The History and Current State of Digital Preservation in the United States

Peter B. Hirtle

An earlier version of this paper was given at the conference on "Practical Experiences in Digital Preservation" held at The National Archives, Kew, England, in April, 2003.

Introduction

One of the fundamental functions of all research libraries is the preservation of the literary, scientific, and cultural record of humanity. Libraries have been greatly assisted in this task by the nature of the materials on which much of our cultural legacy has been stored.[1] Information stored on clay tablets or in carved marble can endure for thousands of years. Even paper, when properly manufactured and stored, can have a life measured in hundreds of years. Today, however, much of the information being produced is digital,[2] and digital formats are notoriously fragile. Either the media on which information is stored become unreadable, or the hardware and software needed to read the media become obsolete. Think of that old 8" floppy disk in the back of the

[1] Paul Conway, *Preservation in the Digital World* (Washington, D.C.: Commission on Preservation and Access, 1996).
[2] Peter Lyman and Hal Varian, "How Much Information 2003?" http://www.sims.berkeley.edu/research/projects/how-much-info- 2003/ (8 Mar. 2008).

drawer with your attempt from twenty years ago to write the Great American Novel (in WordStar). The magnetic data may not still be readable, drives that can read the disk are scarce, and few word processing packages today can understand WordStar documents.

To preserve analog information resources, it is often sufficient to house them in a benign, monitored, environment. In particularly bad cases, it might be necessary to make a microfilm or xerographic copy of the original, but copying is the exception rather than the rule. Digital preservation requires much more. Successful digital preservation encompasses a broad range of activities designed to extend the usable life of machine-readable computer files and protect them from media failure, physical loss, and obsolescence.

As part of an effort to understand the unique nature of digital preservation, this paper addresses four issues. First, it provides a brief history of the concept of digital preservation. If no one yet has all the answers to digital preservation, it is in part because the concept is relatively young, and has continuously evolved during its brief existence.

Second, the paper provides a brief overview of some current U.S. initiatives in digital preservation. A word of caution is in order. The focus on U.S. initiatives is entirely arbitrary. It is not meant to suggest that either preservation specialists in the United States know more about digital preservation than people from other countries, or that the nature of the problem is somehow different in the U.S. The challenge of preserving records created or maintained in electronic form knows no national boundaries. Microsoft and the World Wide Web (to cite just two common sources for digital documents) are global in their reach. Additionally, the problems that Microsoft

and the World Wide Web represent for archivists are much the same regardless of the country in which the archivist is based. There are many forums that debate the harms and benefits of globalization, but there is no question that for archivists, globalization has created a set of shared problems, and possibly shared solutions. The bits that make up digital documents are the same, no matter whether they are located in Cambridge, Massachusetts or Cambridge, England, and the problems that specialists face in preserving those documents are also the same. There is much that digital preservation specialists from around the world can learn from one another.

Third, the paper will summarize some of the principles and findings that have emerged over the past decade and are shaping the development of current preservation initiatives. In particular, the paper will cite the importance of preservation metadata as the basis for all successful preservation programs. Finally, the paper will close with some questions that are still outstanding.

Brief History of Digital Preservation

"Digital Preservation" is a relatively recent term in English. One of the earliest references to digital preservation appeared not in the library or archival literature, but in the journal *Theatre Crafts*, where an article entitled "Digital Preservation" appeared in 1992. It described the technical state of the art:

> The Biesemeyer table saw BladeGuard™ system was designed with the convenience of the operator in mind. The see-through guard is wide enough to accommodate any blade angle or dado head, and is counterbalanced so that the guard rides easily over the work. The guard is mounted on a long boom that attaches to any table saw and ensures that the vertical support post is well out of the way of the work. The guard lifts with a finger touch and latches out of the way for changing the blade, but an alarm with a key switch beeps to warn the operator if the saw is started

> with the guard in the storage position.[3] Thus, in 1992 "digital preservation" was just as likely to mean keeping one's fingers when using a table saw as it was to mean the preservation of digital data.

The preservation of electronic records had been of concern to archivists since at least the early 1960s, when the Machine-Readable Records Branch was formed at the National Archives.[4] The Inter-university Consortium for Political and Social Research (ICPSR), a data archive at the University of Michigan, also began in the early 1960s. These early programs, while concerned with the preservation of digital information, tended not to use the term "digital preservation," preferring instead to speak of the archiving of electronic records or the preservation of data sets.

