<table>
<thead>
<tr>
<th>著者</th>
<th>ヘドストラム マリー</th>
<th>デジタル図書館</th>
</tr>
</thead>
<tbody>
<tr>
<td>言語</td>
<td>日本語</td>
<td>日本語</td>
</tr>
<tr>
<td>国名</td>
<td></td>
<td>日本</td>
</tr>
<tr>
<td>場所</td>
<td></td>
<td>図書館</td>
</tr>
<tr>
<td>時間</td>
<td></td>
<td>日本</td>
</tr>
<tr>
<td>検索</td>
<td></td>
<td>デジタル図書館</td>
</tr>
</tbody>
</table>
Digital Preservation: Problems and Prospects

Margaret Hedstrom
School of Information
University of Michigan

In an article that I wrote more than five years ago called, "Digital Preservation: A Time Bomb For Digital Libraries," I drew attention the problems that preservation would create for digital libraries if this issue were ignored [16]. Even though many preservation issues remain unresolved, I am pleased to report that digital preservation has received considerable attention during the last five years as a major obstacle to building sustainable digital libraries and as a critical area for research and development. In this paper, I will outline the main issues that long-term preservation requirements raise for digital libraries, describe some of the current models for digital archiving, and discuss some of the outstanding issues for research and development. I will present a broad overview of the issues rather than an in-depth analysis of specific models or research projects. There are simply too many issues are too much activity to give all of the issues and projects the attention that they deserve.

Digital preservation has become a common topic in digital library research and development for several reasons. First, digital libraries have gone through a transition from research and experimental projects to an important part of the infrastructure for research and teaching. In many scientific fields, research depends on access to persistent stores of digital information that are built and refined continuously. Consistent with the cumulative nature of scholarly research, journals that report research findings and that make references to previous studies constitute a continuous record of research and discovery. As scholarly communications have shifted rapidly from print-based journals to either hybrids of print and electronic journals or to exclusive publication in digital form, the need to preserve a comprehensive record of research and scholarly achievement has not diminished. Moreover, libraries, archives, and other organizations have made considerable investments in acquiring digital content and in converting older print-only materials into digital form to improve access to disparate sources that were difficult to locate and retrieve. According to the 1998-1999 statistics from the Association of Research Libraries, major research libraries in the United States spent between slightly less than three percent to more than twenty-one percent of their acquisition budgets on electronic journals and other digital resources [1]. Few aspects of contemporary society and culture are untouched by changes in the creation, distribution, and access to information. There have also been concerted efforts to raise awareness of the preservation issues associated with new information technologies out of concern over the survival of digital information to document this rapid shift in communications [9, 13, 26, 29].

The rapid acceptance of digital technologies and the growth of digital libraries created three primary motivations for investments in digital preservation research and program development. First, there is an increasing demand for continuing access to the resources that digital libraries make available. Once users become accustomed to accessing information on-line they do not want those resources to be removed or diminished. Second, there is an interest on the part of digital library developers to protect the investments made in digital resources, whether those investments are subscription fees paid to publishers for on-line content, the costs of initial data collection and preparation, or the costs of converting print and other
analog materials into digital form. Third, there is a concern about preserving digital communications for the future study of our present time and culture. This includes the content of digital documents that might be considered ephemeral as well as evidence of the impact of digital communications on many aspects of society.

Despite evidence of increasing concern about digital preservation, there are numerous technical, organizational, legal and economic barriers to a comprehensive infrastructure for protecting and preserving digital assets. The most familiar problems in digital preservation are media failure or deterioration and rapid changes in computer hardware and software that make older systems obsolete on a regular basis. Efforts to preserve digital information have always been challenged by the relative instability and short life of most digital storage media. Media failures and undetected deterioration of storage media remain a problem for digital preservation, but the issue of media longevity has moved into the background. There have been significant improvements in the quality and longevity of almost all digital storage media. Although there is no "permanent" digital storage medium that meets standards of longevity and durability established for "permanent paper" or microfilm, improvements in magnetic and optical media reduce the frequency at which digital materials must be copied to new media in order to prevent deterioration or loss. In some cases, transferring older digital information to new media brings additional advantages, such as increased media capacity and faster access which offset the costs of copying. Established repositories and most digital library designers accept the need for systematic maintenance of digital materials and periodic replacement or "refreshing" of the underlying storage media. Some recent research, discussed below, is developing automatic methods for detecting and repairing damage from media failure and deterioration [7, 10, 22]. Media deterioration and loss remain a problem when digital materials are not integrated into systematic management and maintenance programs and where there is no adequate system of security and back-up.

