC has with respect to deposited works pertain to the physical copies, not to the underlying rights. (For example, LC may not, merely by virtue of its receipt of deposit copies of motion pictures or musical works, authorize public performances of those works.) The statute expressly permits the Copyright Office to make a facsimile reproduction of deposit material before transferring it to LC or otherwise disposing of it,[30] but otherwise there is no license to exercise any other rights with respect to the works. It is reasonable to interpret the law to permit LC to use deposit copies of works such as computer programs or CD-ROMs on a stand-alone computer, just as any other individual user could, even though the computer technically makes a copy when it runs or plays the work. But use on a network would implicate not only the reproduction right but also the rights to publicly perform, display or distribute (depending on the work). There is nothing currently in the law that would permit LC to make deposit copies generally available in digital form on a publicly accessible network.[31]

Some works—large databases, for example—are no longer distributed in complete copies in a portable medium like a book or CD-ROM. Instead, the end user licenses access to the database via the Internet, and generally downloads and prints only the portion of the database relevant to her research. Application of the mandatory deposit provisions to works distributed in this manner, and to websites generally, is far from clear. For example:

- To what extent can such works be considered published, if not all of the work is available for downloading in copies?
- What if material is available to a limited group, with restrictions, and thus constitutes only a "limited publication" technically considered unpublished under copyright law?[32]
- If materials available online are unpublished, to what extent can they be considered "transmission programs" that LC may copy or demand?[33]
- How can the deposit copy of a website be defined, when website boundaries are so amorphous?
- If the work is distributed only with technological security measures, can LC demand it in a different form?
- What is the legal effect of the license agreements that frequently accompany works available online? Can LC reasonably take the position that it is not bound by them? Does it matter whether the copyright owner disseminates copies of the complete work, or merely licenses the right to access it online?
- Should all works that can be downloaded from the Internet in the United States be considered "published" here for purposes of mandatory deposit? This position would substantially broaden mandatory deposit for non-U.S. works.

Even where LC has a clear right to demand copies, in the past it has been sensitive to copyright owners' legitimate concerns about the use of those copies, and presumably would continue to be so. This raises the following additional questions:

- Under what circumstances is it reasonable to request deposit copies of works published online, and with what frequency?
- How can LC's needs be met without imposing serious hardship or risk on copyright owners?
- Regardless of whether LC is bound by license agreements associated with deposit copies (an issue this paper does not address), are there terms and conditions that reflect valid security or other concerns that should nevertheless be taken into account?

There are no clear answers to these questions, and little precedent. This is an area that would benefit from further study.

7.0 Copyright Ownership

Usually the human creator of a work is the author and initial owner of copyright.[34]

Copyright rights can be transferred, either separately or together. For example, someone can transfer the right to reproduce a work without transferring the right to create a derivative work. A transfer of copyright ownership, including the grant of an exclusive license, must be in writing and signed by the grantor.[35] Nonexclusive licenses need not be in writing, but frequently are.

A copyright license can span a very long period of time. Complicated issues can arise when new forms of exploitation are developed during the license term. Usually the grantor will claim she did not intend to include the new rights in the license, and the grantee will claim the opposite. For example, *Random House, Inc. v. Rosetta Books LLC*[36] is an ongoing case concerning whether the words "in book form" in publishing contracts entered into before the advent of electronic publishing cover electronic book rights. The authors contended that electronic book rights were not covered by their existing publishing agreements with Random House, and entered into new agreements with Rosetta to publish their books in electronic form. Recently a federal court in New York agreed, and refused to enter the preliminary injunction sought by Random House to stop Rosetta from publishing the electronic books. Decisions in these "new use" cases usually hinge on the wording of the contract and industry practices at the time it was entered.[37]

Another debate about electronic rights was resolved last year in *New York Times Co. v. Tasini.*[38] The Supreme Court held that the New York Times, in licensing back issues of the newspaper for inclusion in electronic databases such as Nexis, could not license the works of freelance journalists contained in the newspapers. The Times' contracts with the journalists did not address copyright ownership, so it relied instead on a provision in the Copyright Act that gives limited privileges to owners of collective works, such as journals and newspapers, in respect of individual contributions to those works.[39] According to the Court, the New York Times had the right to publish the freelancers' articles in the original issue of the newspaper in which they first appeared, and in revisions of that newspaper, but the authors—and not the Times—retained the rights to license use in electronic databases. The principle announced in *Tasini* affects many other newspapers, magazines and journals. They may not license the works of freelance journalists for individual access through electronic databases unless they have a contract that permits them to do so.

As these two cases illustrate, ownership of electronic rights can be ambiguous, and sometimes widely dispersed.

How does one track ownership of a copyrighted work? The process can be complicated and sometimes frustrating. Usually the Copyright Office registration and renewal records are a good place to start. Registration and renewal of copyrights used to be mandatory, so registration records are more complete for older works. However, even if the copyright is registered, rights may have changed hands subsequent to registration.[40] It is also possible to obtain information from the copyright notice (no longer mandatory but still commonly used) or other materials associated with the work.

Other records in the Copyright Office may be helpful. For example, the copyright law provides for recordation in the Copyright Office of transfers related to copyright.[41] To perfect a security interest in a copyrighted work or to ensure that the first transferee will prevail over a second transferee of the same interest, a license or assignment must be timely recorded in the Copyright Office.[42] Not all copyright owners record their agreements, however; it is most commonly done for works of significant commercial value.

What does someone do if she wants to use a work and has tried without success to identify and locate the copyright owner? Some users are reluctant to use anything without clear rights, but others will engage in risk assessment. For example, if the work is to be used in a database from which it can be removed promptly if there is a complaint, the user may decide as a business matter that the risk is worth running.[43] However, if the work is a short story that is to be the basis of a new screenplay and motion picture, and the investment could be lost if the copyright owner learned of and objected to the project, she may decide the risk is too great to proceed.

8.0 Unpublished Works

A work is published when copies are distributed to the public by sale or other transfer of ownership, or by rental, lease or lending. Publicly performing or displaying a work does not itself constitute publication.[44] There are a number of distinctions in the law between published and unpublished works. The most significant in this context are the treatment of published and unpublished copies for purposes of preservation under section 108 (discussed in section 4.0, above), and fair use. The scope of fair use is narrower for unpublished works than for published works, although the fact that a work is unpublished does not itself bar fair use. The unpublished nature of the manuscript of President Ford's memoirs was a significant factor in the Supreme Court's decision that *The Nation* was liable for copyright infringement in publishing excerpts of those memoirs (quotations that totalled about 300 words).[45]

9.0 Digital Millennium Copyright Act

The Digital Millennium Copyright Act (DMCA) prohibits the act of circumventing a technological measure that "effectively controls access" to a work protected by copyright.[46] Technological access controls are mechanisms such as passwords or encryption that prevent viewing or listening to the work without authorization.

The law also contains two provisions that prohibit trafficking in devices that circumvent technological measures of protection. The first is aimed at devices and services that circumvent access controls. Specifically, it prohibits manufacturing, importing, offering to the public, providing or otherwise trafficking in technologies, products or services

- that are primarily designed or produced to circumvent a technological measure that effectively controls access to a copyrighted work, or
- that have only limited commercially significant purpose or use other than to circumvent such controls, or
- that are marketed for use in circumventing such controls.[47]

There is a similarly worded prohibition against trafficking in devices or services to circumvent rights controls.[48] Technological rights controls are mechanisms that restrict copying the work or playing it in a particular environment without authorization. *There is no prohibition on the act of circumventing rights controls.* Legislators believed if copies made as a consequence of circumventing rights controls were excused by copyright exceptions or privileges, there should be no liability for the circumvention. If, on the other hand, such copies are infringing, the rightholder has a claim under the copyright law.

There are a number of exceptions to the ban on circumventing access controls, and a few exceptions to the anti-trafficking ban. There is no exception for archiving, nor is there a general "fair use" type exception written into the statute.[49] The law does, however, include an administrative procedure for creating new exceptions. Every three years the Librarian of Congress, upon the recommendation of the Copyright Office, is directed to determine through a rulemaking proceeding whether users of any particular class of copyrighted works are, or are likely to be, adversely affected in their ability to make noninfringing uses of those works by the prohibition against circumventing technological access controls. If so, he is to lift the prohibition on circumventing access controls for that particular class of works for the ensuing three-year period.[50]

The DMCA could potentially affect archiving in a couple of ways. First, the law would prohibit an archive from circumventing technological access controls to obtain access to copyrighted works. However, should a situation arise in which that archive has legally defensible reasons for seeking to archive materials to which it has no authorized access, it could seek an exception pursuant to the rulemaking procedure discussed above.

The second potential problem is the DMCA's ban on the circulation of circumvention devices. Even where a library or archive has valid access to a work, that work may be protected by a copy control. Circumventing the copy control will not violate the DMCA (its permissibility would be judged separately under the Copyright Act), but a library or archive may not have the means readily available to make that copy because of the anti-trafficking provision. It is possible that a digital archive could develop the expertise to circumvent technological controls where necessary. Moreover, it may also be possible to engage expert assistance: the law would appear to allow someone to offer circumvention services whose primary purpose and effect would be to facilitate permissible library archiving. The implications of the DMCA for archiving activities is an area that warrants further study.

10.0 International Issues

There are at least three categories of international issues to consider in planning a digital archive. First, international treaties place certain constraints on the United States' ability to create exceptions to copyright protection, or to impose requirements on copyright owners. Second, there are legal and logistical uncertainties that can make it difficult for a copyright owner to obtain redress for copyright infringements committed abroad. These uncertainties should be considered in deciding which works should be included in the digital archive and from where they will be accessible. Third, a digital archive that permits online access outside the United States could itself be vulnerable to suit by foreign copyright owners whose works are included.

10.1 Limitations of Copyright Treaties

Through a series of copyright treaties with other countries, United States nationals have the benefit of copyright laws in many foreign countries, and nationals of many foreign countries have the benefit of U.S. laws. The principal international copyright treaty is the Berne Convention for the Protection of Literary and Artistic Works.[51] In 1996 a new international copyright treaty was negotiated under the auspices of the World Intellectual Property Organization (WIPO). Known as the WIPO Copyright Treaty, it addresses issues raised by new technologies.[52] Many countries are in the process of amending their laws to comply with the treaty. To date more than thirty countries have joined.[53]

These treaties generally provide for (1) national treatment, and (2) minimum standards of protection. National treatment means that when a U.S. citizen sues in another country—Germany, for example—she will be treated as a German citizen, with the benefit of German laws. Those laws will likely be similar to U.S. laws in many respects, due to the minimum standards imposed by the treaties. However, there are still likely to be differences, especially in areas related to new technologies, where international treaties and national laws sometimes have a difficult time keeping pace with technological developments.

There are many standards for copyright protection imposed on treaty members. The principal ones that could be implicated by a digital archive are the prohibition on "formalities," the limitation on exceptions to copyright rights, and the prohibition on compulsory licensing.

Article 5(2) of the Berne Convention provides that the "enjoyment and exercise" of copyright rights "shall not be subject to any formality." Prohibited formalities include such things as mandatory copyright notice or registration. Mandatory deposit is permitted provided it is not a condition of copyright protection.

Article 9(2) of the Berne Convention provides that countries may allow for exceptions to the author's exclusive right of reproduction "in certain special cases, provided that such reproduction does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author." The WIPO Copyright Treaty extends this limitation to all rights provided by that treaty or by the Berne Convention, not just the reproduction right.[54]

Compulsory licenses "obviously run counter to the whole basis of the [Berne] Convention, which is that the rights conferred under it are the author's exclusive rights which he can dispose of as he wishes."[55] A compulsory license reduces the author's freedom to license (or not) to a mere right of remuneration. The Berne Convention expressly recognizes compulsory licenses only in two cases: broadcasting, and recordings of musical compositions.[56]

It is certainly possible to create a digital archive without violating any U.S. treaty obligations. However, it could not be premised on a requirement for deposit or notice linked to copyright protection. Nor could it be premised on an exception to copyright rights that would jeopardize the normal exploitation of a work or harm the author's legitimate interests, or subject works to a broad compulsory license.[57]

10.2 Potential Difficulties in Obtaining Redress for Infringements Abroad

It is assumed, for purposes of discussing this point and the next, that the archive would be located in the United States but accessible online in other countries, and that it would be possible to download works and reproduce them there without authorization. Such an archive could increase copyright owners' exposure to economic harm from infringement.[58] The logistics of bringing a suit based on an infringement that takes place in another country can be daunting. First, it is difficult and costly to sue in another country. Second, as discussed above, even though national treatment is the rule, there may be significant differences in national copyright laws, particularly in areas of new technology. Third, even if the copyright owner were able to obtain personal jurisdiction over the defendant in the United States (which is likely to be difficult), a U.S. court may be reluctant to adjudicate a case involving an interpretation of a foreign country's laws.[59]

Moreover, there are still many unsettled areas. For example, if the archive is limited to authorized users by means of technological access controls, has a user in another country who circumvents those controls to gain access violated any law? Not all countries have laws protecting such measures from circumvention (either by a ban on circumventing, on trafficking, or both), and those that have use different approaches. Can a user in another country be held to an online agreement that restricts use of the archive? Laws on electronic contracts are still developing.

10.3 The Archive's Potential Exposure to Suits Abroad

Finally, the archive itself could be exposed to infringement suits if it were accessible outside the United States. Foreign copyright owners whose works were included in the archive might sue if their works are made accessible in countries where such use is infringing. A court outside the United States could apply the laws of a country (its own or a third country) that regards placing a copyrighted work on a publicly accessible network without authorization to be an infringement.

The international issues are very complicated, and worthy of more detailed study if the archive is to be accessible from outside the United States.

11.0 Summary and Conclusion

Below is a brief summary of the ways in which the archive might acquire a work and the copyright and contract constraints on each.