The earliest reference that I could find in English to the "digital preservation" of data occurs in the context of the research that Anne Kenney and Lynne Personnius undertook in 1990 at the Cornell University Library in conjunction with the Xerox Corporation. In their research, "digital preservation" meant using digital technologies to reformat analog media as part of the preservation process of those media. Reformatted information, or what came to be known as "re-born digital" documents, were the heart of early digital preservation initiatives. The concept of digital preservation originally developed in libraries, not archives, as an aid to ongoing library analog preservation efforts. Furthermore, it initially did not concern itself with the preservation of information that was "born digital." In fact, M. Stuart Lynn, the Vice President for Information Technologies at Cornell University, declared in August, 1990, in a standard

[3] "Digital Preservation," *Theatre Crafts* 26 no. 4 (April 1992): 56.
[4] On the early history of the electronic records movement in the U.S., see Richard J. Cox, *The First Generation of Electronic Records Archivists in the United States: A Study in Professionalization (*New York: Haworth Press, 1994), and Bruce I. Ambacher, *Thirty Years of Electronic Records* (New York: Rowman & Littlefield, 2003).

glossary that "original documents that are of concern for library preservation purposes are not normally encoded in a digital electronic medium."[5]

The library and archival communities concerned with the management of information in electronic form came together in 1994 with the creation of the Commission on Preservation and Access/Research Libraries Group (CPA/RLG) Task Force on Archiving Digital Information.[6] While formed by two groups primarily associated with libraries, the Task Force included two archivists among its members, along with publishers active in electronic publishing and others responsible for information created in electronic form. By 1998 information created in electronic form was redefined as "born digital" resources.[7] "Re-born digital" materials, meaning those materials that have been electronically reformatted, have become such a subset in digital preservation that a background report on digital preservation for the Library of Congress could reverse M. Stuart Lynn's definition of slightly more than a decade earlier and define the scope of the digital preservation issue to encompass only material that is born digital.[8]

The creation of the field of digital preservation as an activity of the library community has meant that some issues that are important to archivists may have initially received less attention. Early library-based digital preservation initiatives, for example, followed the analog preservation model used in libraries, which emphasized

[5] M. Stuart Lynn, *Preservation and Access Technology: The Relationship Between Digital and Other Media Conversion Processes: A Structured Glossary of Technical Terms* (Washington, D.C.: Commission on Preservation and Access, 8/1990), Section 1.1.6, http://www.clir.org/pubs/reports/lynn/index.html (8 Mar. 2008).
[6] Task Force on Archiving of Digital Information, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (Washington, DC: Commission on Preservation and Access, 1996).
[7] The Word Spy, "born-digital," http://www.wordspy.com/words/born-digital.asp (8 Mar. 2008).
[8] Amy Friedlander, "Summary of Findings," in *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program. Appendices* (Washington, D.C.: National Digital Preservation Program, 2002, p. 18).

the transfer of information to stable media that could be managed with minimal effort. Early on, archivists realized that all electronic media requires active management. Additionally, librarians focused on capturing and preserving the information found in documents, whereas archivists were also interested in preserving the integrity, authenticity, and reliability of records. In addition to preserving the information content of records, archivists have emphasized the need to maintain the ability of electronic records to serve as evidence.

Current Initiatives in Digital Preservation

While the initial interest in digital preservation may have begun in libraries and then merged with the interest of archivists, it has grown beyond these communities. Digital preservation is the subject of articles in the popular press, and governments have come to recognize the importance of digital archiving. The supplement to the federal budget in 2002, for example, noted that "strategies to assure long-term preservation of digital records constitute another particularly pressing issue for research."[9]

In several new initiatives, the differences between how libraries and archives approach electronic information are becoming less distinct. Most prominent among these programs is the National Digital Information Infrastructure and Preservation Program (NDIIPP).[10] An initiative of the Library of Congress, the legislation creating the NDIIPP insists that the Library consult with other agencies as well, including the

[9] Interagency Working Group on Information Technology Research and Development, National Science and Technology Council, *Networking and Information Technology Research and Development Supplement to the President's Budget for FY 2002:A Report* (Arlington, VA: National Science and Technology Council, 2001), 23, http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/19/e7/eb.pdf (8 Mar. 2008).
[10] National Digital Information Infrastructure and Preservation Program, http://www.digitalpreservation.gov/ (8 Mar. 2008).

