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

Commissioned for and sponsored by the National Digital Information Infrastructure and Preservation Program, Library of Congress

April 2002

Council on Library and Information Resources
Washington, D.C.

and

Library of Congress

About the National Digital Information Infrastructure and Preservation Program

The mission of the National Digital Information Infrastructure and Preservation Program is to develop a national strategy to collect, archive, and preserve the burgeoning amounts of digital content, especially materials that are created only in digital formats, for current and future generations.

ISBN 1-887334-91-2

Copublished by:

Council on Library and Information Resources
1755 Massachusetts Avenue, NW, Suite 500
Washington, DC 20036
Web site at http://www.clir.org

and

Library of Congress
101 Independence Avenue, SE
Washington, DC 20540
Web site at http://www.loc.gov

Additional copies are available for $20 per copy and may be ordered through CLIR's Web site.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

Contents

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 9. (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 inter-

ested 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 for-

mats 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

[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.

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

[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.

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

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

- 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).

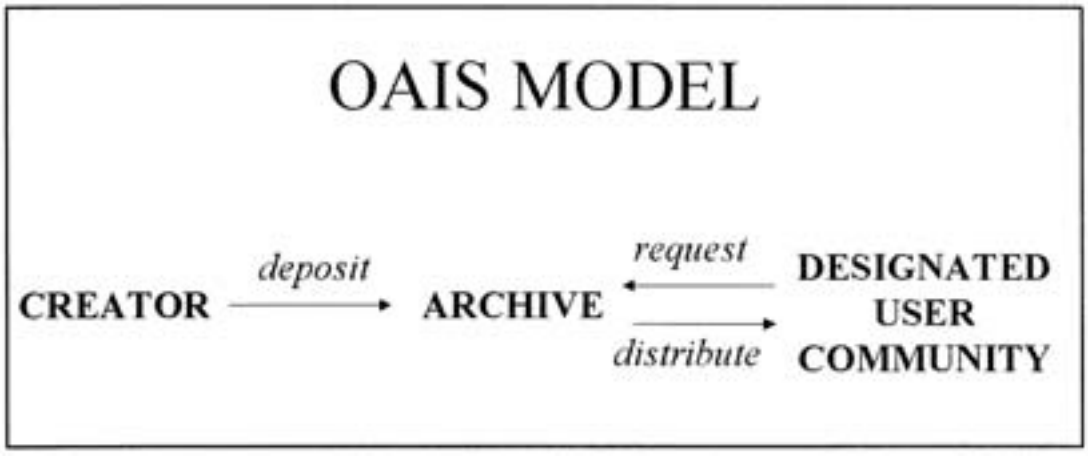

Fig. 1. OAIS model of players and roles

[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.

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 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 ma-

terial 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

---

[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.

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-

---

[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.

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).

[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.

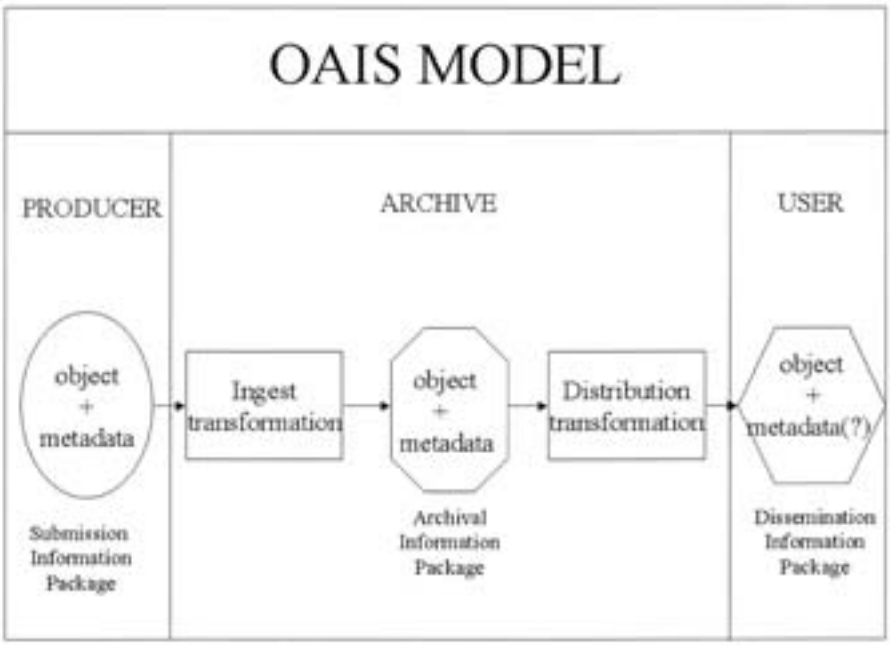

Fig. 2. Information packages in the OAIS model

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 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.

