Background
Summary of Results from Interviews and Essays

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 has chosen to open its investigations with a series of probes into six principal areas in which the Library faces collection management issues: large Web sites, electronic books, electronic journals, digitally recorded sound, digital film, and digital television. This paper summarizes what a series of interviews and essays, conducted and written during the late summer and early fall 2001, have told us about a complex and shifting landscape.

We have conducted formal, 30-minute interviews and had shorter conversations and e-mail exchanges with individuals who represent a range of interests and organizations across publishing, film, entertainment, news, electronic books, computer science, libraries, corporate research, non-profit organizations, professional and trade associations, and academe. Their names and primary affiliations are listed in Appendix 1. Most people talked to several concerns and formats; thus, we have abandoned efforts to characterize them exclusively by format (e-books or e-journals, Web sites, digital film, digital TV, digitally recorded sound), profession, or organization. (Note that corporate representatives frequently sit on the boards of non-profit and cultural organizations; thus their perspective is informed across communities.) This information was complemented by six "environmental scans," which are intended to provide a baseline of information for concerned groups outside of the library, preservation, and archival communities. Their intent has been 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 paragraphs, we summarize some of the key findings:

BORN DIGITAL VERSUS DIGITIZED

The scope of the effort was defined as material that is "born digital," that is, objects that have been created in digital form rather than works (or objects) that have been digitized or converted from analog to digital. This distinction 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 older entertainment products partly or wholly in digital form, as either new editions of older works or re-used 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," writes 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 in the recording industry agree that appropriate 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 itself, there are many sites offering digitized versions of print works, so archiving the Web itself can be seen as encompassing both “born digital” and “digitized” materials. One publishing executive argued that “being 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 goes to the concept of completeness and with that concept were associated notions of “copies,” “versions,” and other ideas critical to managing works and their associated rights.

THE SCOPE

In addition to the blurred distinction between “digitized” and “born digital,” the notion of scope reappeared at many levels, from the definition of the object to the scope of the effort. Several people in the library community and outside of it urged planners to consider the scope of the effort carefully, including what was selected for the collection (if it were even a single collection), the longevity of the collection (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 itself. 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. (Note that in the case of an interactive Web site, the range of potential rights holders extends well beyond those behind its creation.)

Several people in both the technical and 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 Library 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 that might not be reflected either 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 DVD releases have provided new life for outtakes and other associated production materials. The relative utility of material over the cultural life of a film or a performance changes, and the first public release does not necessarily capture all of the 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. It becomes very clear from the discussions of Web sites, e-books, e-journals, and digital television that boundaries are difficult to draw. 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 back to the Web for ancillary augmentation, and the advent of interactive television also invites new forms of multimedia that combine both resources built for the Web and those created for broadcasting in digital form. Moreover, what 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, all of which may be encoded in different formats. 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. As one person pointed out, this is related to the complex intellectual property considerations that surround digital information. Even in a format as carefully studied as electronic scholarly journals, creation and deposit can involve numerous stakeholders, and the number of interested parties multiplies in sound, television, and film, in which individuals and entities have traditionally had rights in the processes of creation and distribution. Indeed, Frank Romano points out that the e-books world is witnessing changes in traditional roles and functions in which writers can self-publish and thus become distributors and software companies can behave like publishers. Similar shifts and realignments can be seen in some metadata discussions, where, as Peter Lyman points out, 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, consensus around the basic issues, if not necessarily around solutions. The issues are technical obsolescence and standards, metadata, information security, and the overall architecture of the system. These elements 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 just making sure that bits survive but rather the preservation of the 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. As another technical expert described it: think of the degradation as similar to the way that a photograph ages. The image fades—unevenly—and 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, JPEG or MPEG format), the extent to which migrating the information introduces new errors if the data are resampled, and the implications of migrating formats for version control and integrity. When a digital work is migrated from one format to another (e.g., 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 not truly the artist's treatment of a work and it misrepresents the recording technology at an early stage?

At least one technical expert does not consider this to be a serious problem but acknowledges that the rules for the successive formats must be retained. On the other hand, the team from the WGBH Education Foundation notes that while standard archival practices call for refreshing the data through migration and emulation, these strategies may 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 organizations as well as in traditional libraries and has highlighted the importance of metadata as a management tool (see discussion below).