National Archives and Records Administration (NARA). The mission of NDIIPP is simple: to develop a national strategy to collect, archive, and preserve for current and future generations the burgeoning amounts of digital content, especially materials that are created only in digital formats. Initially the program intends to focus on some specific areas, including electronic journals, geographic information systems, and other interactive objects.[11]

In December 2002, Congress accepted the planning report from NDIPP.[12] At the same time, it released an additional $35 million of the $99.8 million it had authorized for the NDIPP program. With the funding in hand, the program has begun to implement many of the recommendations in the report, including the establishment of a national network of preservation partners. In the late 1990s, much of the most exciting work in digital preservation was taking place in Europe and Australia in projects such as CEDARS,[13] PADI,[14] PANDORA,[15] and NEDLIB.[16] Thanks to the NDIIPP initiative, the Library of Congress is now well positioned to become a leader in digital preservation.

A second American initiative likely to be of great interest to the digital preservation community is the National Archives and Records Administration's work on an Electronic Records Archive (ERA).[17] Building on its long history in working with electronic data archives, NARA is designing the ERA so that it will authentically

[11] Laura Campbell, "National Digital Information Infrastructure and Preservation Program," *RLG DigiNews* 7 no. 3 (June 2003), http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070513:000006285814&reqid=1033#feature1 (8 Mar. 2008).
[12] *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program: A Collaborative Initiative of the Library of Congress* ([Washington, D.C.?]: The Program, [2002]), http://www.digitalpreservation.gov/library/pdf/ndiipp_plan.pdf (8 Mar. 2008)
[13] "CURL Exemplars in Digital Archives," http://www.leeds.ac.uk/cedars/ (8 Mar. 2008).
[14] Preserving Access to Digital Information, http://www.nla.gov.au/padi/ (8 Mar. 2008).
[15] Preserving and Accessing Networked Documentary Resources of Australia (PANDORA), http://pandora.nla.gov.au/ (8 Mar. 2008).
[16] Networked European Digital LIBrary (NEDLIB), http://nedlib.kb.nl/ (8 Mar. 2008).
[17] National Archives and Records Administration, Electronic Records Archives, http://www.archives.gov/electronic_records_archives/index.html (8 Mar. 2008).

preserve and provide access to any kind of electronic record, free from dependency on any specific hardware or software. In its planning documents, NARA has stressed that any solution must be able to deal with the immense volume of records created by the federal government, that it must ensure the authenticity of those records, and that it must provide access to the records. For instance, a conservative 1999 estimate indicates that the yearly volume of e-mail traffic in the U.S. federal government is approaching 36.5 billion messages per year.[18] Although only a percentage of those messages may be permanently valuable, the volume is still orders of magnitude larger than what NARA has had to manage in the past.

The National Archives and Record Administration's preliminary work has been heavily focused on the technology and infrastructure needed to build such an archives. It has established strong partnerships with some of the leading government research institutes and initiatives, including the San Diego Supercomputer Center, U.S. Army Research Laboratories, the National Initiative for Standards and Technology, and the National Aeronautics and Space Administration. The value of such partnerships is demonstrated by the development of the Open Archival Information System (OAIS) reference model, which was developed by the space science community with the strong support of NARA. [19]

The ERA initiative has been funded with $38 million dollars in FY2003 budget. In August 2004, NARA announced that two companies, Lockheed Martin and the Harris Corporation, had been issued contracts valued at $20.1 million to design "a

[18] Jason R. Baron, "E-mail Metadata in a Post-Armstrong World" (paper presented at META-DATA '99: The Third IEEE Meta-Data Conference, April 6-7, 1999, National Institutes of Health, Bethesda, Maryland), http://www.archives.gov/era/pdf/baron-email-metadata.pdf (8 Mar. 2008).
[19] http://ssdoo.gsfc.nasa.gov/nost/isoas/ (8 Mar. 2008).

technological solution to the challenge of preserving electronic information across space and time."[20] Prior to the awarding of the contract, a study committee of the Computer Science and Telecommunications Board of the National Academy of Sciences assessed NARA's plans, and its report is an important analysis of some of the technical problems associated with maintaining software-independent and authentic electronic records.[21]

While the NDIIPP and ERA initiatives are now well underway, several other groups have been supporting important work in digital preservation. The National Science Foundation (NSF), for example, has supported several research projects in digital preservation through its Digital Libraries funding, and there is an implied (though as yet unrealized) preservation component to its National Science Digital Library initiative. In conjunction with the Library of Congress, the NSF recently convened a meeting of fifty- one specialists in computer science, archives, and libraries to develop a research agenda in digital preservation.[22]