A far more challenging problem is hardware and software obsolescence. Information technology continues to evolve and each new generation of hardware and software tends to displace the hardware and software of the previous generation. At its most basic level, this means that even if digital objects are copied perfectly and transferred to new storage media, it may be impossible to retrieve, render, or interpret these objects because of incompatibilities between the systems used to create them originally and the current generation of systems used to retrieve these objects. Consequently, digital preservation requires both maintenance of an accurate signal (byte stream) and the ability to retrieve and recreate that byte stream using current or future technology.

The problem of dependency on rapidly changing hardware and software seems intractable unless it is broken down into a number of smaller discrete problems and issues. Despite debates in the digital preservation community about the best method for ensuring longevity of digital materials [2, 28], most recent progress is the result of a focus on particular aspects of the problem and attempts to find solutions to smaller pieces of the puzzle. This approach also has the advantage of bringing research on high performance mass storage systems, metadata and representation schemes, rights management, and user evaluation to bear on the challenges of longevity of digital materials. At the same time, a number of useful models have been put forward as an overall framework for digital preservation that allow us to understand how different approaches contribute solutions to digital preservation issues and help us to identify which pieces are missing.
Technical Strategies

Digital preservation to date has relied on two main technical strategies: standards and migration. Technical standards form a foundation for much of what makes digital libraries possible. Standards and protocols for storage, data formats, bibliographic control, display, retrieval, transport, and distribution are imbedded in the infrastructure that make digital libraries accessible, manageable, and useable. In the area of digital preservation, standards issues primarily concern encoding, data formats, and representation schemes. Archivists and librarians tend to favor open standards over de facto or proprietary standards for two reasons. First, open standards are published and readily available whereas de facto standards (e.g. standards by virtue of common use or marketplace dominance) tend to be proprietary. If a digital preservation strategy depends on proprietary standards, then the long-term future of digital resources in that proprietary standard is dependent upon the longevity of the firm that owns the standard or on its continued market dominance. Second, open standards are developed through a consensus process where various parties with an interest in the standard have opportunities to contribute to or comment on its merits. In an ideal world, institutions that are building digital libraries and individuals with particular expertise in this area would participate actively in standards developments that affect the ability to preserve digital information and lower the costs of doing so. Digital libraries would then adopt policies that limit the materials in their collections only to those that conform to open standards.

In the real world, there are a number of limitations to reliance on standards alone as a digital preservation strategy [28]. First, there are many areas for which no technical standards yet exist. Commonly, new types of media, new forms of representation, and other innovations precede the development of either open or proprietary standards. Typical examples today include digital audio, digital video, color representation schemes for images, and multi-media objects, all of which are being generated in competing formats. Second, in the absence of open standards for many aspects of digital objects, proprietary standards become the de facto standard. Even where open standards exist, they may not be effective because a proprietary standard is technically superior to an open standard or because few or no vendors produce products that conform to the open standard. Third, standards change and evolve over time. Even strict adherence to standards will not eliminate the need eventually to convert digital materials in obsolete, but standard formats, to newer formats. Finally, even if there are accepted and implemented standards for the types of materials that a digital library collects and wishes to preserve, the digital library developers may not be able to enforce those standards on the firms, organizations, and individuals that supply information to it.

Despite these limitations on standards as an underlying digital preservation strategy, there are many areas where standard play an important role in digital preservation. Digital libraries that build their collections around particular types of materials such as technical reports, electronic journals, or image collections, can benefit from standards that have been adopted or are emerging for those particular types of materials. Use of the Standard Generalized Markup Language (SGML) and now the Extensible Markup Language (XML) provides one example of successful standardization. SGML has proven a particularly robust standard for storing textual documents in digital libraries in spite of the fact that there are relatively few commercial products that allow users to generate documents in SGML-compliant formats. In the area of image files, the Tagged Image File Format (TIFF) is popular despite concerns about consistency in such areas as color matching [20]. In summary, standards are useful if there is a consensus in the preservation community about which standard or standards are effective for a particular
type of digital material, if there are readily available products that conform to or support the standard, if
the standard has a demonstrated track record, and if managers can require compliance with the standard
as a condition of inclusion in the digital library collection. Although these conditions may apply only
in a minority of cases, it is one starting point for reducing the long-term preservation costs of digital
libraries.