- Copies received through mandatory deposit. LC receives copies under the mandatory deposit provisions of the Copyright Act. Copies, including digital copies, can be made pursuant to section 108, but the circumstances under which they can be made and used are restricted, as discussed above. Placing a digital copy (whether made by LC or received in that form) on a publicly accessible network can violate a copyright owner's rights.[60] (A network accessible only from a limited number of locations can be "public" for these purposes.)

- Copies obtained by gift or purchase. Copies of works that are purchased raise the same issues as copies received through mandatory deposit. This is also true of copies received by gift, unless the gift embraces not just the physical copies but also corresponding rights.
- Copies obtained through subscription or license. Copies of works obtained through subscription or license may be subject to additional requirements of a subscription or license agreement, which may restrict use of the work beyond what the copyright law would allow.
- Copies made or received under agreements with copyright owners. Some copyright owners may simply be willing to allow their works to be included in a digital archive. Others may agree if they get something in return, e.g., more favorable treatment in the registration process such as "group registration." Many copyright owners would want to ensure that there are appropriate security measures, and limitations on use such as restrictions on where the works can be accessed, limitations on downloading, or user agreements. LC has in the past entered into agreements with copyright owners to place deposit copies on a local area network.[61]

What about copying, or "harvesting," publicly available websites? There is no specific exception in the law for this type of copying, and its permissibility would likely depend on whether it qualified as fair use. That determination would have to be made on a case-by-case basis, based on factors such as the nature of the material copied, the scope of the copying, who would have access, how the archival use could affect the copyright owner's market, and so on. LC's ability to obtain website material under the mandatory deposit provisions is considered above in section 6.0.

As this list illustrates, there is no clear road under existing law for collecting the works proposed for a digital archive and placing them on a publicly accessible network.[62] A more detailed assessment of the copyright implications of a digital archive requires further information about how the archive would operate and what it would include.

Finally, as noted throughout this paper, there are areas that would benefit from more detailed study. Additional research will not necessarily yield clear legal answers, since many of the uncertainties come from applying laws to technologies and methods of distribution they were not designed to address. Such studies could, however, narrow the issues and suggest constructive ways to achieve the goal of creating and operating an archive to ensure long term preservation of works in digital form for the benefit of society.

June 20, 2002

Endnotes

1. I have assumed that the archive will be created by or in cooperation with the Library of Congress ("LC").

2. It is my understanding that six types of works are currently contemplated for inclusion (although the list may expand as the effort progresses): e-books, e-journals, websites, digital motion pictures, digital television, and digital sound recordings. However, it appears no decision has yet been made on whether the archive will attempt to include all works in these categories or a subset of them, or on the related question whether participation will be voluntary or mandatory.

Background information provided to me suggested that the archive could include published and unpublished materials. For purposes of this exercise, I have assumed that those materials on publicly accessible websites available for downloading are published.

3. Copyright law is contained in Title 17 of the United States Code. All statutory references herein are to sections of Title 17, unless otherwise noted.

4. §302(a). Certain categories of works, e.g., works first published prior to Jan. 1, 1978 (the effective date of the current Copyright Act), and works made for hire, which are discussed below, have different terms of protection. §§304, 302(c); see also §303.

5. §101.

6. E.g., *MAI Systems Corp. v. Peak*, 991 F.2d 511 (9th Cir. 1993), cert. dismissed, 114 S. Ct. 671 (1994) . In a recent report to Congress, the Copyright Office observed: "Every court that has addressed the issue of reproductions in volatile RAM has expressly or impliedly found such reproductions to be copies within the scope of the reproduction right." U.S. Copyright Office, *DMCA Section 104 Report* 118 (August 2001) (available on the Copyright Office website at *http://lcweb.loc.gov/copyright/*).

7. §109(a). There are exceptions for computer programs and sound recordings, designed to deter the development of a commercial rental market.

8. See, e.g., *Playboy Enters., Inc. v. Webbworld, Inc.*, 991 F. Supp. 543 (N.D. Texas 1997), aff'd without opinion, 168 F.3d 486 (5th Cir. 1999); see Robert A. Gorman & Jane C. Ginsburg, *Copyright* 549-52 (Foundation Press, 6th ed. 2002).

9. In its recent *DMCA Section 104 Report*, supra note 6, the Copyright Office rejected the argument that receipt of a copy by digital transmission should be treated the same as receipt of a physical copy, with the recipient free to dispose of the digital copy at will. Digital transmission involves making a copy, not merely transferring a copy. The report expressed concern that application of the first sale doctrine would require deleting the sender's copy when it was sent to the recipient, a feature not generally available on software currently in use and unlikely to be done on a systematic basis by users. The Office also rejected the assumption that forward-and-delete is completely analogous to transferring a physical copy, because delivery and return of a digital copy can be done almost instantaneously, so fewer copies can satisfy the same demand. Id. at 96-101.

10. §101.

11. Id.

12. E.g., *Kelly v. Arriba Soft Corp.*, 280 F.3d 934 (9th Cir. 2002); *Playboy Enters., Inc. v. Frena*, 839 F. Supp. 1552 (M.D. Fla. 1993).

13. The donor frequently does not own the rights and therefore cannot convey them. For example, the copyright in letters is owned by the writer, not the recipient, though the recipient owns the physical copies. Even when the donor owns the rights, they are transferred to the library or archives only if the gift includes a license or assignment.

14. Technically, copies of sound recordings are referred to as “phonorecords” under the Copyright Act. §101.

15. §115.

16. §106(6), §114.

17. §108(b). There are other conditions as well to the library privileges under section 108. For example, the reproduction may not be for commercial advantage; the library must be open to the public, or at least to researchers in a specialized field; and the library must include a copyright notice or legend on copies.

18. §108(c). There are other privileges granted to libraries in section 108, subject to certain conditions. They may reproduce articles and short excerpts at the request of users, and they may reproduce out of print works at users’ request if those works cannot be obtained at a fair price. §108(d), (e). However, they may not engage in systematic reproduction and distribution of copies. Libraries may enter into interlibrary arrangements provided the copies they receive under the arrangement do not substitute for a purchase or subscription. §108(g). Libraries and archives have broad privileges to copy and use many types of published works during the last twenty years of their copyright term for preservation and scholarship purposes, if the works are no longer being commercially exploited and cannot be obtained at a reasonable price. §108(h).

19. §108(f)(4).

20. Copyright law has no “public figure” exception: this is a libel law concept. Nor is there any special exception to permit copying of highly important or newsworthy works. As the Supreme Court stated in *Harper & Row, Pubs. v. Nation Enterprises,* 471 U.S. 539, 559 (1985): “It is fundamentally at odds with the scheme of copyright to accord lesser rights in those works that are of greatest importance to the public.”

21. For example, in *Sony Corp. v. Universal City Studios, Inc.*, 464 U.S. 417 (1984)—commonly referred to as “the betamax case”—the Supreme Court held that private in-home copying of free television programs for time-shifting purposes was fair use.

22. A copy or adaptation that is an essential step in using the program in the computer is also permissible, as are copies made in the course of computer maintenance and repair. §117.

23. In its *DMCA Section 104 Report,* supra note 6, the Copyright Office concluded that copies of digital works made in the course of periodic back-ups of computer hard drives likely qualified as fair use, but recommended a statutory change to make clear that such copies may be used exclusively for archival purposes and not for distribution. Id. at 153-61.

24. §407. The “best edition” is the edition published in the United States that LC deems most suitable for its purposes. §101. What constitutes publication will be considered further below and in section 8.0.

25. §407(d). Certain types of works are exempt from the deposit requirement in whole or in part, either because LC is not interested in acquiring them or because the requirement imposes

a hardship on the copyright owner. For example, three-dimensional sculptural works and works published only as reproduced in or on jewelry, toys, games, wall or floor coverings or other useful articles are exempt from the deposit requirement. 37 C.F.R. §202.19 (c)(6). In the case of motion pictures, only one deposit copy is required, and LC may (and does) enter into agreements to return that copy to the depositor under certain conditions. Id. §202.19 (d)(2)(ii). Copyright owners may also request "special relief" in the event that deposit requirements pose a particular problem for them. §202.19 (e).

26. §407(e).

27. §101.

28. §704(b).

29. Id. Unpublished works are not subject to mandatory deposit (except transmission programs, as noted above), but may be deposited with the Copyright Office as part of a registration application.

30. §704(c).

31. LC does put some deposit copies on a local area network pursuant to agreements with copyright owners. When LC first announced its intention to require deposits of CD-ROMs, copyright owners objected because they feared economic harm might result if their works were readily available through LC for copying and downloading. Their concern was heightened by LC's position that as the owner of the CD-ROMs pursuant to section 704(a), it was not bound by the terms of the associated license agreements. After lengthy negotiations, the parties achieved a compromise under which copyright owners could deposit a single copy under the mandatory deposit provisions, or could opt instead to enter into an agreement with LC either (1) to provide two copies of each CD-ROM for use on a stand-alone computer on LC premises (three copies if they are "copy protected"), or (2) to provide one copy for use on a local area network covering LC premises and a limited number of additional locations in the D.C. area, for use by a limited number (up to five, if the copyright owner agreed) of simultaneous users. Under the agreements, which are rather complex, the copyright owner is required to provide the deposit within 60 days, rather than three months as required by §407. LC, in turn, agrees to undertake various security measures to limit downloading from or transfer of the CD-ROMs.

32. For a discussion of the doctrine of limited publication under the 1976 Copyright Act, see 1 Melville B. Nimmer & David Nimmer, *Nimmer on Copyright* §4.13[B] (LexisNexis 2001).

33. The provisions in the law concerning transmission programs were intended "to provide a basis for the Library of Congress to acquire, as part of the copyright deposit system, copies or recordings of non-syndicated radio and television programs without imposing any hardships on broadcasters." H.R. Rep. No. 1476, 94th Cong., 2d Sess. 152 (1976). A transmission program is "a body of material that, as an aggregate, has been produced for the sole purpose of transmission to the public in sequence and as a unit." §101. This definition is arguably broad enough to encompass some of the materials transmitted over the web. However, the requirement that the body of material be transmitted "in sequence and as a unit" could rule out many websites taken as a whole, where the materials and the sequence in which they are viewed are determined by the user.

34. The one exception is a "work made for hire." Works made for hire are works created by employees in the course of their employment, in which case the employer is deemed by law to be the author, and certain types of commissioned works, provided that the parties agree in writing that the work will be a work made for hire owned by the commissioning party. §§101, 201(a), (b)

35. §204(a).

36. 150 F. Supp. 2d 613 (S.D.N.Y. 2001), aff'd, 2002 U.S. App. Lexis 3673 (2d Cir. Mar. 8, 2002).

37. Even though copyright law is federal law, contract disputes are decided under state law.

38. 533 U.S. 483 (2001).

39. Section 201(c) of the Copyright Act provides: "Copyright in each separate contribution to a collective work is distinct from copyright in the collective work as a whole, and vests initially in the author of the contribution. In the absence of an express transfer of the copyright or any rights under it, the owner of copyright in the collective work is presumed to have acquired only the privilege of reproducing and distributing the contribution as part of that particular collective work, any revision of that collective work, or any later collective work in the same series."

40. For example, the rights may be assigned, transferred by bequest or through bankruptcy. The copyright law also provides circumstances in which a contract assigning rights can be terminated and the rights reverted to the author or her heirs. The provisions of the law dealing with copyright transfer, including renewal, termination and restoration, are extremely complicated and beyond the scope of this paper. However, it is important to bear in mind that when a publisher refuses to grant a license for a particular use, it may be because it does not own the necessary rights, or ownership is ambiguous.

41. §205(a).

42. §205(c), (d).

43. However, even if the work is removed promptly the user may be liable for damages suffered by the copyright owner as a result of an infringing use.

44. §101.

45. *Harper & Row, Pubs. v. Nation Enterprises,* 471 U.S. 539 (1985).

46. §1201(a)(1)(A).

47. §1201(a)(2).

48. §1201(b).

49. There is an exception that permits a nonprofit library, archive or educational institution to circumvent a technological access control to make a good faith determination whether to acquire a copy of the protected work. However, the institution may not retain the copy so accessed longer than necessary to make that determination, nor use it for any other purpose. §1201(d).

50. §1201(a)(1)(B)-(E).

[51] The most recent version is the Paris Act, 1971. This is the version to which the United States has adhered. While the Berne Convention itself has no enforcement mechanism, the requirements of Berne were incorporated into the General Agreement on Tariffs and Trade, Agreement on Trade-Related Aspects of Intellectual Property Rights (GATT TRIPS) and are now subject to the enforcement procedures of the World Trade Organization (WTO).

[52] There is a companion treaty known as the WIPO Performances and Phonograms Treaty, or WPPT.

[53] The United States implemented the WIPO Treaties in the DMCA and has joined both treaties.

[54] WIPO Copyright Treaty, Art. 10.

[55] Sam Ricketson, *The Berne Convention for the Protection of Literary and Artistic Works: 1886–1986* §16.27 (Kluwer, 1987).

[56] Arts. 11*bis*(2) and 13(1). It can be argued that certain limited compulsory licenses are permissible under Berne, and some countries do employ levy schemes (e.g., charges on blank tapes and equipment to compensate rightholders for home audio and videotaping; the United States has such a measure in the Audio Home Recording Act, chapter 10 of Title 17). See Ricketson, supra note 55 at §16.28. However, any compulsory license that would subject copyright owners to broad unconsented-to use of their works could potentially violate Berne obligations.

[57] It is theoretically possible to treat U.S. and foreign works differently. Although the Berne Convention requires that a country provide these minimum standards to works of foreign nationals, a country remains free to accord its own citizens lesser rights. Berne, Art. 5(3). The U.S. applies differential treatment concerning copyright registration; when a lawsuit is based on a work of U.S. origin, the copyright must be registered before suit is commenced. §411(a). This is not true for a work of foreign origin, whose copyright need not be registered at all. However, differential treatment can problematic where it is unclear when a work is of U.S. or foreign origin.