The Research Libraries Group and OCLC, Inc. have also undertaken important work in digital preservation. They have jointly prepared a fundamental report on the attributes of trusted digital repositories.[23] One conclusion that has emerged from the

[20] National Archives and Records Administration. "National Archives Names Two Companies to Design an Electronic Archives." Press release, 3 August 2004. http://www.archives.gov/press/press-releases/2004/nr04-74.html (8 Mar. 2008).
[21] National Research Council (U.S) Committee on Digital Archiving and the National Archives and Records Administration, *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development* (Washington, DC: National Academies Press, 2003), http://www.nap.edu/books/0309089476/html (8 Mar. 2008).
[22] *It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation. Final Report*. (Paper presented at the Workshop on Research Challenges in Digital Archiving and Long-term Preservation, April 12-13, 2002, sponsored by the National Science Foundation, Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering, and the Library of Congress, National Digital Information Infrastructure and Preservation Program), http://www.digitalpreservation.gov/library/pdf/NSF.pdf (8 Mar. 2008).
[23] *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report* (Mountain View, CA: RLG, May, 2002), http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf (8 Mar. 2008).

research on digital preservation is that no one institution will be able to preserve everything. Digital preservation will have to be collaborative. But how is the British Library to know if Cornell University, for example, has "preserved" a resource in such a way that the British Library need no longer worry about it? The definition of the attributes is a first step in this regard. RLG has joined with NARA in an international effort to develop certification criteria for digital archives. According to its charge, the Task Force on Digital Repository Certification will identify a digital certification process:

> A digital repository certification process should address the range of activities, functions, and responsibilities associated with repositories while providing layers of trust for all involved. It should yield a high degree of confidence that the information a repository disseminates is the same information that was ingested and preserved. And the certification process or framework must address the consequences of failure, including fail-safe mechanisms that would enable a certified archival repository to perform rescue of endangered digital information.[24]

The initial report of the Task Force is expected to appear shortly.

In addition to their work on the attributes of a trusted digital repository, OCLC and RLG have also been looking at the need for a standard set of preservation metadata. A joint working group synthesized much of the preservation metadata work conducted by CEDARS, Pandora, NEDLIB, and others to prepare a metadata framework for digital preservation.[25] It is currently at work on a new project, PREMIS, which seeks to determine the best way to implement a preservation metadata system.[26]

The PREMIS group has two specific goals. First, it seeks to define a "core" set of preservation metadata elements, applicable to a broad range of digital preservation

[24] Research Libraries Group, Task Force on Digital Repository Certification, http://www.oclc.org/programs/ourwork/past/repositorycert.htm (8 Mar. 2008).
[25] OCLC/RLG Preservation Metadata Framework Working Group, *A Metadata Framework to Support the Preservation of Digital Objects*. June, 2002. http://www.oclc.org/research/projects/pmwg/pm_framework.pdf (8 Mar. 2008).
[26] PREMIS stands for PREservation Metadata: Implementation Strategies. See http://www.oclc.org/research/projects/pmwg/ (8 Mar. 2008).

activities. As part of this task, the group intends to develop a data dictionary and guidelines for applying, populating, and managing the core elements. Second, the PREMIS group wants to "identify and evaluate alternative strategies for encoding, storing, managing, and exchanging preservation metadata—in particular, the core metadata elements—within a digital preservation system."[27] One possibility is to exploit the Metadata Encoding and Transmission Standard (METS)—an XML schema for encoding descriptive, administrative, and structural metadata regarding objects within a digital library—to store preservation information.[28]

The list of interested participants in the digital preservation arena goes on. The Andrew W. Mellon Foundation, for example, recently funded six planning projects to determine how electronic journals can be preserved.[29] In the past, libraries preserved the hard copies of journals they purchased. Today, libraries are more likely to license access to an electronic journal rather than purchase it. That access right usually does not give a library the right to make a local copy of the journal for preservation purposes. Continued access to commercially produced data sources may also become of interest to governmental archives as well. Copyrighted information is often used in government investigations, and it may serve as the basis for government actions. How will the

[27] Brian F. Lavoie, "Implementing Metadata in Digital Preservation Systems: The PREMIS Activity," *D-Lib Magazine* 10 no. 4 (April 2004), http://www.dlib.org/dlib/april04/lavoie/04lavoie.html (8 Mar. 2008); Priscilla Caplan, "PREMIS - Preservation Metadata - Implementation Strategies Update 1. Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community," *RLG Diginews* 8 no. 5 (October 15, 2004), http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000068892&reqid=1106#article2 (8 Mar. 2008).
[28] On METS, see the official web site published by the Library of Congress at http://www.loc.gov/standards/mets/ (8 Mar. 2004).
[29] Mellon E-journal archiving, http://www.diglib.org/preserve/ejp.htm (8 Mar. 2008). See also Dale Flecker, "Digital Archiving: What Is Involved," *Educause Review* (Jan/Feb, 2003):10-11.

archives be able to capture this information if copies of it no longer exist in the files of the agencies involved?