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, recreating the experience for future users implies access to the device on which the content was intended to be displayed.

Standards and Technical Obsolescence

The rapid obsolescence of some formats, as well as the plethora of standards, were widely considered to be barriers both to creation and to satisfactory preservation. Those who had opinions on open versus proprietary standards favored open standards 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 mark-up language, compression, and fonts.

Information Security

Before September 11, 2001, few people consulted had strong opinions on this topic, 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 the 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. It was recommended that discussion 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 addresses or the addresses of the 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 inhibited accountability.

Proposals for Storage Architecture

Those who addressed technical issues also tended to favor distributed rather than centralized systems, because they would accommodate a high degree of "local" variation within shared protocols. There were also calls for interoperability, which would enable information to be shared across platforms and among vendors. One publisher thought it was important that the Library do the development in-house without recourse to proprietary software and by employing commercially available tools because it would facilitate future upgrades to the system. Two architectural approaches were set forth: one for e-journals (see essay by Dale Flecker), which is fleshed out in some detail, and a more rudimentary one, which looked at the problem of preservation from a broad perspective in which the Library is one of many entities that might be involved. In general, there is discussion 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 notes, the technology is likely to be developed outside of the traditional library community by other interests. The Library 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 Library's imprimatur.

Metadata

Metadata, which is typically understood as "data about data," is simultaneously a standard, a management and access tool, and a feature of the system architecture. For example, whether the metadata is "bundled" into the "content" or is maintained separately is a question that is being discussed with respect to several formats and affects approaches to interoperability as well as system design. The team from Carnegie Mellon University argues persuasively for the importance of metadata to the management of the archive as well as for providing appropriate access. The essay 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, and 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:

> 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 essays addresses IPR at some level, with perhaps the most general discussion offered in Peter Lyman's essay 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 Library of Congress 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 for both bibliographic records and metadata, which enable scholars to find out 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 acknowledgement of the need to find a new 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 based on estimates of time during which the owner expected to extract the economic value. However, product cycles of re-use would complicate that approach.

Several people felt that existing laws were sufficient and what was required was appropriate enforcement. Others felt that there was a need to revisit and clarify what the law said, particularly with regard to international law, since the Web is an international phenomenon, and to fair use. As of this writing, terms such as “copy,” “publication,” “performance,” and “public figure,” which had had some consensus on their meaning were subject to discussion. Still others pointed to misperceptions that were clouding the discussion in several contradictory ways: people thought that information in digital form had both more value (those who tended to inflate the costs for permissions) and far less value (those who thought information should be free). 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 was becoming increasingly unwieldy. For example, in film, the multiplicity of rights and permissions that affect distribution and re-use of material had derailed educational projects because it was simply impossible to unravel the layers. Recorded sound has similar layers of rights, and Peter Lyman also elaborates on this point with some care in his essay on the Web. Finally, ambiguity over the law is itself becoming a barrier. Faculty members are wary of developing new coursework for online learning in an environment in which there is no consensus about appropriate conduct and the legal ramifications of their decisions are unknown.

Appendix 1: Individuals Consulted

Brindley, Lynne
British Library

Brin, David
Independent Author

Brown, John Seely
Xerox PARC

Carey, John
Professor, Columbia University

Crocker, Steve
Longitude Systems

Daley, Elizabeth
Annenberg Center
University of Southern California

DeMartino, Nicholas
American Film Institute

Eaton, Nancy
Pennsylvania State University

Franey, Colin
EMI

Elizabeth Frayzee
AOL/Time Warner

Garza, Carlos
The Recording Industry Association of America

Grey, James
Microsoft Bay Area Research Center

Hindman, James
American Film Institute

Kinder, Marsha
Annenberg Center
University of Southern California

Leones, Edrolfo
The Walt Disney Company

Mink, Allen
National Institute of Standards and Technology

Powell, Adam Clayton, III
The Freedom Forum

Rudick, Richard
John Wiley

Roper, Ray
Printing Industries of America

Schline, John
Penguin Putnam, Inc.

Weissman, Larry
Random House

Wickham, Woodward
MacArthur Foundation

Williams, Troy
Questia