Migration is the most widely deployed technical strategy in repositories that have established digital
holdings. Migration has been defined as "a set of organized tasks designed to achieve the periodic trans-
fer of digital materials from one hardware/software configuration to another or from one generation of
computer technology to a subsequent generation" [28]. Subsequent research on migration has demonstra-
ted that there are several different types of migration, each of which may affect the digital products
that result. Dollar and Wheatley each developed taxonomies for migration based on the degrees of
transformation to the original byte stream and amount of human effort or technical complexity involved
Wheatley's categorization is based on the impact of various migration strategies on the functionality and
look and feel of the original object and it includes both data and software migration [38]. Lawrence, et.
al., developed a framework for a risk assessment of various migration strategies that allows archivists
and librarians to assess the risks associated with migration from a specific source format to a new target
format and that places risk the context of larger organizational, hardware, software, and metadata fac-
tors [21]. This study identified several risk categories associated with migration including, content fixity,
security, context and integrity, references, cost, staffing, functionality, and legal considerations. These
refinements to the concept of migration are especially useful because they provide a technical framework
for differentiating solutions and approaches based on the type of source file and the available options for
a target file. They also consider a variety of institutional and other context-specific factors that may
influence the uncertainty of digital preservation strategies.

Migration as a digital preservation strategy has been criticized on two grounds [28]. First, migration
involves some transformation of the original byte stream, except where migration is so simple that it
involves only making an exact copy of the original byte stream on new media. Yet even with simple
copying and exact replication, the byte stream may be corrupted by software bugs, mishandling of data,
or mechanical failure of the input or output devices [21]. Changes to the original byte stream may involve
loss of information, loss of functionality, the introduction of errors into the target files, or changes in the
way the information is rendered to users. For example, conversion of documents created in a common
word processing application, such as Microsoft Word or WordPerfect, to a more portable format may
involve loss of formatting, changes to page layouts, loss of meaningful content such as comments and
annotations, transformation of color images to gray scale, and loss of the ability to further edit or
extract information from the document. Although such changes may not diminish the value or potential
reuse of certain types of digital resources, for some types of digital information these transformations may
undermine the potential for reuse of the preserved object. Moreover, as digital objects have become more
complex in their structure and behavior, only a small percentage of digital materials can be preserved
meaningfully simply by transferring the original byte stream to a new medium, making some type of
conversion necessary in most uses of migration as a preservation strategy.