The Mellon projects identified issues associated with the preservation of electronic journals, worked with publishers to try to better understand their issues and concerns, and explored business models that could support preservation activities. The Mellon Foundation is funding two follow-up activities. One is a grant to the JSTOR project that will explore whether a centralized repository makes the most sense.[30] The second is the Lots Of Copies Keeps Stuff Safe (LOCKSS) project, which is discussed later in this paper.

Other initiatives worth noting include the rise of interest in institutional electronic repositories.[31] Most notable, perhaps, is the DSpace Initiative at the Massachusetts Institute of Technology.[32] DSpace is a software implementation that creates a common storage environment for any work in electronic form produced at an institution. While initially developed with electronic publications and data sets in mind, there is a growing interest in using DSpace to include a wide range of electronic institutional records including such things as an archived version of electronic or distributed courses. The records management and archival implications of such a use

[30] JSTOR, "E-Archiving Born Digital Content," *JSTORNEWS*, http://www.jstor.org/news/2002.10/EarchivingBornDigitalContent.html (8 Mar. 2008).
[31] A good recent overview of institutional repositories is Mark Ware Consulting, *Publisher and Library/Learning Solutions (PALS) Pathfinder Research on Web-based Repositories: Final Report*, January, 2004, http://www.palsgroup.org.uk/palsweb/palsweb.nsf/79b0d164e01a6cb880256ae0004a0e34/8c43ce800a9c67cd80256e370051e88a/$FILE/PALS%20report%20on%20Institutional%20Repositories.pdf (8 Mar. 2008). See also Open Society Institute, *Guide to Institutional Repository Software v 3.0* (New York: Open Society Institute, August, 2004. http://www.soros.org/openaccess/software/ (8 Mar. 2008).
[32] DSpace Federation, http://www.dspace.org/ (8 Mar. 2008).

have yet to be determined.[33]

Industry, too, has become interested in the issue of digital preservation, though their answers are often shaped by their business needs. Kodak, for example, has argued that it is too expensive to preserve electronic data in electronic form, and proposes instead that electronic data can best be preserved if it is printed out to microfilm so that it can be readily rescanned back to electronic form when needed.[34] Adobe has realized that the Portable Document Format (PDF) format is often used as a de facto preservation format and is consequently leading an ISO committee to develop a PDF specification known as PDF/A (with the "A" standing for "archive").[35] Raymond Lorie at IBM's Almaden Labs has proposed a preservation strategy built around what he calls a Universal Virtual Computer (UVC). Developers of software and formats would write their code such that it could be read and displayed on the UVC. Future generations would only need to ensure that the UVC could run on contemporary machines – and not have to migrate all pre-existing software to a new environment.[36]

[33] 33. Clifford Lynch has written a number of papers related to this topic. See "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL Bimonthly Report* 226 (February, 2003), http://www.arl.org/newsltr/226/ir.html (4 Oct. 2004); "The Afterlives of Courses on the Network: Information Management Issues for Learning Management Systems," *Educause Center for Applied Research (ECAR) Research Bulletin* 2002 issue 23 (26 November 2002), http://www.cni.org/staff/cliffpubs/ECARpaper2002.pdf (4 Oct. 2004); and "Editor's Interview with Clifford A. Lynch," *RLG DigiNews* 8 no. 4 (August, 2004), http://www.rlg.org/en/page.php?Page_ID=19481#article0 (4 Oct. 2004).

[34] H. Andrew Lawrence, *Digital Insurance for Information at Risk: A Strategic Overview of Digital Preservation* (Rochester, NY: Eastman Kodak, 2000), http://www.microfilm.com/images/article_17.pdf (8 Mar. 2008). Kodak's digital preservation solutions are described at http://www.kodak.com/US/en/dpq/site/TKX/name/Digital Preservation (15 Mar. 2008).