[58] Obviously, this will differ with the type of work; if the archived work is an unprotected website, then the copyright owner already has such exposure.

[59] Usually the law of the country where the infringement takes place is applied, but the Internet raises complex choice of law questions.

[60] The fair use defense may be available in some circumstances, but would have to be evaluated on a case-by-case basis.

[61] See note 31, supra.

[62] This paper does not address whether or how the law could be modified to facilitate the development of a digital archive.

APPENDIX 7

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

It’s About Time: Research Challenges in Digital Archiving and Long-Term Preservation

Report on a Workshop on Research Challenges in Digital Archiving: Toward a National Infrastructure for Long-Term Preservation of Digital Information

Executive Summary of Pre-Publication Draft

Organizing Committee
MARGARET HEDSTROM, SHARON DAWES, CARL FLEISCHHAUER, JAMES GRAY, CLIFFORD LYNCH, VICTOR MCCRARY, REAGAN MOORE, KENNETH THIBODEAU, AND DONALD WATERS

Executive Summary

In April 2002, a group of computer scientists, information scientists, archivists, digital library experts, and government program managers met to examine the prospect of advancing computer and information technology research through a research program that addresses the unique challenges of long-term preservation of digital information. 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 many current technologies and information management methodologies. Digital archiving research is based on the premise that computer and information technology will continue to evolve at a rapid pace as long as many of the country’s best minds concentrate on information technology (IT) research and development, and as long as the IT sector continues to serve as an engine for economic development and growth. Some of the information created yesterday and today may move through many generations of information technology before it is reused at some point in the future. Other resources may be in continuous demand over many decades while new systems and technology evolve around the data. Long-term digital archiving requires systems, institutions, and business models that are robust enough to withstand technological failures, 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 and the Library of Congress. Some of the key recommendations of the workshop include:

- The National Science Foundation, the Library of Congress, and other government agencies should undertake a massive research effort to improve the state of knowledge and practice for long-term preservation of digital information.
- Important new research opportunities have emerged in computer and information science to address issues of storage and processing capacities, interoperability among heterogeneous systems, automation of many intake and preservation management processes, and complex metadata and semantic representation requirements.
- Long-term preservation issues will not be resolved through better tools and technology alone. Research opportunities abound around questions of economic and business models for affordable and sustainable long-term preservation programs. Research is also needed on policies and incentives for long-term preservation and on the economic, social, and legal impediments to digital archiving.
- Research in almost every discipline depends on well-managed, reliable, and readily accessible digital resources. Future research capabilities will be seriously compromised without significant investments in research and the development of digital archives.
- A pressing and urgent need exists to develop better solutions for long-term digital preservation in government agencies, libraries, archives, museums, private corporations, and even among private citizens who rely increasingly on the Internet to transact business and to communicate with colleagues, friends, and family members.

The report describes new challenges and opportunities in digital archiving, explains what is at stake if these challenges are not addressed, and sets out a research agenda with priority research areas and a discussion of research modalities and necessary investments.

New Challenges in Digital Archiving

Digital collections are vast, heterogeneous, and growing at a rate that outpaces our ability to manage and preserve them. One of the marvels of the information technology revolution is the continuous improvement in computer, memory, and storage performance and the simultaneous drop in costs. Thanks to what has been called “silicon scaling,” the processing power of a 1980s vintage mainframe computer now fits on a small silicon chip that can be embedded in any number of capture devices from complex remote sensors to consumer digital cameras. Digital storage devices and media have benefited from similar performance improvements and cost declines. More and more individuals can afford laptop and desktop computers with multiple gigabytes of storage. Larger organizations regularly add terabytes of storage capacity. One might suspect that archiving 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.

A few examples illustrate this problem. Internet search engines crawl the Web, copy Web pages, and then index them automatically so that users have a reasonable chance of finding information relevant to them on the Web. Large search engine companies, such as Google, index more than 2 billion Web pages and store copies in a cache as a backup in case the requested page is not available. But search engine companies are in the business of providing tools for searching and navigating. They are not in the business of long-term archiving of the Web or even a significant portion of it, nor should they be expected to take on this responsibility. Who will?

The Internet Archive, a public nonprofit organization, was founded in 1996 to preserve content distributed on the Web. In six 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. Although the Internet Archive has a policy to migrate its collections to new media at least once every 10 years, it has not yet undertaken one complete migration. As a small organization without a predictable, steady flow of resources, it is also seeking stable institutional partners, including the Library of Congress and the Smithsonian Institution, to collaborate in its long-term preservation endeavors.

Much more digital content is available and worth preserving; researchers increasingly depend on digital resources and assume that they will be preserved. During the last decade, many scientific, academic, and cultural organizations as well as government agencies and private enterprises have assembled valuable collections of digital information, either in the normal course of business or as special projects. Under the American Memory program, the Library of Congress led an effort to digitize more than 100 historical collections from materials in its own holdings and in libraries, archives, and museums across the country. The more than 7 million items in the American Memory collections are used daily by teachers, students, scholars, genealogists, and private citizens. The Digital Library Initiatives, sponsored by the National Science Foundation, Defense Advanced Research Projects Agency (DARPA), the National Library of Medicine, the Library of Congress, the National Aeronautics and Space Administration (NASA), and the National Endowment for the Humanities (NEH), fostered research and development for hundreds of digital libraries. Many digital library projects started as test beds 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 continuing research and teaching and to protect several hundred millions of dollars invested to digitize, organize, and provide access. Scholarly journals, preprints and even raw research data have moved online and become the preferred means for keeping up with new research in many fields. Such resources are emerging as the vital venue for scholarly communications. Society's ability to preserve a continuous record of research and scholarship will require an infrastructure for archiving digital scholarly communications that is as affordable and as robust as the complex networks and relationships among libraries, and between libraries and content creators, that have served reasonably well to preserve the published output for the last 400 years.

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 diversity 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. Indeed, the increasing use of complexity theory and integrated models in scientific research has generated the demand for massive datasets and complex analytical tools. Recently, NASA investigators had to use a combination of data from current satellites and from satellite instruments launched in the early 1980s in order 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. Digital preservation is important because it allows new data to be derived from unexpected uses of previous data. 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. Without better systems and methodologies for long-term preservation, integration of older and more recent data is costly and cumbersome, and many valuable resources remain at risk.

Government, commerce, and personal communications rely on digital information and communications. Critical needs for digital archiving strategies extend into almost all aspects of modern society. Whether carrying out business-to-business transactions, using the Internet to purchase goods and services online, communicating via e-mail, or using stand-alone computer systems, electronic transactions generate enormous quantities of information, some of which is worth saving for the long term. 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 are in digital form, must be kept as long as a particular model of aircraft is in service—a period that can exceed 50 years. The Food and Drug Administration (FDA) 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 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 those digital records that have long-term uses for keeping the business running and for maintaining accountability. The entertainment industry is shifting rapidly to digital masters of recorded sound, movies, and television programming. Within a few years, digital television and digital movies will be the preferred delivery method. Even private citizens are seeking ways to manage and preserve their e-mail, online accounts, and digital photographs.

What Is Unique About Digital Archiving Research? It's About Time.

Digital preservation shares many requirements with well-designed information systems, such as security, authentication, robust models for representation, and sophisticated information retrieval mechanisms. Nevertheless, unique long-term preservation requirements raise many interesting research questions that demand innovative solutions. One unique aspect of preservation is its concern with *the long term,* where long term may simply mean long enough to be concerned about the obsolescence of technology, or it may mean decades or centuries. 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, management, and funding.

The funding and business models for digital archives differ considerably from common business models that are based on relationships between investments, operating costs, and the utility of goods and services. Repositories may be expected to preserve digital resources even though their utility may not become apparent until well into the future. In this respect, the economic models for digital archives resemble the economics of public goods, where the primary beneficiaries of current investments may be future generations. Future users of digital archives will have different needs, expectations, technologies, and analytical tools from those of the communities that created the digital content initially. This raises challenging research questions in the areas of semantics and description and in knowledge-management technologies that will enable future reuse of digital archives. Another factor that distinguishes digital preservation research from many other types of research is the difficulty of knowing whether or not we have solved the problems successfully, because the ultimate test of success will be the new knowledge and discoveries that result at some future date. This problem requires some very challenging thinking about success measures and evaluation criteria, and it will demand an extended research effort over the next decade.

A Digital Archiving and Long-Term Preservation Research Agenda

Digital archiving challenges are ubiquitous and multifaceted. As a consequence, a significant, multidisciplinary research effort is needed to produce new knowledge in computer and information science, economics, and policy. Solving this complex problem will require many different approaches. We do not anticipate that a single solution will emerge or would be appropriate for the wide variety of collections, technologies, and organizational arrangements governing digital archiving requirements. At the same time, we believe that concerted research efforts will produce basic principles, new technologies, and new curatorial methods that will enable long-term preservation of vast resources at a fraction of the cost of today's immature and customized strategies. Opportunities for research partnerships abound between academic researchers, researchers in industry, and the many government agencies, cultural institutions, and private companies that are seeking solutions to long-term preservation problems. These research opportunities fall into four closely related categories: attributes of digital repositories, attributes of archived collections, tools and technologies, and economic and policy models.

Attributes of Digital Repositories

Even with a common conceptual framework, it seems unlikely that a single approach will satisfy all the digital preservation needs of various organizations and individuals. The development of infrastructures for digital archiving is strongly driven by the need to support multiple communities. Each community has unique requirements that will influence the design of the digital archive. Computer, information science, and engineering research is needed on a spectrum of archival repository designs. Variations in archival repository models raise many different research issues.

Data Model-driven Architecture

This model is used to preserve specific types of data for future reuse. Associated research issues include capacity and scalability of multiple petabyte repositories and methods for automated acquisition, quality control, and description.

Controlled Access Repositories

Research questions derive from stringent requirements for auditability, authentication, and access controls.

Archives of Temporally Changing Data

These archives preserve data that are continually changing, either through regular additions that are streamed into the archive or through updates and changes.

Research is needed on definitions, methodologies, and tools for time-based capture and representation, for taking useful snapshots of dynamic databases, for versioning, and on the identification of knowledge models to represent temporal or procedural relationships.

Archives of Evolving Data

Preservation and management of many types of digital information require transformation of the original data to new formats or canonical forms. Research is needed to better define and characterize transformation processes so that they can be automated, and so that transformations made on the original data can be documented.

Archives of Derived Data Products

Archives are not limited to the original materials. In the scientific community, processing may be done on archived collections to create derived data products to address scientific questions. Research issues include the ability to characterize the derived data products with descriptive metadata. This descriptive metadata can include the type of processing algorithm that was applied, the mathematical expression of the related operation, and the associated software implementation.

Repurposing of Archives

Many archives will need to enable new access mechanisms so that their collections can be used for different purposes from those originally envisioned. Repurposing of archived material may require the ability to stream the entire collection through processing steps. This requirement illustrates the need to think of archives as repositories of information and knowledge that may need to be updated at periodic intervals. Archives in the future may be dependent upon the ability to support generation of new semantic indexing through the processing of every digital entity.

Although this spectrum may not capture all potential types of archival repositories, it illustrates the need for research that more closely examines the relationships between the purpose of the archive, the types of data and information that it acquires, and the needs of its producer and user communities.

Attributes of Archived Collections

A great deal of information is "saved" in digital form on file servers, on personal hard drives, and in large repositories of tapes and optical disks. Nevertheless, archived collections have additional attributes that enhance their quality, utility, trustworthiness, and longevity. Archival collections don't just happen when someone clicks on the "save" icon—dumps of saved documents offer precious little for future researchers because they lack critical contextual and content-oriented metadata. Rather, archival collections are created through curatorial processes that include selection, organization, description, and quality control, and they require individuals or organizational entities that will take on formal responsibility for long-term 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—from text, to images, to recorded sound, to computer models and simulations, to digital video, plus all combinations of these object types. Research is needed in several key areas to better define the attributes of archival collections and curatorial processes, including:

Selection and Preservation of Complex Digital Objects

Methods exist today to preserve simple, static digital objects, but managing and preserving complex multimedia objects and dynamic objects that change on a regular basis present significant challenges. An increasing percentage of born-digital content falls into this category.

Aggregation of Items and Objects into Collections

With the need to capture materials from the Web before they are updated or deleted, research is needed to determine the appropriate extent and depth of Web-based collections, to bring coherence to widely distributed collections, and to further develop effective and economical collection-level metadata schema that describe attributes common to all items in a collection and provide for inheritance of metadata from the collection to the item level.

Decision Models for Selection

Long-term preservation does not imply that everything is worth saving. Most libraries, archives, and museums have well established collecting policies for physical items, but selection decisions in the digital realm are becoming more complex. An increasing amount of the content that libraries deliver to users is held in publishers' repositories and is not owned physically by the library, raising 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. Collecting policies that were designed for physical materials do not encompass new types of digital objects and collections (such as Web sites and multimedia productions). Formal models of selection decisions are needed so that tools can be developed to assist curators with selection responsibilities and to automate some selection decisions, but not to eliminate the considerable human judgment that goes into collection development.

Resolution of Naming Hierarchies

Multiple naming conventions are used 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, to relationship names used to describe properties of the collection. Preservation requires the ability to manipulate each name space at some arbitrary point in the future. A major research question is whether the generalization of name spaces as ontologies that characterize either semantic relationships, structural relationships, or logical relationships will lead to a simpler way to preserve the information and knowledge content of archives.

Tools and Technology

Human labor is the most expensive component of digital archiving systems. Therefore, research and development of better archiving tools and technologies will not only make digital archives more robust and reliable, but also drive down the costs of this endeavor. Some of the priority areas of research and technology development include:

Acquisition and Ingest

Archives can use automated Web crawlers and harvesters (the "pull" method) or formal submissions (the "push" method) or some combination of these to acquire digital content. Both models would benefit from research that allows finer tuning of ingestion tools and that are better integrated with selection criteria and subsequent preservation management requirements. Given the vast quantities of data likely to flow into digital archives, tools are needed for automated indexing, metadata extraction, validation, and quality control. Tools are also needed to transform disparate types of objects into the formats, standard forms, and data models that a repository can manage over the long term.