A second criticism of migration concerns the potential for migration to scale up to the size and dimensions
of the digital preservation problem. Migration is not a unitary one-time process because, as Rothenberg
points out, "[m]igration requires a unique new solution for each new format or paradigm and each type
of document that is to be converted to a new form" [28]. Wide scale adoption of standards by the producers, distributors, repositories, and users of digital documents may reduce the intervals at which new conversion plans have to be developed and carried out, but this level of agreement on standards appears to be the exception rather than the rule when we consider the full range of communities creating digital materials that may warrant preservation. Without well established standards, each migration requires a customized approach that involves an analysis of the source file format, a selection of a target file format, and a conversion using either off-the-shelf products or programs written specifically for the conversion. There is no generic off-the-shelf software available to implement migration as a general preservation strategy, although some commercial products include several generations of backward compatibility as well as import and export functionality for related classes of documents, images, or data files [21]. Where custom design of conversion programs is necessary, there is little relationship between the quantity of digital information that needs to be converted and the cost of the migration. A digital library might easily invest as many resources to migrate one small but complex file as it does to migrate several large, but consistently structured databases. There are also concerns about the technical feasibility of migrating very large databases in the multiple terrabyte range. A recent report from storage experts at the NASA Goddard Space Flight Center pointed out that input and output (I/O) transfer rates have not accelerated as rapidly as storage capabilities. For very large databases, even if migration is carefully planned and undertaken as soon as a new storage system is in place, there might not be enough time to transfer all of the data to the target system before the target system itself becomes obsolete [15]. Emulation is under investigation as an alternative to migration. One model for emulation, proposed by Rothenberg, entails preservation of the original digital document (byte stream) and the original application software and operating system used to create the document, the creation of an emulator specification that can be used at some future date to write an emulator that will run on a future hardware platform, and the creation of documentation that remain human-readable and provide instructions for future access and use of the document [28]. In theory, this approach has several distinct advantages over migration. First, no transformations of the original byte stream are necessary. By preserving the original document along with the software necessary to reproduce it, a digital document can be presented to users in the future with its original "look and feel" and functionality intact. Emulation is also considered to be more cost effective because a single emulator for a particular hardware configuration could be used to access all digital information that requires that configuration for retrieval and viewing. Finally, according to this model, emulators do not have to be written for each change in hardware configurations, but only on an as needed basis when a user wishes to retrieve information dependent on a particular hardware platform [28]. Research underway on the technical feasibility of emulation raises questions about emulation as a unified concept [14, 23]. Analyses and tests of the Rothenberg proposal reveal potential flaws or a high degree risk if future recovery of preserved digital information depends on the ability to write emulators on an as needed basis at some point in the future. In one experiment that developed an emulator for an operating system which had been obsolete for nearly two decades, the investigators would not have been able to develop an emulator based on available documentation. Rather they relied on access to some of the original developers of the system and their own knowledge of it to cope with many undocumented work-arounds that programmers had introduced into the original system to overcome what were, at that time, hardware and memory limitations. [37]. One could argue that such problems are to be expected.
when writing emulators for obsolete systems because there is no framework for expressing emulator specifications. Lorie, however, contends that creating a perfect and complete description of any system architecture is a notoriously difficult task, implying that even if emulator specifications were in place and widely adopted, it would be difficult to anticipate all of the nuances, work-arounds, and tacit knowledge that would have to be specified formally [23]. Lorie advocates an alternative to the writing emulators on demand by writing emulators in the language of a Universal Virtual Computer (UVC). Our research in the CAMiLEON project testing emulation as a preservation strategy for an obsolete computer game further revealed that there are many versions of emulators available and that they vary considerably in their ability to reproduce the original "look-and-feel" of complex digital objects. In a recent experiment, users who compared a migrated or emulated version of a computer game to the same game running in its original native environment, generally found the migrated version more like the original in part because a thorough migration (complete rewrite of the original code to run on a current platform) performed more like the original program than a poor emulation [32].

While it is premature to reject emulation as a digital preservation method, recent research and experimentation indicates that this method falls short of Rothenberg's ideal solution which "should provide a single, extensible, long-term solution that can be designed once and for all and applied uniformly, automatically, and in synchrony" [28]. Current research on emulation is attempting to define where emulation may be the most effective or the most cost-effective approach, such as for objects where the behavior of the original must be archived, for retrieval or salvage of data stored in obsolete formats, and perhaps for classes of digital objects where a single emulator can provide access to a large number of objects.

There are three other technical approaches to digital preservation that warrant mention because they may have some role to play in preserving certain types of digital objects or in hybrid solutions. The strategy of print-to-paper is so primitive that it barely deserves attention, except for the fact that printing to paper may serve as a stopgap measure for preserving digital information in very simple formats, such as an exact page image, in institutions that lack the technology infrastructure and capacity to pursue other methods. This approach however, is becoming increasingly limited in its application as digital objects become more complex and as they contain more features and/or behaviors that can only be preserved or replicated if the materials are preserved in digital form. Another approach relies on "computer museums" or repositories of obsolete computer hardware, peripheral devices, operating systems, and application software [30]. Computer museums have been ruled out as a long-term preservation strategy because of the impossibility of keeping computers or peripheral devices in operating condition for the long run. There are attempts however in a few museums to extend the useful life of obsolete hardware and these could intersect in a meaningful way with efforts by librarians and archivists to salvage obsolete materials. Finally, digital archeology refers to a variety of methods that can be used to access data on media that have been damaged in disasters, from age or neglect, or where the hardware and software are either no longer available or are unknown. Ross and Gow summarized the results of several case studies of data recovery using digital archeology and concluded that with sufficient resources, much material that seems lost when media are damaged or allowed to deteriorate can be recovered [27]. The operative question for digital archaeology is the question of sufficient resources. The case studies show that data recovery is very expensive and often only partially successful. Therefore, digital archaeology is useful only as a last resort and only when data to be recovered has extraordinary high value.