[35] Michael Looney, "The Need for Digital Archiving Standards," *Syllabus: Technology for Higher Education* (March 2003), http://www.syllabus.com/article.asp?id=7362 (8 Mar. 2008); William G. LeFurgy, "PDF/A: Developing a File Format for Long-Term Preservation," *RLG DigiNews* 7:6 (December 15, 2003) http://digitalarchive.oclc.org/da/ViewObjectMain.jsp;jsessionid=84ae0c5f82406a58ea63de8540119e7ccc4ed769a23a?fileid=0000070511:000006280063&reqid=1522#feature1 (8 Mar. 2008).

[36] Raymond A. Lorie, "A Project on Preservation of Digital Data," *RLG DigiNews* 53 (June 15, 2001), http://digitalarchive.oclc.org/da/ViewObjectMain.jsp;jsessionid=84ae0c5f82406a58ea63de8540119e7ccc4ed769a23a?fileid=0000070511:000006279465&reqid=1522#feature2 (8 Mar. 2008); Raymond A. Lorie, "Long Term Preservation of Digital Information," in *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: ACM Press, 2001), 346-352, http://portal.acm.org/citation.cfm?id=379726 (8 Mar. 2008); Raymond A. Lorie, "A Methodology and System for Preserving Digital Data," in: *International Conference on Digital Libraries: Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital libraries, 2002, Portland, Oregon, USA* (New York: ACM Press, 2002).

Basic Principles in Digital Preservation

In spite of all the current research underway, as well as the identification by the NSF panel of additional areas for research, there are certain truths upon which all digital preservation specialists agree. In the area of technology, there is a growing consensus that digital preservation will require a combination of methodological approaches. To address the issues associated with hardware and software obsolescence, the techniques of format migration, emulation, encapsulation, and simple bit preservation must all be employed in order to preserve the usability of documents. There is even a place for the simple preservation of bits, and the concomitant development of what I have called "digital paleographers." Cultural repositories are filled with medieval manuscripts written in scripts and languages and using abbreviations that even most scholars cannot understand. Paleography has developed as a scholarly tool that assists experts as they read manuscripts. Similarly, digital paleographers will have the skills to be able to read other lost languages and scripts, such as HTML 2.0 encoded in ASCII, twenty years in the future.[37]

Which technical approach—migration, emulation, encapsulation, or bit preservation—digital preservation specialists adopt for any particular digital object will depend on the nature of that object. With some objects, maintaining their usability may be what is important. For example, it may be sufficient to be able to search, read, and print e-mail messages from the White House without having to know exactly what sort of computer system was being used, how the text appeared on the screen, etc.; it is the informational content that matters. For other digital objects, however, it may be

[37] Peter Hirtle, "Digital Paleography," *D-Lib Magazine* 6 no. 4, http://dlib.org/dlib/april00/04editorial.html (8 Mar, 2008).

necessary to maintain the original interface. We may wish to know, for example, how a modern Shakespeare actually interacted with her text as she wrote it on a computer monitor, just as we often like to listen to the music of Beethoven as it sounds when played on the kind of fortepiano he would have used. As with any reformatting initiative, archivists will need to select the proper method based on the nature of the original source material.

Equally important is the growing recognition that digital preservation is more than just a technological challenge. The organizational and social issues associated with digital preservation are more important than the technology. The definition of digital preservation adopted by the RLG/OCLC Working Group on Digital Archives Attributes highlights this issue: "Digital preservation," they note, "managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents."[38] Technologies are not mentioned in this definition; the focus is on digital preservation as a set of "managed activities."

Digital preservation needs to address issues associated with the organization of the digital archives and the procedures used to certify the integrity, authenticity, and reliability of the archived data. Furthermore, there is no assurance that these management activities will go away; digital preservation will require the *continuous,* active management of digital objects, especially as the technological and organizational approaches to digital preservation evolve. Thus digital repositories will also need a clear administrative mandate for their activities, and the financial sustainability to continue.

[38] Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report (Mountain View, Calif.: RLG, May, 2002), http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf (8 Mar. 2008).

Figure 8.1 highlights the component parts of an organizational view of digital preservation.