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 over time. Research is needed to develop methods for unique and persistent naming of archived digital objects, tools for certification and authentication of preserved digital objects, methods for version control, and interoperability among naming mechanisms used by different content providers. The emergence of data grids that create global name spaces is an example of a technology for persistent naming. This technology needs to be extended to support persistent naming of the information and knowledge content of the collections.

Decision Models and Metrics

In addition to decision models to support selection, research is needed to develop models and tools that will support decisions regarding preservation formats and standards, choice of preservation strategies (normalization, migration, emulation), and on the costs and benefits of various levels of description and metadata. Key research areas include metrics for measuring the quality and fidelity of preserved digital objects and for documenting the consequences of archival processes on them.

Metrics need to include the maximal sustainable archive size (as a function of the access rate), the archival bandwidth (amount of material that can be moved forward into the future as a function of the type of storage technology), and the repurposing rate (the amount of time needed to process the entire collection to derive new collection attributes).

Standards and Interoperability

Standards for data formats, data models, metadata, and many other aspects of digital information are useful for long-term preservation, but standards change over time and archived digital entities will have to be migrated to new standards in the future. Longevity of digital information will be enhanced through research on standard and long-term methods for representing text, sound, image, video, and other object components and for characterizing their semantic, temporal, spatial, and procedural relationships. Archived digital entities will have to be migrated to new standards in the future. A migration can be viewed as "lossless" if the new standard provides a superset of the features of the old standard. A goal for standard encoding formats is the creation of lossless feature conversions when migrating between standards. Research is also needed to support interoperability among different competing standards and for developing models that help predict which standards are likely to achieve widescale adoption over extended periods of time.

Policy and Economic Models

Even the most effective tools and technology will be useless without a policy and economic environment that is conducive to long-term preservation. The area of policy and economic models is ripe for research. Some of the key research areas include:

Incentives for Long-Term Preservation of Digital Information

Research is needed on a variety of incentives that would encourage organizations to develop digital archiving capabilities, build repositories, provide archiving services, and create content in ways that facilitate its long-term preservation. A variety of mechanisms warrant investigation, including direct public subsidies, tax incentives for placing content in the public domain prior to the expiration of copyrights, philanthropic donations, and market mechanisms that provide for cost recovery or revenue streams to support the repository.

Incentives for Deposit of Digital Content into Archives

Conversely, content creators need incentives to deposit content in repositories for long-term preservation. Research in this area is closely tied to the concept of trust. Depositors must have a very high level of trust in a repository based on secure technology, a track record of performance, and consistent application of rules and agreements.

Metrics

There is a critical need for research that will produce metrics and methods to measure almost every aspect of digital archiving, from the performance of storage media over the long-term, to the effectiveness and costs of different preservation strategies, to the market value of archiving services and market analysis of user demand. Evaluation of digital archiving is impossible without concrete measures of the costs, benefits, and value of digital objects.

Intellectual Capital

Archives need to become the repositories of intellectual capital that are viewed as the driving resource for economic growth. This emphasizes the view of archives as information and knowledge repositories. The goal of the archive is to make the information and knowledge content as readily accessible as possible, and to make it easy to repurpose the collection for a new use. Digital archiving research is needed to achieve this goal in ways that are sustainable, manageable, and cost effective.

Research Modalities and Scale

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. A concerted, focused effort is needed now that engages a sufficient number of researchers, involves government agencies and other partners with substantial digital archiving needs, and mobilizes an appropriate level of investment to address the problem effectively. We anticipate that a minimum investment of $5 million to $8 million per year is needed for a focused research program for the next 10 years. The 10-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 10-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 launch this research initiative; encourage sponsorship from other government agencies, private foundations, content providers, and industry; and participate in active partnerships with researchers from many disciplines.

One exciting aspect of research on digital archiving and long-term preservation is that the research is amenable to many different research methodologies and innovative approaches. Possible research methodologies cover a whole spectrum, from small, single investigator projects to test beds involving many researchers and multiple participating institutions. Another attractive feature is that, although oriented to the long term, 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 by creating business opportunities for new technologies and services. Therefore, we recommend support for a wide variety of research modalities, ranging from small, single-investigator projects to the creation of two or three large test beds involving multiple institutions, large teams of researchers, and experimentation with existing digital collections with obvious long-term value. Many opportunities exist for partnerships between researchers and organizations of all sorts that hold significant digital collections and face pressing digital archiving needs. There may be benefit to creating one or more centers for digital archiving and long-term preservation research to serve as focal points for this effort and to address issues of technology and knowledge transfer, education and training, and capacity building.

Conclusion

It's about time to launch a new research initiative that will advance research in computer and information science, information economics, policy, and social and organizational behavior while addressing critical needs in government, the private sector, universities, and cultural institutions to find reliable, sustainable, and cost-effective means to preserve valuable digital information resources that are critical to near-term and long-term discoveries of new knowledge. A concerted research effort undoubtedly will advance our knowledge in many disciplines while also contributing to the foundation and infrastructure for the discovery and generation of new knowledge in the future. The full report presents a more thorough discussion of needs and opportunities for digital archiving and preservation research.

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, United Kingdom

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

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

Note

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.
A final report from the workshop will be available pending further external review in September 2002 at *www.si.umich.edu/digarch/* and in printed form.

APPENDIX 8

Highlights of the Library of Congress's Scenario Learning Process on the Future of Digital Preservation

Highlights of the Library of Congress's Scenario Learning Process on the Future of Digital Preservation

CHRIS ERTEL AND CHRIS COLDEWEY
Global Business Network

Introduction: The NDIIPP Scenario Learning Process

The National Digital Information Infrastructure and Preservation Program (NDIIPP) Plan raises a number of uncertainties in the external world that could significantly affect the future of digital preservation. How quickly will today's digital technologies be rendered obsolete by even newer technologies? What will those future technologies look like? How might public attitudes and practices related to intellectual property rights evolve in an increasingly digital world? How will economic trends and shifts in the federal budget affect the ability to make wise public investments in the future?

These are just a few of the many important questions relevant to the future of digital preservation that must be reckoned with, but which cannot be answered definitively. In general, the need to grapple with uncertainty becomes all the more important—and difficult—as the time horizon of a strategic challenge grows longer. In the case of digital preservation, the time horizon ultimately extends out as far as the digital collections themselves—which, one hopes, will be on the scale of decades or even centuries. Surely no expert can claim to be able to answer any of these important questions—much less all of them—over the long course of the preservation time horizon.

Yet, even in the face of such daunting uncertainties, we must act. Scenario learning—also known as scenario planning—is an approach commonly used by leading organizations in the private and public sectors to craft adaptive strategies in such a climate of high uncertainty.

In scenario learning, an organization creates a small number of detailed stories—or scenarios—about how the future might unfold based on different outcomes of critical uncertainties in the external environment. The organization then uses these scenarios as a platform from which to identify a high-level vision of a desired future state that it would like to achieve, and to design a course of action toward that vision that can be adaptable in multiple environments. Finally, the scenario learning approach helps an organization to be perceptive and flexible in correcting its course of action when future surprises in the external environment inevitably occur.

Scenario learning thus offers a sensible middle path between the two extremes that too many organizations fall victim to in their planning—of either pretending that they know the one true future that will unfold, or of being paralyzed by uncertainties altogether.

As part of creation of the NDIIPP Plan, the Library engaged Global Business Network (GBN), a leading futures organization based in Emeryville, California, to facilitate a scenario learning process to inform and help shape the Library's larger strategy and planning process. This scenario process was designed to be consistent with the Library's desire *not* to search for "the one right answer" to the challenge of digital preservation—which all expert informants agreed cannot be found at this time—but rather to identify a high-level vision of a desired future state, and then to chart a course of action that will allow the Library and its partners to learn their way into the future in a collaborative and iterative fashion. GBN's specific roles were to facilitate the learning process and to assist the Library in reaching out to a broader range of external expertise and potential partners for possible collaboration. The GBN team also provided "talking partner" assistance on the future direction; however, the ultimate crafting of strategy and the action plan was left to the NDIIPP team, which will be charged with future execution.

This scenario learning process ran from approximately January to August 2002, and included the following major steps:

- pre-work: Project Design and Initial Issue Inventory
- Major Workshop No. 1: Exploring the Future Environment and Defining Possible Solution Spaces
- convening of the Preservation Architecture Task Force
- key Stakeholder Interviews
- Mini-Workshop in Hollywood
- Major Workshop No. 2: Testing the Proposed Preservation Architecture and Exploring Possible Pilot Project Experiments
- Synthesis of Results for Strategic Direction and Master Plan

Without dwelling on the details of each of these steps, the highlights of this scenario learning process are discussed in the sections that follow.

Major Workshop No. 1: Exploring the Future and Defining Possible Solution Spaces

The scenario learning engagement began shortly after the conclusion of the stakeholder meetings that were held in Washington, D.C., during November 2001. Team members from GBN attended these three convening sessions to ensure continuity between the scenario learning process and the important work that preceded it.

The purpose of the convening sessions and much of the work by the Library prior to the scenario learning process was to elicit a broad a range of information and opinion on the topic of digital preservation from as many different informed stakeholder viewpoints as possible. With the scenario learning process, the Library and GBN together began the process of engaging a subset of these stakeholders in the challenging work of defining some possible broad solutions that could take into account the diversity of interests and concerns, as well as the landscape of uncertainty in the external environment. Using this approach, GBN and the Library intended to elevate the current discussion on digital preservation beyond a sharing of perspectives and toward the acceptance of shared responsibility for creating solutions that could be acceptable to a wide range of stakeholders.

The initial scenario creation workshop, held February 13–14, 2002, in Berkeley, California, was facilitated by Peter Schwartz, Chairman of GBN, and included 26 highly skilled participants representing a wide range of expertise from the content-generating industries (e.g., book publishing, news media, music recording, film production), archival and scholarly institutions, other government agencies, and others with experience in fields related to digital preservation (e.g., technology development, the economics of information). The workshop followed a time-tested methodology whereby scenarios were developed to illustrate diverging views of the future based on different outcomes of critical uncertainties and the different perspectives of workshop participants themselves.

The initial discussion at the workshop yielded a focal question to frame the issues relevant to the future of digital preservation: What will be preserved, how, for what purpose, by whom, and who is going to pay? The time frame agreed upon for the scenario-building exercise was 15 years, to the year 2017. The group then brainstormed an exhaustive list of influential external forces (social, technological, economic, environmental, and political), and with these in mind, outlined a "drift scenario" that imagined the implications of a future where the Library of Congress took no action regarding digital preservation. With this null hypothesis brought to life, the group then took on the task of developing detailed, alternative visions of how the future of digital preservation might unfold over the next 15 years. Three scenarios ultimately emerged from this session and were expanded in detail by smaller teams of participants.

Each of the scenarios developed describes a particular strategy for digital preservation that the Library could undertake, depending in part on the future course of external forces such as the pace of technological change, future agreements on intellectual property rights, future levels of federal government spending, and so on. Each of these scenarios assumes a different external environment, especially in terms of the pace of technology development, the intellectual property regime, general economic conditions, and the political climate, especially regarding the evolving role of the state. However, for the sake of brevity, this narrative will tread lightly on these external driving forces in order to focus on how each scenario took a different approach to the solution space.

The Scenarios

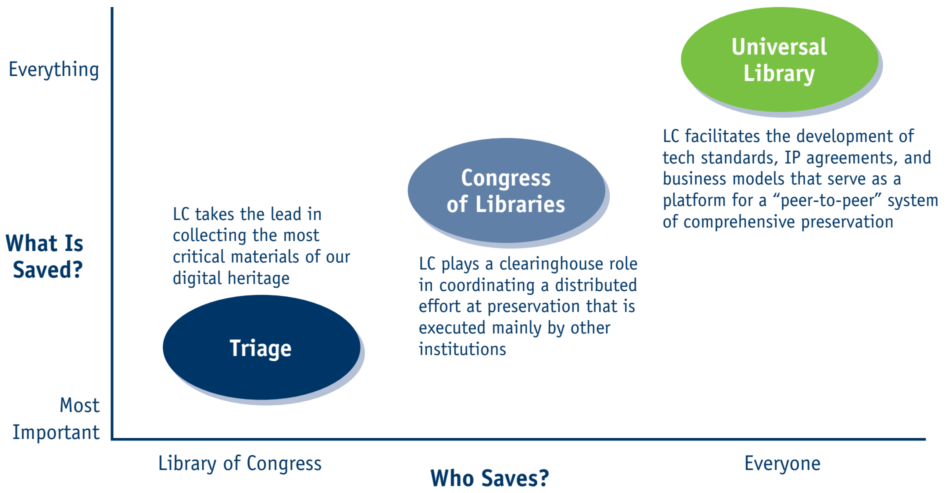
Using different assumptions in the external environment as starting points, the three scenario breakout groups explored a variety of approaches to preservation. As the groups compared their initial work in plenary conversation, a simple framework of two axes was developed that captured the emerging range of preservation options. The horizontal axis in the graph (below) shows the extent and distribution of the preservation effort, ranging from only the Library of Congress to everyone. The vertical axis represents the comprehensive scope of the preservation effort, from only the most important content to everything.

Triage—The Library of Congress as Central Repository for the Most Critical and At-Risk Collections

The triage scenario is most consistent with a world in which technology change is relatively slower, the growth in the economy and federal budgets is modest, the role of the state is not expanding, and new economic models and intellectual property regimes to support digital preservation are slow in coming.