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. 34-35.)

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 re-recorded 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 online, 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 digi-

tal 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. Mi-

crosoft 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 1.

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).

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

Fig. 1. Sampling of e-book publishers

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 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 genera-

[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/.

tions 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 ar-

chiving 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

---

[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.

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

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

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.

---

[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.

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

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

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

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 44.1 kHz 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 *Au-*

dible 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 (44.1 kHz, 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 kHz 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 kHz.

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 re-

dundancy. 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 Yorke*r, 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 incompati-

ble 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 at-

tached 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 sat-

isfy 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. *Television 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)

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 1.

Fig. 1. Effects of seeking directly to a match point on "Lunar Rover," courtesy of tight transcript to video alignment provided by automatic speech processing

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 2.

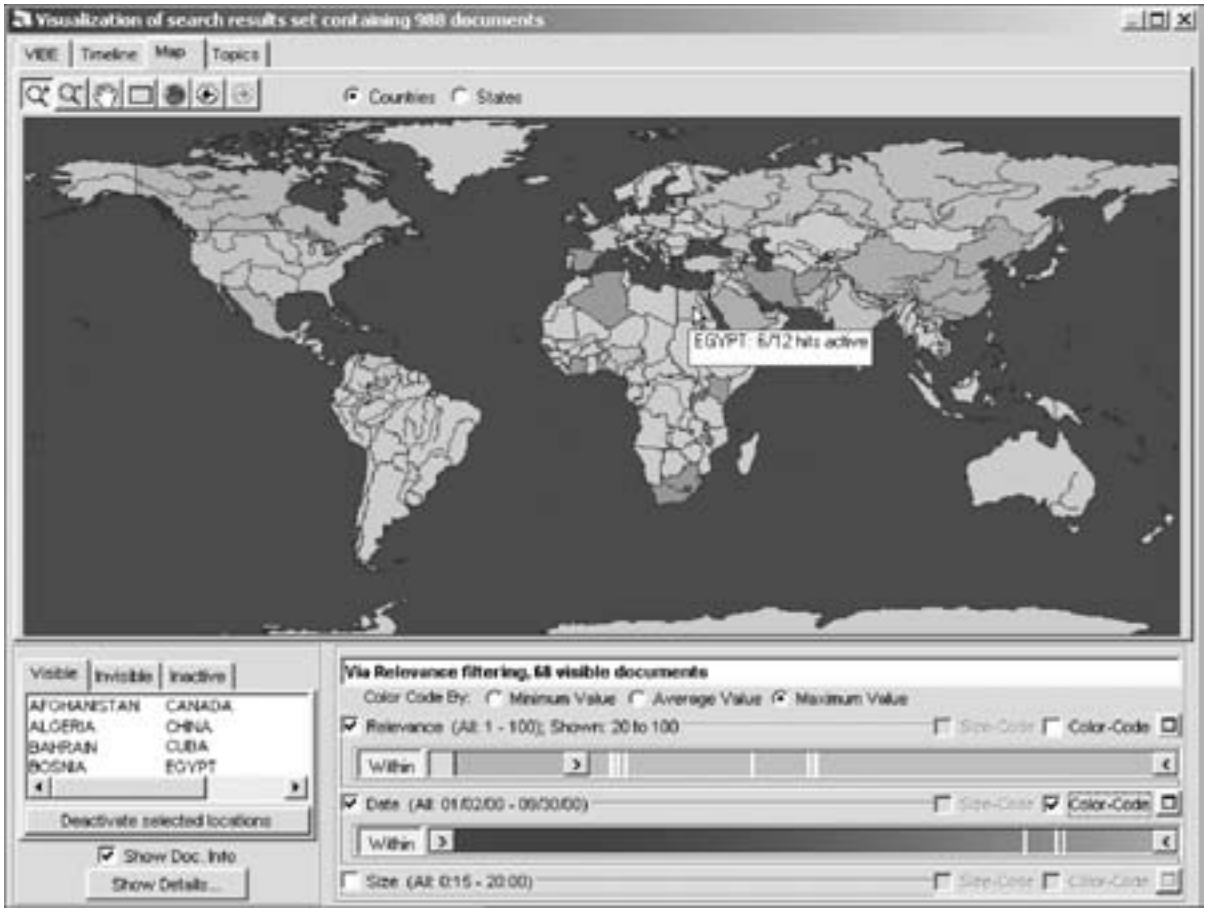

Fig. 2. Map visualization for results of "air crash" query, with dynamic query sliders for control and feedback

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 3, 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 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 3.