Making progress on the technical front depends on abandoning the notion of a single one-size-fits-all
solution to digital preservation. Librarians, archivists, digital library designers, and even private citizens need guidance now on the most appropriate, cost-effective, and risk-averse options to use while the research community develops more effective technical approaches. Some general guidelines are available based on such factors as file formats, available standards, and cost [3, 18, 21]. Although the technical methodologies are radically different, the process of selecting appropriate methods from a suite of possible strategies is similar to preservation administration for conventional materials where librarians or archivists consider such factors as the nature of the original material, its current condition, and its potential uses before choosing a particular approach to preservation. On the research front, there is considerable work to be done in evaluating how different technical approaches can work together to best address digital preservation under a particular set of circumstances. Rather than working toward the one best method for preserving digital materials, I envision a suite of methods and tools and an associated set of decision-making tools that will help collection managers select those that are most appropriate for their specific conditions.

Operational Models and Programs

Digital preservation is difficult to untangle from many other digital library functions. Preservation concerns may affect collection development if a digital library limits the materials it will acquire to those that conform to designated standards. Likewise, preserving digital materials would be pointless if a digital library could not provide a means for accessing the materials. The inter-relationships between collection development, preservation, and access, however, add levels of complexity to digital library development that are difficult to address without some method of breaking the inter-dependencies down into reasonable and solvable problems. These closely related issues also influence how digital libraries are designed and administered. Recent conceptual work on archival functions and efforts to develop operational digital repositories provide some useful frames of reference for integrating digital preservation into digital library design and operations.

The Open Archival Information Systems (OAIS) Reference Model has been adopted by a number of digital libraries as a basic framework for defining digital archiving functions and for organizing digital repositories [6]. The OAIS reference model, a proposed ISO standard, was developed by the Consultative Committee on Space Data Systems as a framework for designing repositories capable of storing and managing the types of data collected and archived by member space agencies, but it has resonated with a variety of digital preservation initiatives. The model is a high-level description of the environment for an OAIS, the types of information that an OAIS must be capable of handling, and functions that are necessary to maintain an OAIS. An OAIS operates in an environment where there are producers (people or client systems) that create data to be archived, management that sets overall policy for the OAIS but does not oversee its day-to-day operations, and consumers (people or client systems) that interact with the OAIS to find and acquire preserved information. An OAIS contains of a number information objects (called packages) that consist of content information, preservation descriptive information, packaging information, and descriptive information. The OAIS further defines three types of information packages: a Submission Information Package (SIP) that is supplied to the OAIS by the producer, and Archival Information Package (AIP) that has all of the qualities needed for permanent or long-term preservation of an information object, and a Dissemination Information Package (DIP) that is distributed to a consumer upon request. The high-level functions of the OAIS include ingest (acquiring SIP’s and preparing them
for archival storage), archival storage, data management, administration, and access.

The OAIS model has been used in a number recent projects to develop repositories that are capable of acquiring digital information, preserving it for long-periods of time, and delivering digital information to users. Project NEDLIB, the Networked European Depository Library, used the OAIS reference model to model the functions needed for a Depository System for Electronic Publications. According to the NEDLIB developers, the OAIS model provided a common conceptual framework and terminology that made it possible to communicate with partner institutions and with other libraries that are developing depository systems for electronic publications [24, 35]. As a high-level reference model, the OAIS does not specify particular standards for the information packages, nor does it require a particular technical strategy for archival storage and data management. This leaves digital libraries and other repository developers with a great deal of flexibility in the precise implementation of an OAIS-compliant repository, but it also requires a great deal of work on the part of each digital library or concerned community to adopt standards or develop specifications that satisfy the needs and requirements of their particular communities of producers and users.

Several projects have added further refinements to the OAIS model or to particular aspects of it. NEDLIB, for example, developed standards for its Depository System that covered submission, archival storage, data management, access, administration, and preservation [12]. The CEDARS Project in the United Kingdom, developed draft specifications for preservation metadata that further refine the requirements for content information, preservation descriptive information, and representation information [5]. In addition to the use of the OAIS model in the conceptualization and design of digital repositories, it has the potential to create a larger market for storage systems and other products that vendors might produce to support preservation functions. Finally, the model could be used by existing digital libraries to assess whether their current architecture, administration, and data handling procedures meet the requirements for long-term preservation.