In this scenario, there is a strong focus on developing clear, explicit standards for what limited items should be collected and preserved. The collection effort would sit mainly within the Library itself and emphasize discriminating, curatorial selection, rather than an exhaustive approach. A three-tiered system of data classification would

Figure 1. Scenario Framework


Everything
What Is Saved?
Most Important
Library of Congress
Who Saves?
Everyone
Triage
LC takes the lead in collecting the most critical materials of our digital heritage
Congress of Libraries
LC plays a clearinghouse role in coordinating a distributed effort at preservation that is executed mainly by other institutions
Universal Library
LC facilitates the development of tech standards, IP agreements, and business models that serve as a platform for a “peer-to-peer” system of comprehensive preservation

be used to determine which data to preserve. The highest tier would consist of data that are desired, but already being preserved by other means, perhaps commercial. The middle tier is data that are desired and are not otherwise being preserved, where the Library would become the collector of last resort. The bottom tier consists of data that are deemed not worthy of being saved, because it would be too costly, or because the data are too obscure, or too ephemeral. The Library would have several roles in this scenario: ensuring preservation of this middle tier of content, managing preservation standards, enabling preservation tools, and experimenting with a variety of approaches and partners. These experiments would focus on finding successful approaches regarding technology, partners, business models, and perhaps higher-level services such as registries, self-reporting, auditing, and metadata standards. The goals of this effort are to build trust and establish clear progress in digital preservation, recognizing that a larger solution is impossible to create, at least at first.

Congress of Libraries—The Library of Congress as Portal and Keeper of "The List of Lists"

The Congress of Libraries scenario is most consistent with a world in which technology development is relatively fast and there is gradually expanding trust over emerging intellectual property agreements, but in which the resources and role of the federal government are in a general retreat.

This scenario describes an approach to digital preservation wherein the Library of Congress functions mainly as a convener of, and a portal to, other preservation organizations. Recognizing the vast scale of the broader digital preservation enterprise and the difficulty of taking a lead role, the Library directs its efforts toward managing a "list of lists," pointing to information rather than creating a central repository. The Library might also develop standards for deposition and collection that the network may adopt to maximize its scope. This approach strongly relies on good communication among a loose network of ever-growing preservation organizations—public and private libraries and collections, nonprofit organizations, and content-generating businesses. Instead of managing a centralized collection, the Library of Congress would serve as the portal to these other resources and point to information, if not actually guaranteeing access to the material itself.

This scenario of preservation would require less financial backing than a more centralized or actively curatorial approach. The workshop participants presented it as an option that might weather budget cuts and administrative priority shifts, even if it might not be an ideal role for the Library of Congress to play. Detractors feared that this approach would lead to significant losses of digital material, as neither the Library nor anyone else would assume responsibility for ensuring that the public interest was served in the decisions about what materials would be preserved for the future. Ultimately, the effectiveness of this approach would depend largely upon the wisdom and capacity of other organizations to do the lion's share of the work in preserving digital content.

Universal Library—The Library of Congress as Key Enabler in the Creation of a Robust, Peer-to-Peer, Distributed Network of Preservation

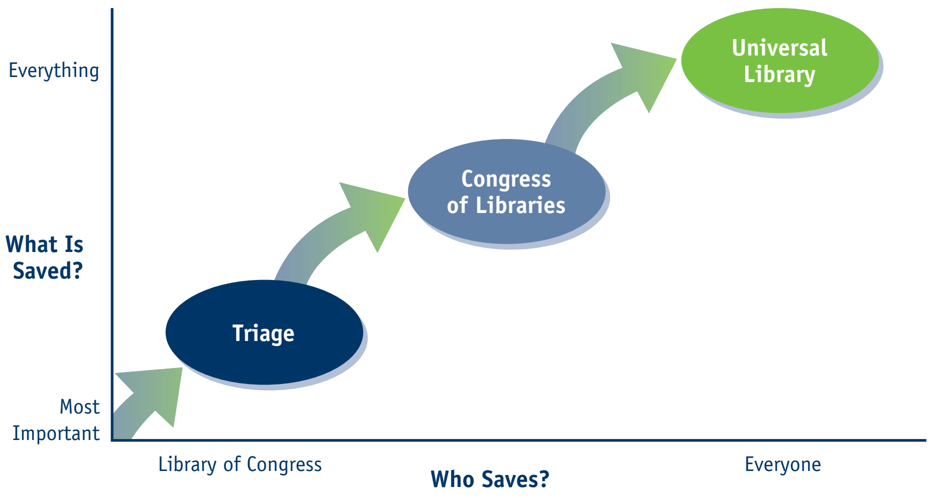
The Universal Library scenario is most consistent with a world in which technology development is especially rapid, economic growth is solid, there are expanding agreements related to intellectual property that increase fluidity in the system, and the role of the state is evolving in new directions.

In this high-profile approach, the Library of Congress takes a lead role in catalyzing and enabling digital preservation nationwide. To engage all stakeholders, the Library would create a central organizing body: the National Digital Preservation Board. The NDPB would coordinate collection development, design the architecture of the repositories, set policies, and ensure standards of excellence. For example, media providers could interact with the NDPB to ensure preservation of their materials and negotiate deals to charge for access. These providers, major state library systems, and private collections would make up the bulk of a "Unilibrary," which actually would provide content, rather than just point to it (as in *Congress of Libraries*). Although the system would ensure access, the potential for content losses still remains, as participants in the system ultimately decide what is to be preserved. However, there is greater leverage for increasing the scope of what is preserved and accessed here than in either of the other scenarios, and the high level of redundancy of preservation within the system should reduce the risk of loss. The system would derive revenue from access fees and tax incentives.

Sequencing the Scenarios

As described above, these three scenarios can be understood as separate approaches that are differentiated by the extent of preservation and the actors involved. However, as illustrated in the graph below, the scenarios also form a possible sequence of complementary approaches—a potential model to show how efforts by the Library to preserve digital materials could progress and evolve over time.

Figure 2. Scenario Sequencing



The sequence begins with *Triage*, as that scenario articulates clear short-term goals and immediate steps for action by the Library of Congress. The experimental partnerships described in Triage would widen the scope of preservation, lay the groundwork for establishing trusted linkages with other preservation organizations, and even catalyze some progress on rights and access issues.

These steps would pave the way for a *Congress of Libraries* to develop. As institutions begin to link up, the Library would potentially face a steep learning curve in facilitating a federated system of institutions, or could encounter scaled-back Congressional support.

This model allows and encourages the network to develop during this period. In this phase, the Library takes small steps into the use and management of distributed systems, allowing for experiments to occur regarding appropriate models for access and institutional involvement. As the system matures, different needs and capabilities arise, overseeing preservation standards and further distributing the preservation burden among others.

Universal Library is the culmination of this sequence, and here the Library reemerges as an active leader. Though the preservation effort is most widely shared here, the large number of constituencies and very broad focus increases the need for central coordination, guidance, and support. This scenario is designed with continued federal and tax-incentive support in mind, which places it further along the continuum of possible approaches. This scenario sequence articulates one possible vision of how a national system of digital preservation might evolve.

To be sure, the sequence described above is a highly stylized story, and reality is likely to be more complex. Most often, after good scenario work has been done, the actual future that occurs is some combination of the imagined futures—with different scenarios often playing out simultaneously in different places and situations. Still, as a stylized story, the sequence of scenarios provides one view of how the future might unfold from one solution toward another over a longer period of time.

Scenario Implications

After this critical scenario development workshop was completed, core team members reflected on the rich output from the session and identified a number of key implications that emerged across the divergent scenarios.

- The workshop participants quite deliberately did *not* create a scenario in which the Library of Congress saves everything, or anywhere close to everything. The issues involved in replicating the comprehensiveness of the Library's physical assets were considered to be far too complex to imagine this in the foreseeable future. At the same time, participants did imagine an important role for the Library in all future solutions to the challenge of digital preservation.
- Each scenario developed by the group would ultimately require that the Library of Congress make significant changes in key investments and core competencies—

although the specific nature of many of these would differ importantly depending on the scenario.

- As the amount of material that is being saved increases, the role of the Library as a central repository decreases. Yet, even within a broad-based, peer-to-peer system, there is a great need for a focusing of effort—for standards, making connections, etc.—but without central control *per se.* Open source systems require a strong focusing actor to make the system work, but that center must also invite and *encourage* active outside participation (within well-defined limits) in order to succeed.
- Depending on the outcome of various external factors (e.g., the pace of technology change or rights agreements), different scenarios may be seen as playing out in different domains of digital content at the same time in the future. For example, a "Triage" approach may be necessary to save content in one area where there is weak institutional support for preservation (e.g., in the case of Web coverage of major world events), while at the same time an area with clear rights understandings might be ready to embrace a "Universal Library" approach (e.g., some publishers of e-journals).
- Most workshop participants supported the notion that the best strategy is to get into the learning loop as quickly and strategically as possible. While it is impossible to know now what approach will be best, it is very realistic to make step-wise and iterative progress toward a better future.
- In order to begin the learning loop, there is a strong need to define the playing field in a way that clarifies roles, offers flexibility, and provides a focusing device for institutions to make clear choices about if and how they would like to participate in any broad-based, national effort at digital preservation.

Convening of the Preservation Architecture Task Force

After the scenario development workshop, the GBN-Library core team discussed at length what had been learned, and agreed on a next course of action. The team quickly agreed that the most important implication from the session in terms of next steps was the need for the Library to more clearly define a more specific context—a high-level preservation architecture design—around which subsequent conversations around concrete next steps could occur.

This learning emerged most clearly at the first workshop during the small-group breakout work on the "Universal Library" scenario. In the course of fleshing out this story—the most ambitious high-level preservation solution discussed by the group—it became clear that there was a great need for a shared framework around which different parties with very different interests could find a choice of roles and levels for participation within a clearly defined system. Without such a framework, it would be extremely difficult to "scale up" a national approach with so many different types of content, players, and objectives.

As a result of this learning, a special task force of seven participants was convened at GBN on April 3–4, 2002, to create a high-level design for a preservation architecture that could serve as this shared framework for future progress. The main results of this work are described in Appendix 9. It is important to emphasize that, although much of the description of this preservation architecture relates to technological infrastructure, from a strategy perspective the most critical element of the preservation architecture is that it creates a neutral forum, or platform, for stakeholders with a wide range of interests and concerns to: (1) determine how and where they want to contribute to and benefit from the emerging preservation system; and (2) understand the roles, responsibilities, and boundaries that are associated with the level of participation that they find most appropriate. Of course, the preservation architecture also serves as the conceptual backbone for the creation of an appropriate technological infrastructure. But this infrastructure should be built once the playing field is well defined and the roles, rules, and responsibilities of key players are made clear. For this reason, the task force spent much of its time discussing different kinds of roles and responsibilities that the architecture might need to accommodate, and less time discussing technological specifications like the kind of servers that would be needed, and so on.

Major Workshop No. 2: Testing the Proposed Preservation Architecture and Exploring Possible Pilot Project Experiments

On April 29–30, 2002, a second major workshop was held in Arlington, Virginia, to further explore the implications of the scenarios for the future of digital preservation, to scrutinize the first draft of the high-level preservation architecture, and to brainstorm potential next steps in terms of possible pilot project experiments. Thirty-three participants attended this second workshop, representing a similarly broad range of interests and expertise as at the first workshop. Again, Peter Schwartz, Chairman of GBN, was the lead facilitator.

Much of the first day of the workshop was spent in animated discussion around the proposed preservation architecture. While many refinements were suggested, by and large the participants found the architecture to be a very useful starting point for discussion around different possible solutions, roles, rules, and responsibilities related to the challenge of digital preservation. The current version of the preservation architecture, described in the NDIIPP Plan, has benefited significantly from revisions suggested during this session.

Once the workshop participants critiqued the proposed preservation architecture, they brainstormed a larger number of possible pilot projects. Next, the participants divided into self-organized small groups to create a smaller number of draft projects that could begin to populate and bring to life the preservation architecture.

Upon reflection after this second major workshop, a few major implications from the session became clear.

- While the preservation architecture described in the NDIIPP Plan is still specified at a high level, and will require further definition over time to become operational, the initial high-level design has proved to be sufficiently detailed to serve as a useful starting point for meaningful interaction among key stakeholders.

- Given the high level of diversity of projects that were brainstormed and developed, the group showed optimism that the architecture design could be scaled up or down depending on the specific needs of the participants. Specific project ideas were developed in relation to each of the scenarios and each of the four levels of the preservation architecture (Repositories, Gateways, Collections, and Interfaces).
- Perhaps most important, the number of ideas for pilot project experiments that emerged from the group of participants—as well as the level of energy with which they were created—attested clearly to the high level of interest among this group of key stakeholders to move beyond exploring issues related to the challenge and toward specific actions toward a future solution. Indeed, months after the second workshop was complete, interest among many participants in moving forward with specific digital preservation projects continued to be strong.

Conclusion

As a planning approach to digital preservation, the scenario learning process described above can be seen as foreshadowing, in many ways, the much larger learning journey that the Library can expect to undergo in the months and years ahead. The scenario learning process was very iterative and at times even messy, taking unexpected twists and turns along the way, and even chasing down a blind alley or two. In the end, though, the process served its purpose in leading toward a clear strategic direction and course of action. Most importantly, the process provided a "soft structure" for continuous progress—that is, just enough structure to focus attention, yet not so much structure to restrict options or discourage creative solutions. Such an approach is especially important in tackling a challenge that will require a high level of active collaboration with many diverse stakeholders in order to succeed.

In the coming months and years, the Library intends to create a collaborative network to ensure the preservation of digital culture, which represents a growing—and highly at-risk—share of our modern heritage. Along the way, the Library means to use a series of pilot project experiments to bring to life the preservation architecture, while conducting basic research on key issues. No doubt, many new lessons will be learned along the way, and some unexpected turns will be taken. If the scenario learning process of the past nine months is any indication, the learning process will not always be easy, and perhaps even a few wrong turns will be made. But in this Plan, a course has been laid that should provide the kind of guidance needed to keep the learning process moving forward in very productive directions.