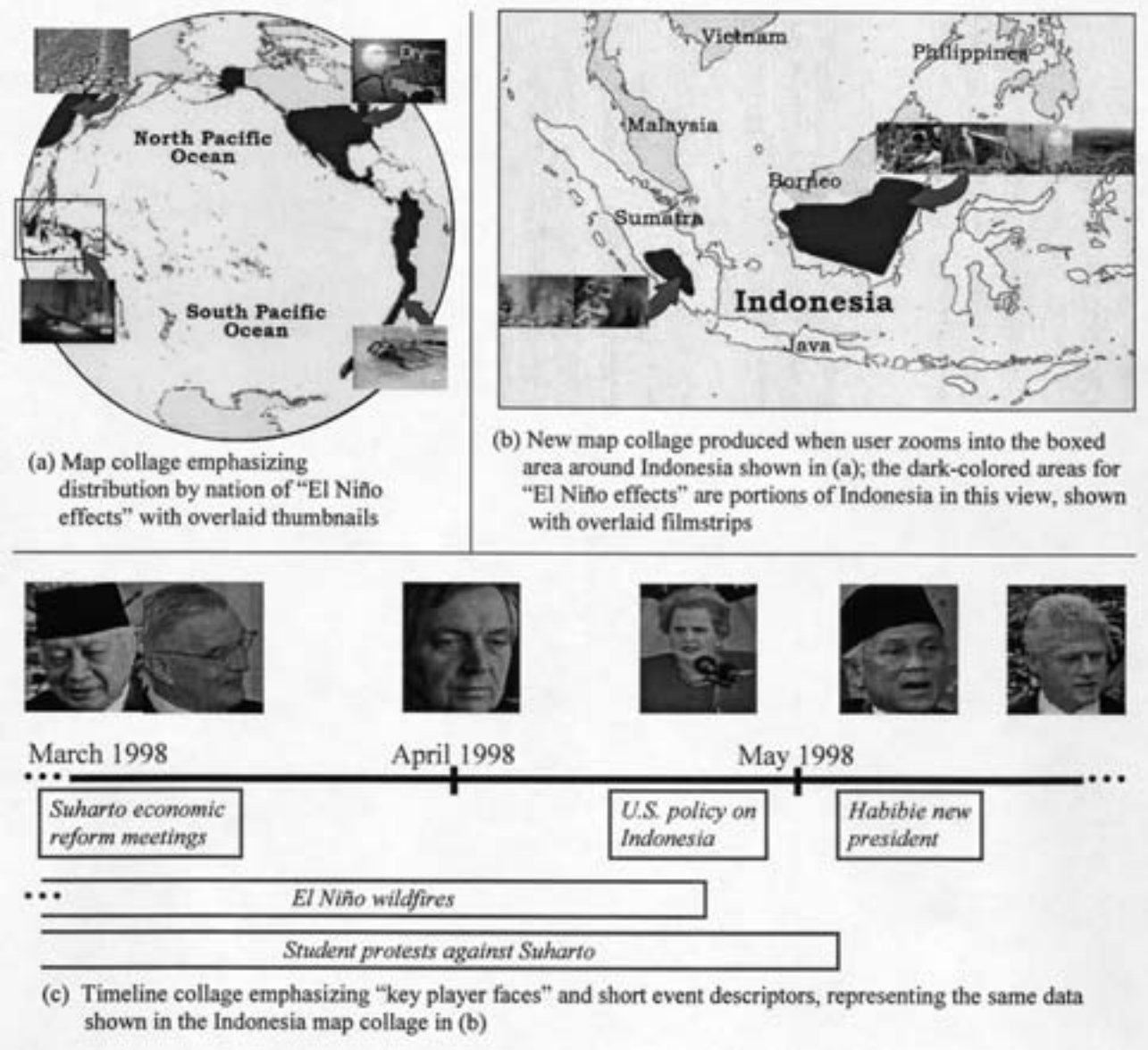

Fig. 3. Prototype of Informedia-II collage summaries built from video metadata

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 outmodes 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