The OAIS model provides a useful conceptual framework for considering digital archiving requirements and developing repositories that satisfy those requirements. But models such as these are useful only to the extent that the organizations responsible for developing digital libraries also create management structures that consider preservation an important requirement for digital libraries and that can mobilize the resources necessary to support digital preservation. Research on digital preservation needs and requirements indicates that while almost all major research libraries are both acquiring digital materials and converting portions of their analog collections to digital form, most are still treating preservation issues as an afterthought [17]. There are several factors that have worked against development of clear management structures for digital preservation. First, many early digital library programs and many projects to develop digital collections were launched as research, experiments, or special projects with no consideration for preserving the assets that each program or project developed. Second, beyond conceptual frameworks, specific standards for some data formats, and a few persistent repositories that are under development, there are no working models of well-integrated and cost-effective digital preservation programs. Third, competing claims about the technical feasibility or effectiveness of different preservation methodologies has caused confusion among program administrators and possibly promoted a "wait-and-see" attitude [17]. Finally, there are no well developed financial models for digital archiving that can help managers assess the overall costs of adding preservation capabilities to digital libraries or to compare the relative costs of different strategies.
Current Research and Ongoing Issues

The increasing amount of research underway that directly or indirectly addresses concerns about the longevity of information is another encouraging sign that digital archiving has become an important issue for digital libraries. The first round of NSF-sponsored digital libraries research in the United States (DLI-1), for example, did not include any projects related to long-term preservation, whereas digital preservation is included in several of the projects funded by the current program (DLI-2). In addition to the NEDLIB project in Europe, several other national libraries have launched experimental programs to develop digital archiving capacity [4, 31, 34]. A number of non-government organizations in the United States and Europe are also supporting digital archiving activities including the Digital Library Federation, the Coalition for Networked Information, the Council for Library and Information Resources, OCLC, the Research Libraries Group, and several private foundations. Several publishers have made commitments to archiving the electronic journals that they own, control, or redistribute and to conduct the underlying research necessary to meet these commitments. The extent to which industry is investing in research on this problem remains unclear, although there are some industry-based projects underway as well [23]. Rather than providing a comprehensive overview of ongoing digital preservation research, I will touch upon some of the issues that are being investigated and then conclude with some thoughts on unresolved issues.

Several research projects are investigating ways to improve basic digital preservation methods by making them more robust, scalable, reliable, and cost-effective. The PRISM project at Cornell University is investigating long-term survivability of digital information, reliability of information resources and services, interoperability, security and metadata that makes it possible to ensure information integrity in digital libraries. Researchers at Cornell are translating traditional preservation strategies to policies and practices for the digital realm in order to develop digital preservation tools and mechanisms that, to the extent possible, can be integrated into the digital library services and operations. By integrating preservation into the digital library services, more of the decision-making can be automated and enforced by the system, leading to more cost-effective and consistent preservation practice [8]. Researchers at Stanford University are designing and implementing a modern, scaleable digital library repository by investigating several fundamental problems including identification of digital objects in a distributed and changing environment; the replication of digital objects for archiving; the management of metadata; distributed indexing mechanisms; and robust and scaleable awareness schemes. The Stanford researchers have developed a model for an archival repository and a simulation tool that can model risk features of digital libraries including media and format failure for multiple formats and components of a digital library [7, 10]. LOCKSS (Lots of Copies Keep Stuff Safe) is a project at Stanford University that is testing a system for permanent publishing and access to immutable digital content, such as peer-reviewed electronic journals. LOCKSS uses off-the-shelf open source software to manage a web cache at each participating library. The system detects and recovers from failures by ensuring that a minimum number of copies are available in the network of caches, by identifying copies that have changed, and by replacing damaged copies with what the system agrees is an undamaged copy. LOCKSS requires minimal equipment, maintenance, or human oversight, although its use is limited to stable types of digital objects [22].

Research is underway on the questions of integrity, authenticity, and user requirements for digital archives. Authenticity and integrity of digital information have been underlying concerns in digital
preservation because of the ease of altering digital objects and the dynamic nature of digital information. Moreover, preservation strategies that transform digital objects into standard formats for ease of preservation or that convert data from one format to another in a migration process may compromise the authenticity of the original object. The issues of integrity, authenticity, and user requirements for digital archives were addressed mostly on a theoretical or hypothetical basis until digital libraries developed to a point where preservation became a fundamental concern. The projects discussed above are all addressing integrity of digital data at the level of exact replication of digital objects when they are copied or used as redundant back-up. Some of the digital preservation research is going beyond this to consider how users react to digital objects that may have been altered in some way and how to represent digital objects to users in a way that they can trust the validity of the digital information they receive when exact replication is not feasible over the long-term. The Data Provenance Project at the University of Pennsylvania is developing methods to allow users of databases to determine the source and provenance of any piece of data that appears in the result set of a database query and to assess why this particular data element appeared in the result set [33]. In the CAMiLEON Project, we are investigating user preferences for different versions of preserved digital objects and analyzing how users define authenticity of objects running in their native software environment, running under emulation, and delivered in migrated versions [32]. Research on user requirements will enable digital libraries to provide transparency with regard to changes in digital objects that were necessary for long-term preservation without undermining the basic integrity and meaning of the information that has been preserved.