APPENDIX 9

Preliminary Architecture Proposal for Long-Term Digital Preservation

Preliminary Architecture Proposal for Long-Term Digital Preservation

CLAY SHIRKY

Introduction

During the scenario planning sessions, it became clear that in order for the various stakeholders to be able to collaborate on long-term digital preservation, the National Digital Information Infrastructure and Preservation Program (NDIIPP) would have to provide a technical architecture that would support their various efforts. In response, the Library convened an Architecture Group, made up of representatives from the Library, other library and archiving efforts, and the computer industry. This group developed a conceptual framework for supporting the technical functions of NDIIPP, which we are calling the preservation architecture.

The architecture began with a foundational set of assumptions, listed below, and describes both the components and some basic rules for their interconnection. It assumes that the NDIIPP will be built over time, and that its construction will involve both public and private institutions as well as the Library. It also assumes that the NDIIPP will never be finished in any static sense, but will instead need to be able to evolve continually to be able to integrate new forms of hardware and software, and to preserve digital material of new formats and types. In order to accomplish these things, the architecture proposes building the necessary infrastructure in four layers, with each layer embodying a different set of functions and a related set of rules for use. These layers and their interconnections are designed to allow preserving institutions to customize the architecture to their particular needs, and to make it possible to adjust the architecture as those needs change.

The following document outlines the architecture in its current state. It is important to stress that the architecture outlined below represents only a theoretical starting point, and the actual infrastructure that is built to support NDIIPP will require significant work on the individual components, the protocols that hold them together, and the interaction of the infrastructure as a whole in a wide variety of circumstances. It will also require more basic research into areas such as digital security and methods for validating that the stored digital materials are kept "fit for use" over time.

Basic Assumptions

The Architecture Group began this process with some basic assumptions about the appropriate approach to the problem, both about the environment in which the project was taking place, and about core technical considerations.

The two environmental assumptions were:

Don't Reinvent the Wheel

The Architecture Group was keenly aware of the excellent work being done on digital preservation and related issues by a wide variety of federal agencies, nonprofit and educational establishments, and commercial concerns. In particular, the Open Archival Information System (OAIS), digital standards bodies such as the World Wide Web Consortium (W3C) and the Internet Engineering Task Force (IETF), the National Science Foundation (NSF)–led interagency Digital Libraries Initiatives, and the Digital Library Federation, among others, are all doing important work in this area.

If the Library did no work whatsoever on problems relating to long-term digital preservation, some aspects of the problem would nevertheless improve, because of the work being done elsewhere. Therefore, the Architecture Group began with the assumption that the Library should research existing work, and that the NDIIPP should be designed so that this work could be used where available.

Accept the Inevitability of Legal, Cultural, and Economic Change

In the current climate, issues surrounding ownership and use of digital materials are quite volatile. The legal background for digital intellectual property is changing rapidly, with the appearance of new laws like the Digital Millennium Copyright Act (DMCA) and new technologies that rely on these legal protections, like Digital Rights Management (DRM) schemes. Seeing this volatility, the Architecture Group began with the assumption that any proposed architecture should be flexible enough to adapt to changing legal, cultural, and economic norms for the sale and use of digital materials.

In addition to these environmental assumptions, the Architecture Group also began with a number of assumptions about the technological requirements for the NDIIPP:

Use a Modular Approach

We began with the assumption that any infrastructure that supports the NDIIPP should be built modularly, rather than monolithically. We looked at the design and deployment of the Internet generally, and the World Wide Web specifically, as examples of the value of modularity in large-scale systems.

A modular approach provides several advantages for an undertaking of this scale and duration, including allowing the infrastructure to be built in part and over time, allowing it to be built out of components made by different technology suppliers rather than a single vendor, allowing it to be upgraded in pieces over time, rather than all at once, and making it possible to integrate new technologies as they arise, without forcing the reengineering of the whole infrastructure.

Define Minimal Requirements for Each Layer

Because the possible complexity of modular infrastructure grows very quickly, the benefits of modularity can only be captured if the relations between the components are simple enough to be built, debugged and maintained by a diverse group of participants. Having adopted modularity as a basic design goal, the Architecture Group then assumed that the protocols governing the conversation between the different components must be defined as simply as possible.

This is not to say that the information carried between layers of the infrastructure cannot be complex, but rather that this complexity should be optional and defined by the participating institutions where and as needed. To keep this complexity optional, the protocols that are required for all users, irrespective of their needs, must be kept very simple.

Assume Heterogeneous Components

The Architecture Group assumed from the beginning that the infrastructure should not only be modular, but that it should be as tolerant of a variety of hardware and software as possible. This reduces the risk of commercial capture, as well as the risk that an undisclosed flaw in a particular kind of component could threaten the entire infrastructure. As with biological systems, diversity confers a large degree of resistance to catastrophic failure.

Never Optimize the Infrastructure for Any One Instantiation

As a corollary to heterogeneity, the Architecture Group assumed that the infrastructure would never be finished, for the same reason that large cities are never finished. Instead, we assumed that for the foreseeable future, some part of the NDIIPP would be undergoing additions or alterations. This in turn means that the NDIIPP should never be completely optimized for any particular state or any particular arrangement of components, because such optimization would always be premature.

Design the Infrastructure to Survive the First Migration

Finally, a principal design goal of the architecture is that it should be able to survive over the course of several migrations, where all parts of the infrastructure are replaced while it remains functional, and the best predictor of this facility generally is to be able to survive the first migration. For a modular infrastructure to survive, all its parts must be able to be replaced piecemeal, while the infrastructure remains in operation. This is true of the Internet today, for example, where none of the original computers connected to the network are still in operation, but the Internet as a whole has nevertheless survived.

With these environmental and technical principles in mind, the Architecture Group proposed a hypothetical four-layer architecture that will allow for the creation of a flexible, useful, and secure infrastructure that can be assembled and upgraded in pieces over time. The layers and their functions are presented below.

Layered Architecture

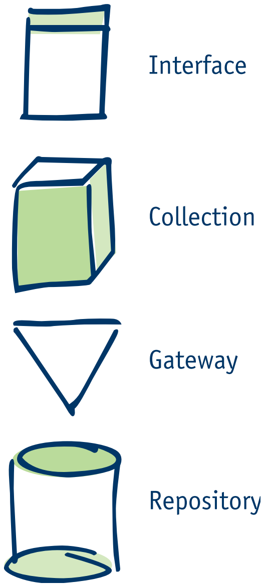
The proposed architecture is a stack of four layers, each of which performs a particular set of functions, and each of which interfaces with the layer above and below. From the bottom of the stack up, the layers are:

Repository, for storing bits,

Gateway, for governing access to Repositories,

Collection, where human judgment about the nature and value of the material is kept, and

Interface, where patrons access currently available material. (Note that there will be restricted-access material preserved within this infrastructure for possible future use, but which will not be accessible to the public in the present.)



This modular and layered approach creates several advantages related to the technical principles listed above, including:

- Allowing the infrastructure to be built in pieces and assembled over time, rather than requiring monolithic assembly,
- Allowing many different types of hardware and software, from multiple sources, to be used to build the infrastructure,
- Allowing issues of preservation of digital materials to be handled separately from issues of public access, so that commercially valuable materials can be preserved securely, and
- Creating the kind of modularity necessary to allow the infrastructure to be upgraded piecemeal, rather than all at once.

The following sections detail the design rationale for each of these four basic components, from the bottom of the stack up. Each section begins with a drawing of the four layers, and an illustration of where the particular layer fits in the overall infrastructure.

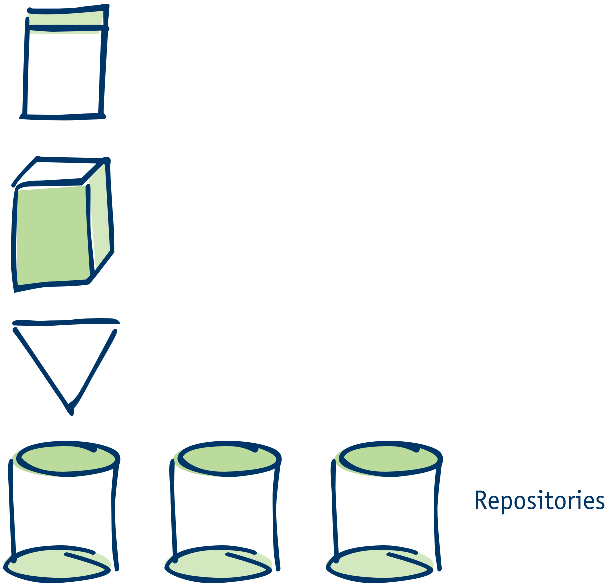
Repositories

A Repository is the lowest level of the proposed architecture, and it has the simplest function: storing bits. Because of the large and growing number of data types, encryption and compression formats, and DRM schemes, there is no way to specify in advance the format of data to be stored within the infrastructure. The only certain thing that can be said of digital data created in a decade's time is that it will, by definition, be made of bits.

Therefore, the principal function of a Repository is simply that it associate a unique identifier (ID) with a string of bits, and that it return or otherwise act on the bits connected to that ID whenever it receives an authorized request from a Gateway. The stored bits can be encrypted or not, compressed or not, with imbedded DRM or not, and so on. The Repository is never accessed by the Interfaces that serve patrons directly, and may not even be directly accessed by Collections that have arranged to preserve the material, depending on the access controls put in place at the Gateway layer.

Repositories are not required to be connected to one another, or to know about one another's existence in any way. There can be several Repositories holding the same material, in order to ensure the availability of backup copies. Repositories are also not required to be directly connected to any network. In the case of secure Repositories, they may be protected by an "air gap," where they exist as unconnected stand-alone machines, or on a local network unconnected to the Internet. Some repositories may be owned and operated by libraries, some by rights holders to hold their own materials, and some by third parties.

Repositories may—but are not required to—offer additional services, such as redundancy, versioning, check-summing of content, and so on. Repositories do not need to contain metadata about the data they store, other than the address. It is the functions of Collections and Gateways to maintain and associate metadata about the data stored in the Repositories.



Choices about who owns and operates a Repository, and what functions other than pure storage and retrieval it offers, are to be arranged by the original holder of the digital data and the collecting institution at the time of ingest.

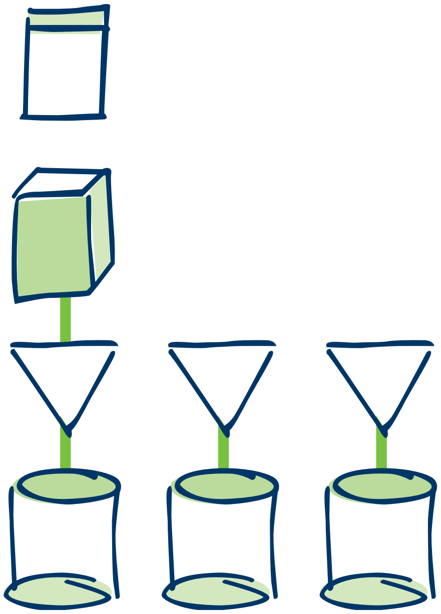
The Repository holds the canonical version of the digital content. It is therefore the likeliest target for various forms of attack, whether designed to copy, alter, or destroy digital content, and needs to be the most secured layer of the stack. Given these constraints, the Architecture Group believes that the Repository should be intentionally "stupid," and should ideally perform only a very small number of functions, in order to make it easier to both secure and to audit.

Gateways

A Gateway is a broker between Collections and Repositories. A Gateway takes requests from a Collection, validates it and, provided it is valid, passes the request to the appropriate Repository. When the Repository replies, usually with a string of bits it has retrieved, the Gateway takes the bits and returns them to the requesting Collection.

A Gateway is a translation layer, which bridges the purely digital view of the Repositories ("Here is an address for some bits, and here are the bits.") and the much richer and more human view of the Collections ("This is a digital photo of the Earth, in JPEG format."). Requests made from the Collection layer for particular pieces of material are translated by the Gateway into simple requests for the string of bits associated with a particular digital ID.

A Gateway is also a potential barrier. In cases where the Repository is storing commercially valuable intellectual property, a Gateway acts as a proxy, where any request to a particular Repository must pass through a particular Gateway, thus allowing for careful logging and auditing of requests, as well as providing an additional layer of authentication. A Gateway might have to authorize itself to a Repository. A Gateway might be required to authorize requests from Collections.



Gateways

(Green lines indicate connections to layers above and below)

A Gateway can be the forward edge of "air gap" repositories, where requests for validation pass off a network to disconnected storage. This means that material can be preserved in a Repository without ever being made available in real time.

Gateways may, but are not required to, offer additional functions, such as load balancing or maintaining reference counts of which Collections point to which Repositories. Like Repositories, Gateways are critical to providing security for the content stored within the system, so their functions should be minimally defined.

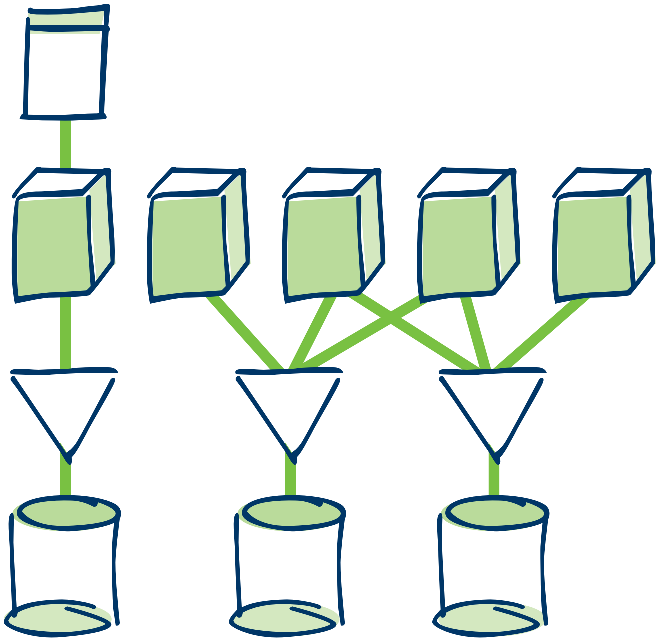
Collections

A Collection is a set of pointers and associated metadata for digital material. These materials can be related by theme, subject, or other semantic grouping—Shaker design, Virginia history, the 1930s, whatever. Collections are the site of both judgment—"These materials are interesting or valuable"—and stewardship—"I will see to it that these materials are preserved, not merely stored."