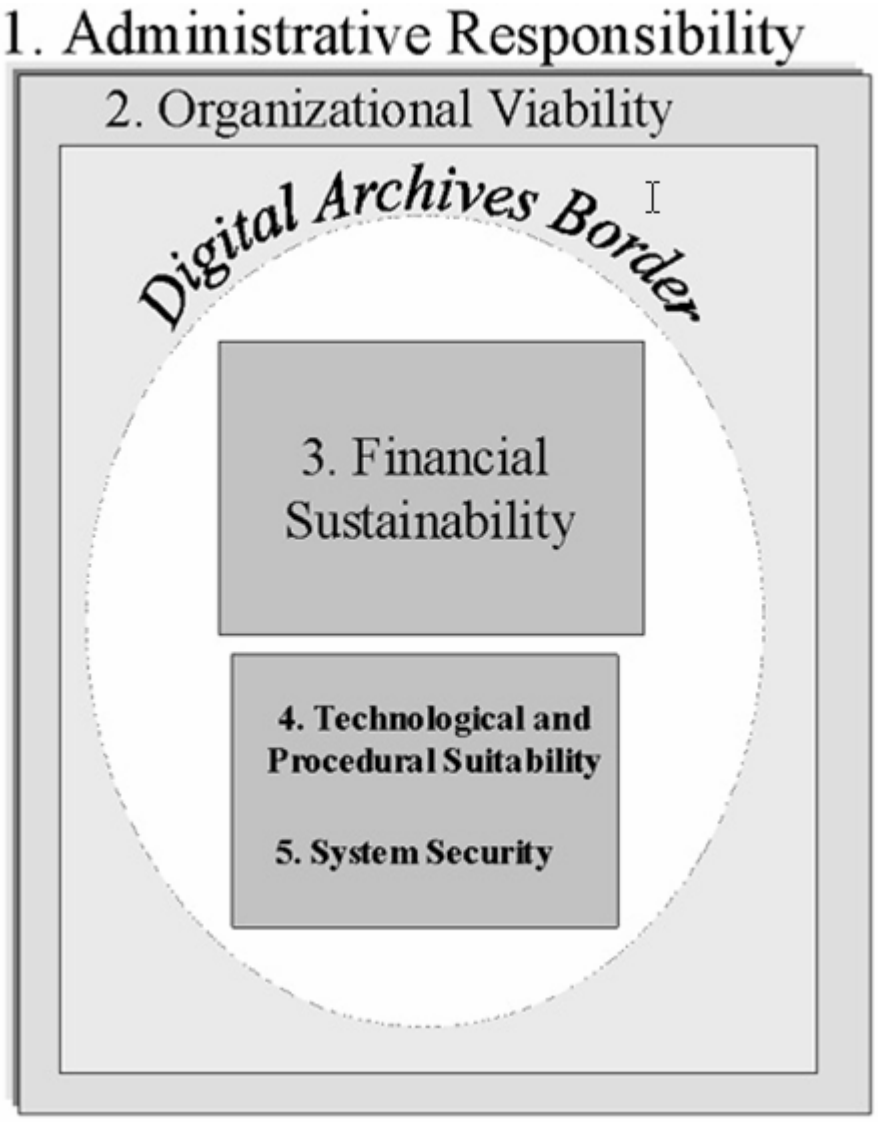

Figure 8.1: Trusted Digital Repositories (TDR) Framework Model[39]

[39] Anne R. Kenney and Nancy Y. McGovern, "The Five Organizational Stages of Digital Preservation," in Patricia Hodges, Maria Bonn, Mark Sandler, and John Price Wilkin, eds., *Digital libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee on the Occasion of Her Departure from the University of Michigan* (Ann Arbor, Michigan: Scholarly Publishing Office, The University of Michigan University Library, 2003), http://name.umdl.umich.edu/BBV9812 (8 Mar. 2008).

Another conclusion that has emerged from the past decade is that institutional collaboration must be at the heart of digital preservation initiatives. The amount of digital information in the world overwhelmingly exceeds the amount of information on paper. Even if one assumes that much of it does not need to be preserved, the remainder is still of such a large scale that it is unlikely that any single institution or funding source can assume the entire burden.

Furthermore, there is a growing recognition that, given the fragility of digital information, attempting to preserve a single copy may not be the best solution. Creating and preserving multiple copies may be the safest way to ensure that digital information endures. Stanford University's LOCKSS project is a major initiative designed to test the efficacy of a decentralized, distributed approach to managing digital assets. LOCKSS stands for "Lots of Copies Keeps Stuff Safe." It is based on research by Hector Garcia-Molina of Stanford on the level of redundancy needed to protect digital information from accidental or malicious harm. LOCKSS to date has not addressed issues of digital preservation, but a new grant from the Mellon Foundation will allow the project to explore this issue.[40]