Despite considerable research and development during the past five years, there are several outstanding questions regarding the scaleability of digital preservation methods, inter-operability among digital archives, and the extent to which the methodologies developed to date can be applied in general to the wide variety of digital materials that may warrant long-term preservation. Most of the research and development projects have concentrated on particular types of digital materials, such as space science data, electronic journals, or image files. In general, research and development has not addressed the vexing problem of dynamic digital objects that are frequently or continuously changed and updated. Programs for web archiving, such as the Internet Archive or the Swedish Web Archiving Project, use web harvesting tools to collect periodic snapshots of all available web pages for all or a portion of the World Wide Web [19, 31]. These projects have collected and preserved valuable resources for understanding the evolution of the Internet, its growth, and its varied uses, but web harvesting tools cannot capture every iteration of all digital documents as they are altered and revised by human authors or generated "on-the-fly" by software agents.

Another concern is the inter-linking of digital documents and the long-term management of url's and other hyperlinks. Research is underway on persistent url's, digital object identifiers, and resolver services that could provide a means for uniquely identifying each iteration of a digital document. Research is needed in this area so that hyperlinks and citations to digital objects will lead a user to the document as it was at the time a link was established or at the time that a document was cited. So far, however, such methodologies and services have only been deployed in relatively small and focused communities. Many repositories have addressed this problem by adopting policies that limit what is preserved in the repository to top level pages or to a limited number of levels of linked objects. In many cases, access restrictions and rights management concerns prevent acquisition of all of the documents that are linked from the top level, even if the repository has a policy of harvesting more deeply. The problems of interdependencies and linkages among digital objects is an extremely difficult one that will require research,
standards, and policy development on technical, organizational, and legal issues.

The current methods for digital archiving also are most effective where there is a strong relationship between the producers of digital information, the digital library, and the user community. It is more feasible to impose standards through market mechanisms, formal agreements, or appeals to the interests of the community when the digital library has strong ties to the producers and users of digital information. Several commercial publishers and scholarly societies are participating in research and development projects for long-term preservation of electronic journals and other forms of scholarly communications [24, 25]. These efforts are attempting to develop mechanisms to enable long-term preservation without undermining the needs of publishers to recover costs, generate revenue, or maintain market share for current materials.

In conclusion, recent progress in digital preservation is a consequence of the growing awareness of longevity as a critical issue for sustainable and useable digital libraries, increased investments in research and development, and efforts to focus on discrete and potentially solvable aspects of the problem. As a result of these efforts, there are much improved models, mechanisms, tools, and organizational strategies for building and maintaining digital libraries that may persist for the long term. One particularly effective approach is to work with a specific community of providers and users to develop standards, acceptable formats for acquisition, storage, and dissemination. Yet, it is also important to remember how much digital information falls outside the domains where sound digital preservation strategies and methodologies are being developed. The collecting realm for most digital libraries and digital archives includes a vast array of heterogeneous material, created by a wide variety of producers but still of potential value to heterogeneous communities, including users far in the future whose interests, needs, and requirements are impossible to anticipate. We have developed effective methods for preserving the content of some well-established forms and genres of material, but we have had little success with new and emergent forms of documentation that may represent some of the most creative and innovative uses of digital technologies. Our current methods are strained or inadequate to address problems such as digital art and performance, complex visualizations, and dynamic objects that are generated "on-the-fly." Finally, we have little data and few models for estimating the costs of digital preservation, defining and identifying more cost-effective methods, or for distributing the costs among producers, digital repositories and users of today, not to mention users in the future. These issues will provide ample opportunities for research, experimentation, and development for years to come.

References


24. NEDLIB, Networked European Depository Library http://www.kb.nl/coop/nelib/


— 15 —