A Collection might point to several Gateways. A Gateway might be accessed by several Collections. Collections can cache or store copies, depending on agreement with rights holders. A Collection might have to authorize itself to a Gateway. A Collection might be required to authorize requests from Interfaces.

Collections are responsible for pre-ingest of digital material through creation of an archival information package (AIP, from the OAIS model), and again from creation of a dissemination information package (DIP) through dissemination to an Interface.

A Collection is concerned with semantics, not storage. It contains information about photos, movies, e-mail correspondence, not about the bits that make up those digital objects. It keeps rich metadata that describes an object (who, what, where, when, why, whence). The Collection is also the layer where the determination is made, in concert with rights holders, concerning who can use the content, how, and under



Collections

(Green lines indicate connections to layers above and below)

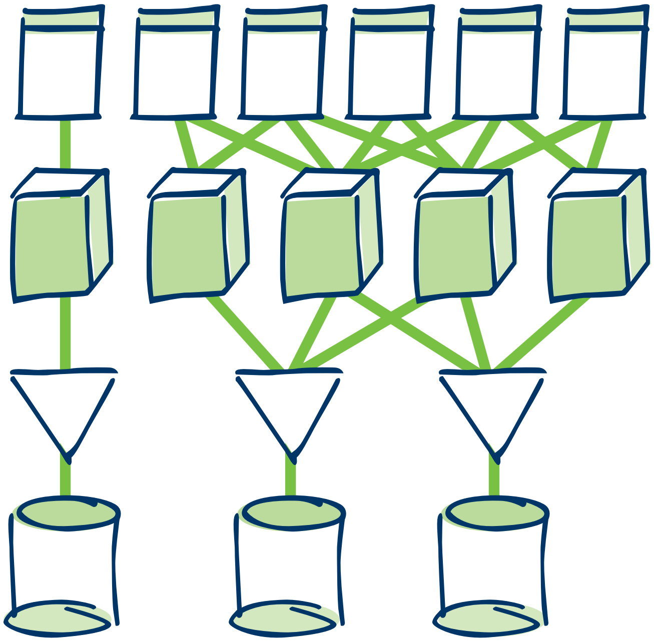
what terms and conditions. The Collection records key decisions about what preservation strategies will apply to the content (emulation, migration, etc.) and how the content should be managed and made available to end users and end user services. Who can access this content, under what terms? What software can interpret these bits, what software must not interpret these bits? How is “archival copy” or “best edition” defined? and so on.

A Collection is also where ingest of digital material takes place. Operators of Collections make their own judgment about what is interesting or valuable. The only requirement, if they are to participate in digital preservation, is stewardship of the materials they point to.

The minimal role for a Collection is to present a related set of references to authorized Interfaces, and to know which Gateways can supply the data referred to. However, unlike Gateways and Repositories, whose minimal functions are narrowly defined in order to ensure ease of implementation, a Collection will almost never exist solely for its minimal functions. Instead, by separating out the technical challenges of storage and retrieval, the technical architecture frees the creators of Collections to do the complicated work of interpreting, valuing, and protecting digital material, without requiring them to take on all the technical challenges of operating and upgrading Repositories and Gateways.

Interfaces

An Interface is whatever patrons access legally distributable content through. Interfaces are the upper edge of the infrastructure and have minimal requirements, other than a requirement not to violate the terms of use for a particular piece of digital material, as set by a Collection. The “American Memory” site at the Library of Congress, for example, is an Interface. Interfaces could also be kiosks, Web interfaces, phone trees, printouts, or other means of display not yet invented or imagined.



Interfaces

(Green lines indicate connections to layers above and below)

An Interface might point to more than one Collection. An Interface draws some but not all of its material from Collections. An Interface can cache or store material, provided it does not violate the terms of a Collection.

Connections and Protocols

The view of the infrastructure presented above has focused on the four components of the architecture. The other critical element is the protocols that connect these components together. The protocols that allow the different layers to communicate are more important than the actual layers themselves.

The design constraints imposed by modularity, and the goal of surviving the first migration, make the careful design and implementation of the protocols themselves far more important than any particular instantiation of the infrastructure. The hardware and software that make up the individual layers is designed to be heterogeneous, and to be deployed and upgraded piecemeal. In order to allow for this flexibility, however, the protocols will all have to exist in a Version 1.0 form early on and will have to be upgraded very slowly and carefully, with great importance placed on backward compatibility.

As with the technical architecture as a whole, we began with a set of principles to guide us in designing the connections between these layers:

Each Connection Should Be Defined as a Protocol

Though each layer will encapsulate a great deal of complexity, the conversations between layers should all be defined as protocols. A protocol is a definition of the ways in which two pieces of software should communicate with one another and exists separate from any particular instantiation. As an example, the transport protocol HTTP (hypertext transport protocol) is used by every Web server and Web browser in existence, but the definition of HTTP exists separately from any of these versions. (By contrast, the API [application programming interface] specifies the rules for a particular piece of hardware or software.)

By defining the essential conversations within the infrastructure as protocols, programmers and engineers will be able to build or adapt interoperable hardware and software without having to coordinate with a central authority, or with one another. This will improve the infrastructures' heterogeneity and ability to be assembled and upgraded piecemeal.

Each Protocol Should Be Minimally Defined

The temptation when designing a new infrastructure is to specify an enormous number of possible behaviors within the protocols. The Architecture Group set as a design goal simplicity in the required aspects of the protocols, with complex behaviors being added to the infrastructure optionally and where needed. The creation of a general and flexible infrastructure for ingesting and preserving digital materials would have enormous advantages outside the national digital preservation infrastructure itself, analogous to the advantages created by the Machine Readable Catalog (MARC) record. Therefore, in order to make the technical architecture as easy to adopt, even by institutions and committed individuals outside the national preservation infrastructure, the protocols should be as simple and easily implemented as possible.

Each Layer Should Allow for Potential Access Control

In order to allow collecting institutions the freedom to craft various approaches toward the twin questions of preservation and access, each layer of the infrastructure should be at least potentially able to implement access controls, whether by username and password, or Kerebos ticket, or some other security system not yet invented or imagined. There can be Collections that only serve material to authorized Interfaces, as well as Collections not connected to any Interface at all. There can be Gateways that only connect to a single Collection, and then only for the purposes of auditing the preserved materials. Likewise, there can be Repositories that only connect to a single Gateway, and then only for the purposes of auditing the preserved materials.

Dark Archives Will Be Triply Access-Controlled

Because this architecture is meant to preserve digital material across the entire range of commercial value and legal encumbrance, from freely available public domain material to material of extreme historical sensitivity or commercial value, such as personal papers or digital films, the infrastructure must allow for “dark archive”–style preservation. A dark archive function is achieved within this version of the technical architecture by placing material in an authorized Repository, which speaks to one and only one authorized Gateway, possibly across an air gap. That Gateway in turn would speak to one and only one authorized Collection, and that Collection in turn would speak to no Interfaces whatsoever.

No Horizontal Connections Will Be Required

In addition to the obvious vertical connections between Repository and Gateway and so on, many diagonal connections within the infrastructure can be created as well. Many Interfaces can talk to many Collections, many Collections can talk to many Gateways, etc. However, horizontal connections are different. There is nothing in the preservation architecture that forbids horizontal connections; indeed, we imagine that Collections in particular will find some horizontal connectivity desirable.

However, it is critical that no horizontal connections be *required*. One of the most significant problems that arise in rapidly growing infrastructure design is handling scale. There are many ways of designing networks that work well with a few hundred nodes but break for networks of a few million, and in the present case, requiring horizontal connections such as requiring all Collections to know about one another risks having the number of required connections grow insupportably vast as the infrastructure as a whole grew even moderately large.

Guided by these principles, the Architecture Group detailed some assumptions about the four possible interlayer connections: Patron and Interface, Interface and Collection, Collection and Gateway, and Gateway and Repository.

Patron and Interface

The protocol governing the transfer of digital material from Interfaces to the public is subject only to negative definitions: The Interface must not offer patrons materials in a way not in keeping with the rules for that material established by the Collection. Thus, if an Interface provides access to digital photographs whose conditions of access insist that they can only be shown in medium-quality JPEG format, the Interface may not offer high-quality TIFFs of the same materials.

Beyond these essential controls—if an Interface makes material from a Collection publicly accessible, it must abide by the Collection's rules for that material—the conversation between an institution hosting an Interface and their patrons is outside the scope of the national preservation infrastructure. This is done in part to allow institutions participating in the national preservation infrastructure to focus on issues of collection management and long-term storage (though we understand that a number of institutions will operate both Collections and Interfaces), and in part because the site of the user-interface is typically where a great deal of effort is put into innovation. The national preservation infrastructure is not about the nature of institutional presentation, so we have designed the preservation architecture to allow for the maximum amount of innovation in that area, while creating the fewest demands on the institutions running Collections, Gateways and Repositories.

Interface and Collection

An Interface is anything that offers the materials held in a Collection to the public, in any way. The protocols that govern the conversation between an Interface and a Collection must allow the Interface to present the Collection with a request for digital materials, along with any sort of authorization credentials if required. It must also allow the Collection to reply to the Interface with whatever materials the Collection has that match the description and authorization request from the Interface. The protocol must also support legally binding assertions by the Collection to the Interface about the terms and conditions for the use of those materials and must allow the Interface to offer legally binding assent to those terms.

Collection and Gateway

A Collection maintains the metadata surrounding a digital object—creator, date, format, playback requirements, and so on—and associates this metadata with a particular digital object. The protocols that govern the conversation between a Collection and a Gateway must allow the Collection to ask for a certain operation to be performed on a digital object referenced by the Gateway (anything from "Give me the following digital photo" to "Please run a checksum on the following digital photo to ensure it is still fit for use"), along with any sort of authorization credentials if required. The protocol must also allow the Gateway to present the Collection with an answer to its request (anything from "Here are the bits that make up the file you asked for" to "You are not authorized to access that content"). The protocol must also support legally binding assertions by the Gateway to the Collection about the terms and conditions for access to those materials and must allow the Collection to offer legally binding assent to those terms.

Gateway and Repository

A Gateway translates from human-form requests made by a Collection—"Give me this photo"—to the underlying digital location in a Repository—"Give me the bits associated with this ID." The protocols that govern the conversation between a Gateway and a Repository must allow the Gateway to ask for a certain operation to be performed on a digital object referenced by the Repository (anything from "Give me the bits associated with the following ID" to "Please run a checksum on the following ID to ensure it is still fit for use"), along with any sort of authorization credentials if required. The protocol must also allow the Repository to present the Gateway with an answer to its request (anything from "Here are the bits you asked for" to "You are not authorized to access that content"). The protocol must also support legally binding assertions by the Repository to the Gateway about the terms and conditions for access to those materials and must allow the Gateway to offer legally binding assent to those terms.

Protocols and Certified Preservation

One notable aspect of the preservation architecture is that, although it is designed for preserving digital materials, there is no "Preservation Layer." The assumption of the Architecture Group was that preservation is not simply a matter of long-term storage, but of institutional commitment to ingest and keep digital materials fit for use. Therefore, while the individual pieces of the infrastructure concern themselves with functions such as storage, metadata management, or access, it is the infrastructure as a whole that will be used to provide real long-term preservation.

The Architecture Group also assumed that many institutions, governmental, academic and commercial, would be involved in the national digital preservation infrastructure. We imagined some way of certifying those institutions that adopt a set of best practices and are willing to affirm that they will preserve certain materials. Therefore, the definition of preservation within the architecture is threefold.

Digital material is considered to be preserved if it is stored in a certified Repository, referenced by a certified Gateway, and listed in a certified Collection. Note that the Architecture Group anticipates that there will be material within the infrastructure that can be regarded as being stored but not preserved, because it lacks the positive institutional affirmation of preservation.

Open Architectural Issues

Though the Architecture Group believes the preservation architecture represents a good starting point in designing and testing various implementations, it is important to emphasize that the architecture is a conceptual framework at this point and lacks the myriad detail-oriented decisions that it will take to make it a reality. In particular, the architecture presents some obvious open issues that it will be critical to grapple with in any test implementation. Among these issues, three stand out:

Separating Semantics from Presentation Is Hard

The architecture imagines a clear split between preserving and presenting digital material. The Collection layer exists as a kind of a hinge between these two activities, with the creation of metadata and the storage of the materials themselves existing at the Collection layer and below, while the actual presentation of those materials to patrons, in the form of a Web page, PDA screen, or some other type of interface not yet imagined, is the job of the Interface layer, and therefore the job of the institution hosting the Interface.

The history of the Web indicates how difficult it is to separate information *about* the material—semantics—from display *of* the material—presentation. The Web was designed to use HTML (Hypertext Markup Language) as a markup language for digital content, and though HTML was descended from SGML (Standard Generalized Markup Language), SGML insisted that descriptions of the contents of a document be separated from instructions for displaying that same document. Despite this mandate, however, HTML quickly became a mixed-use language, with information about the document and presentation instructions to the browser being directly imbedded in the same file. HTML became mixed use in spite of years of theoretical work concerning why such mixed use was a bad idea, in part because the early Web was implemented largely by amateurs for whom cross-platform visual presentation was the paramount virtue of the Web.

While generalizing from such a singular case always carries some dangers, the Architecture Group assumes that rigorous separation of semantics from presentation is at odds with the simplicity required for mass adoption. As both are goals of the architecture, further work is needed to determine a range of possible compromises.