One other area of emerging consensus is the importance of intellectual property issues to digital preservation. Copyright and other intellectual property laws challenge our ability to preserve digital information. It is unclear if even national libraries have the legal right to capture and preserve publicly accessible web content. The situation with materials whose access is governed by licenses is even less clear. We also do not know if,

[40] LOCKSS Program, Stanford University Libraries, http://www.lockss.org/lockss/Home (8 Mar. 2008).

when, or how the public may access and use digital information owned by parties other than the repository preserving the information.[41]

The emergence of digital rights management (DRM) systems will further complicate our ability to preserve information. Digital rights management promises technologies that will limit and control the ability to access and use digital information. A combination of encryption and access control measures will ensure that only those individuals who have the permission of the creator of material will be able to use it. Currently in the U.S. it is a crime to bypass these access control systems without the permission of the copyright holder even if the intended use of the material would be legal.[42] As noted earlier, even government archives will need to begin to worry about intellectual property issues as more of the resources used by agencies are accessed through digital rights management systems. Furthermore, more of the digital documents created by those agencies are likely to have DRM controls embedded as a default. The implications for digital preservation are horrendous.

Areas for Future Research

While the NSF Workshop on Research Challenges in Digital Archiving and Long-term Preservation fully articulates many of the research challenges still facing the digital preservation field, some are worth highlighting here. Earlier I noted that we must select

[41] Peter B. Hirtle, "Digital Preservation and Copyright," Copyright & Fair Use, Stanford University Library Web Site, http://fairuse.stanford.edu/commentary_and_analysis/2003_11_hirtle.html (8 Mar. 2008); June Besek, *Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment* (Washington, D.C.: Council on Library and Information Resources and the Library of Congress, 2003), http://www.clir.org/pubs/reports/pub112/contents.html (8 Mar. 2008).

[42] Peter Hirtle, "FAQ: The Impact of the Librarian of Congress's Rulemaking on the Digital Millennium Copyright Act," *RLG DigiNews* 76 (December 15, 2003), http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070519:000006289350&reqid=3550#faq (8 Mar. 2008); Heather Briston, "Digital Rights Management and Archivists," *Archival Outlook* (July/August 2003), 14.

the appropriate method of preservation based on the nature of the material. There is still much we need to learn, however, about what we mean when we talk about "usable and interpretable." When is it sufficient merely to keep the information readable, and when must we retain the "look and feel" of a digital document at the time of creation? Electronic documents are complex, and the decisions about what must be saved to create usable archives are difficult. For example, with a website, are the hyperlinks that are included on a web page part of the content that must be maintained to keep the website usable?

Closely related to this issue is the question of selection. In spite of the siren songs of technologists who suggest that storage is becoming so inexpensive that we will be able to save everything, selection will be an important component part of digital archives. The amount of digital information in the world is too much to be preserved, and the need to *manage* the digital information that is preserved, and not just store it, will challenge our limited resources. We need guidelines and appraisal criteria that can help us identify content of enduring value.

Once material is selected, we also need guidelines for how long it should be preserved. The well-known archivist James O'Toole has reminded us that even with paper documents, preservation is a time-constrained concept; "permanent" storage of microfilm, for example, means at best five hundred years.[43] For how long are we going to preserve digital data? Ten years? Ten thousand years? We need to build retention periods into our digital preservation schemes.

[43] James M. O'Toole, "On the Idea of Permanence," *American Archivist* 52 no. 1 (Winter, 1989): 10-25. Reprinted in Randall Jimerson, ed., *American Archival Studies* (Chicago: Society of American Archivists, 2000), 475-494.

There is still much we have to learn about the technologies and standards that should be employed in digital preservation. We need metadata schemes that can support work in selection, appraisal, and retention. We need persistent identifiers for preserved digital objects. We need formats and encoding schemes that accommodate the archivist's definitions of usable and interpretable digital information. And we need processes and procedures that ensure that fragile digital documents do not change over time.

Lastly, we must identify how the managed activities that constitute digital preservation are to be funded. A variety of solutions are under discussion, including the development of metrics that measure the costs and benefits of digital preservation and the creation of incentives, including tax incentives, for the creators of digital content to preserve their works. We also need to develop automatic methods, such as the LOCKSS system, to manage efficiently the preservation of digital objects. We know that we need to have multiple digital repositories, but it is unclear how many we need to have. Is it cost-efficient for every archive to develop an electronic repository, or might it be cheaper and more secure to let a commercial service such as OCLC's digital archiving service assume the responsibility? I have always been amazed by how much archives have been able to do when given so little. It is unlikely that digital preservation can similarly be done as cheaply.