It Is Difficult to Store Dynamic Material

The preservation architecture is centered on file-oriented materials—movies, photos, audio clips, software, and other digital objects that can be stored in a stable form. There are, however, a number of valuable and significant kinds of material, from mailing lists and context-sensitive databases to online collaborative or gaming environments, that only really make sense when they're being used, because they change to reflect user-input. While storing a digital snapshot of these materials in a Repository has some obvious value, the concept of a Best Edition is difficult to ascertain for this sort of dynamic material, and in many ways, the most obvious place for a canonical copy to live is at the Interface layer, not the Repository layer.

As the amount and importance of dynamic digital materials are growing, further research will be needed to determine strategies for preserving these materials within the NDIIPP.

Metadata Must Be Preserved as Well

Metadata is data about data, the ocean of facts and judgments that surround a digital object. It is both factual data—"This file is in JPEG format"—and aesthetic—"This picture is better than that one, in the judgment of the curators." While the curators, archivists, librarians, and other professionals who deal with data treat metadata as a separate aspect of a particular piece of digital material, metadata is data as well. It is made up of bits, and it has its own metadata, such as particular formats associated with it (a curator's notes might be in WordPerfect format, for example). Furthermore, metadata is extremely valuable, both because it is required to decode a digital object, but also because it represents human judgment about value and context. Metadata thus presents all the same dilemmas of preservation as any other kind of data.

This creates the risk of a kind of data cascade, where metadata about a preserved digital object is itself stored in a Repository, and therefore requires meta-metadata, which can itself be preserved, ad infinitum. To keep this from happening, some set of best practices about how metadata is preserved needs to be defined.

There are of course other open architectural issues, some very low level, like understanding the behavior of materials such as DVDs and magnetic tape over periods of decades, and others relating to the technology only peripherally, such as how best to write a contract between collecting institutions and the holders of digital material. A chief goal of the next phase of NDIIPP work will be to test the proposed architecture and several alternatives in an attempt to first surface, and then solve, these issues.

Conclusions

Though the Architecture Group's work is still in its early stages, several conclusions can be drawn from a combination of the initial assumptions and subsequent work on the proposed architecture outlined here. The most important of these conclusions are listed here. In the next phase of the NDIIPP, considerable work will need to be undertaken to treat these assumptions as hypotheses and test them in a variety of real-world settings.

Hardware Stores, Institutions Preserve

The first and most important conclusion reached by the Architecture Group is that hardware stores, but institutions preserve. While we are provisionally pleased both with the architectural work to date and with the feedback we have gotten from the technical representatives of intellectual property rights holders, we also acknowledge the impossibility of solving the digital preservation problem with technology alone. Issues such as how to present current material on new kinds of devices, how to convert existing files and formats for use on new operating systems and software, and even what constitutes a "Best Edition" or whether material is fit for use all require human judgment, and the requirement to periodically verify that stored material is actually preserved in a useful way clearly requires some form of ongoing maintenance.

Preservation is a process, not a product, and the architecture provides tools for that process, but not a full solution, because without institutions that sign up for stewardship of digital materials as part of their mission, all the technology in the world will not solve the problem. The architecture is best thought of as a set of tools that will allow institutions to preserve the material they cherish, in ways that suit their mandate.

Preservation and Access

The next conclusion is that any infrastructure must treat the preservation and accessibility of digital materials as separate questions, and must be able to allow for the present-day preservation of digital materials whose date of public access is unspecified and may lie very far in the future. The legal and economic climate for production and distribution of digital content is in flux, and holders of commercially valuable intellectual property will be resistant to participating in an infrastructure that makes immediate public access a requirement for long-term preservation. Therefore, making preservation possible without requiring immediate public access is an essential strategy for ingesting at-risk commercially valuable digital material in the short term, so as to be able to preserve it until such a time as the legal and economic framework surrounding such works is clear.

At the same time, however, the infrastructure must also be flexible enough to hold and make immediately accessible those works that are either in the public domain or are licensed in such a way as to obviate commercial claims. The infrastructure must therefore be flexible enough to preserve any digital material, from the immediately publicly accessible to the commercially valuable and legally protected. But in making this preservation possible, it cannot adopt a one-size-fits-all attitude toward public accessibility of the wide range of materials to be collected.

The Costs of the System Are Mainly Human, Not Technological

Although Moore's Law, explaining the rise of processor speed, is by far the best known example of quadratic improvement in technology, storage is growing at similar rates, with the density of storage doubling roughly annually. The related effect is that storage-per-constant-dollar is growing on a similar curve. This is advantageous, as storage costs can be expected to fall steadily, at least during the period of initial construction.

However, storage alone is not the issue. Preservation requires more than simply filing bits away somewhere. Examples abound of data that have been successfully stored on tape but which cannot now be recovered, because we lack the hardware or software to play the tapes back. Furthermore, secure preservation is even more complex and requires even more human attention, as the arms race of offensive and defensive security techniques means that security is an ongoing process and not a single product. Storage is cheap; preservation is expensive; security is potentially very expensive, depending on the degree of hardening required.

Preservation requires human stewardship, in order to verify that the material stored remains fit for use. Therefore the overall costs of the infrastructure are going to be principally human and ongoing, rather than technological and upfront. While the Architecture Group feels that the architecture offers a good starting point for real-world testing, we want to caution that, although long-term digital preservation is a problem caused in many ways by technology, it is not a problem whose solution is solely or even primarily technological. Thus care must be taken, in the early days of testing and deploying the digital preservation infrastructure, to emphasize the human requirements alongside the technological ones.

APPENDIX 10

Criteria for Projects

Criteria for Projects

Portfolio Criteria

There are near- and long-term activities and investments, in addition to a number of initiatives already under way in many contexts, that can be mobilized as part of a nationwide system. A suite of projects and investments is required both to leverage federal investments effectively and to build functioning systems that will be positioned to take advantage of technological advances as these become appropriate. Collectively, these projects respond both to the design criteria and to the investment criteria. Ten criteria, embodying both investment and technology values, are detailed below.

Not every experiment will meet all 10 criteria, but collectively the portfolio of projects should meet these requirements. The goals of the portfolio overall are:

- to initiate new projects or leverage existing work in key areas that identify and capture at-risk materials,
- to explore technical issues associated with long-term preservation of digital content, including a variety of formats and types of content,
- to examine relevant copyright issues, and
- to build the network of collaborations to meet critical needs of the National Digital Information Infrastructure and Preservation Program (NDIIPP) and maximize use of public funds.

Criterion 1: Does It Preserve Diverse or At-Risk Media?

Rescuing endangered material is one of the motivations for developing a national strategy for long-term preservation of digital content. Discussions with multiple stakeholder groups have reinforced the sense that loss of potentially significant material is imminent. Additionally, the range of media types is extensive, with different complexities posed by different formats. Thus the content that is employed in the tests should either test the diversity of potential types or examine ways that vulnerable materials—orphan collections, ephemera, and so forth—might be identified, captured, collected, and preserved.

Criterion 2: Does It Test Collaborative Network Models?

The scale, complexity, and diversity of digital content formats and ownership regimes mean that collaboration among a range of partners and organizations is a key element of the proposed national strategy. There is also the recognition that the scope of preservation needs reach beyond national boundaries. These collaborations may take many forms—maintaining a certified repository, developing shared metadata, partitioning responsibility for maintaining the integrity of the repository from the issues associated with collection management, and so forth—and involve different national or international partners—professional associations, university and research libraries, nonprofit institutions, commercial enterprises, other public agencies, among others. Given the importance of collaboration to the overall strategy, and given the range of forms in which that collaboration might be expressed, the Library proposes to examine different collaborative structures, relationships, and mechanisms through the various projects.

Criterion 3: Is There Sufficient Capacity to Achieve Satisfactory Execution of the Project?

The timeliness of the practical application and modeling projects requires committed participants who are ready to engage and have identified the requirements, including resources, objectives, and goals of the project. Meetings and workshops with a variety of commercial, federal, and academic stakeholders throughout 2001 and early 2002 provided an understanding of the array of interested, intelligent thought and practice that is evolving across the landscape of preservation activities. The planning team learned that there are existing technologies and projects that can be leveraged to move into the project phase of the NDIIPP.

Criterion 4: Does It Address Pertinent Copyright Concerns?

Managing the intellectual property rights associated with digital works is a challenging topic; a white paper that discusses relevant issues has been appended (Appendix 6). The proposed preservation architecture offers a way in which diverse digital property rights concerns might be addressed, although development of the architecture is neither predicated on resolving these issues immediately nor on selecting now among various proposed and to-be-proposed digital rights management and information security technologies. However, since copyright management is a feature of at least some of the collections and processes, proposed projects must identify relevant copyright issues and should begin to craft solutions to at least one aspect, such as so-called best edition, deposit, or ingest.

Criterion 5: Does It Advance the Development of Standards and Best Practices?

Different elements of the preservation architecture are likely to require different approaches to technical and organizational consistency, coherence, and interoperability. Historically, the conceptual tools for handling these issues include standards, protocols, and best practices. Some features, such as naming, may require formal standards, whereas other features—ingest, for one—might be best handled as protocols. Still other features, such as collection development, might be handled under the rubric of best practices. Given the complexity of the preservation challenge, as well as the range of stakeholders and formats, different circumstances will require different approaches. Thus, the portfolio of projects should collectively test a range of approaches to different technical and organizational contexts, examine the conditions under which the different approaches should be employed, and develop sample representations of each, as appropriate.

Criterion 6: Does It Help Clarify Collection Selection Issues?

Not only is the volume of digital information immense but it is also highly heterogeneous and subject to complex layers of rights and conditions of use. Therefore, partnerships among a broad range of institutions are required as well as specialization among the collections policies of the cooperating entities. Historically, the Library and archival systems in the United States and abroad have evolved formal and informal means of ensuring necessary redundancy among holdings as well as specialization where location, expertise, or access mean that a given institution may be positioned to develop uniquely specialized collections. In this regard, the Library has played an important coordinating role in developing such tools as the National Union Catalog of Manuscript Collections (NUCMC), which enables researchers to locate collections that are physically distributed yet topically allied.

Similar strategies will be required in the digital environment to ensure that the collective scope of the collections is sufficiently redundant to ensure safety, sufficiently broad, and sufficiently deep to satisfy information and research needs now and into the future. Additional issues in defining the scope of individual collections arise from technical issues. For example, what constitutes a Web site, the surface Web or the databases that may support it? Similar questions can be raised about the boundary conditions of other digital resources.

Criterion 7: Does It Test the Digital Preservation Architecture?

The previously discussed four-layer preservation architecture is a critical element of the proposed national preservation strategy. It shows the relationships among the technical layers (Repository, Gateway, Collection, Interface) and suggests the range of organizations that might undertake responsibility for or provide services to different parts of the overall architecture. Although this architecture is consistent with current thinking and has been employed in other contexts, it has not been deployed in a distributed library-archives system such as envisioned here.

Therefore, testing the elements of the architecture—Do they work from a technological perspective? How are the technical elements imbedded in organizations and organizational relationships? What are constraining factors? and so forth—is a critical dimension of the portfolio. Many projects might be proposed that would meet this criteria: for example, testing individual repository architectures; creating and harvesting metadata at the collection level; or developing auditing and monitoring procedures at all levels. It is also important to test alternative preservation architectures to probe for inadequacies or weaknesses in the proposed model.

Criterion 8: Does It Test Scalability?

Digital content is characterized by its immense volume as well as its heterogeneity; both scale and heterogeneity are likely to increase in the future. Indeed, ingest is already a significant challenge in existing systems. Scalability—the ability of systems to handle very large quantities of information or to increase in size in order to handle large quantities of information—is a key feature of future systems and will be an important dimension of the portfolio of activities undertaken in the next phase.

Criterion 9: Does It Test Sustainability?

Getting an experimental system up and running is a challenge; maintaining it over time is a further challenge. Indeed, prior consultations indicate that attention to sustainability early in the research and development process is a key element of successfully operating systems. Sustainability may be understood as having many dimensions: economic sustainability may be related to cost recovery and potential business models; technical sustainability encompasses a range of issues from extensibility (which enables a system to evolve as new and more efficient technologies become available) to metadata definition. Metadata is a particular challenge because creating it can be labor-intensive, yet metadata is essential to the efficient management of the system as well as to long-term and satisfactory use of the content it describes. Finally, sustainability has an organizational dimension: Which collections will be the responsibility of which institution? Is there sufficient redundancy in the system (yet not too much to be wasteful of resources)? Are there tools and services in place at different levels of the preservation architecture to provide sufficient information for the management of the system?

Criterion 10: Does It Leverage Other Efforts?

The process of reaching out to multiple communities in industry, research, higher education, and nonprofits showed that there is awareness as well as projects under way in television (broadcast, cable, commercial, and public), radio, motion picture, and publishing.

There is also fairly widespread concern over archiving the World Wide Web, although the challenges of preserving the Web are substantial. Indeed, the Library has itself initiated a pilot effort in Web capture and preservation, as have other organizations.

There is also some latitude in the degree of readiness evidenced by these separate efforts, which reflects the complexities of preservation as well as the challenges of knitting them into a national strategy. Thus, members of the Recording Industry Association of America (RIAA) engage in an active discussion of the technical issues; representatives of the various record labels have participated actively in consultations with the Library; and the major newspapers have digitized or are in the process of digitizing their archives. Yet proprietary concerns inhibit the degree to which commercially sensitive information can be easily shared. Similarly, Stanford University, MIT, and others have begun to explore technical responses and have initial systems up and running.

These uncoordinated efforts offer resources that may potentially be mobilized into a national system. They are also test beds for technologies and solutions. Given NDIIPP's focus on decentralization and collaboration across interested communities, priority will also be assigned within the portfolio of projects to leveraging existing projects that are compatible with NDIIPP goals and values, where an investment of federal funds is likely to accelerate progress. This represents a prudent and effective investment of federal resources to catalyze a public-private system for the national good.