The Paradox of Digital Preservation

Su-Shing Chen
University of Missouri-Columbia

Preserving digital information is plagued by short media life, obsolete hardware and software, slow read times of old media, and defunct Web sites. Herein lies the paradox: We want to maintain digital information intact, but we also want to access this information in a dynamic use context.

Easy Internet access presumes that everyone can capture, access, and use the world's accumulated digital information. People, businesses, institutions, and governments invest time and effort to create and capture digital information for instantaneous access by anyone. Businesses that produce computers, semiconductors, software, and communications equipment have accounted for a third of the total growth in US economic production since 1992.[1] Researchers are striving to make this information available to communities worldwide.

Unfortunately, we cannot guarantee the continued preservation and accessibility of digital information generated in this context of rapid technological advances.

Despite our information technology investments, there is a critical, cumulative weakness in our information infrastructure. Long-term preservation of digital information is plagued by short media life, obsolete hardware and software, slow read times of old media, and defunct Web sites. Indeed, the majority of products and services on the market today did not exist five years ago. More importantly, we lack proven methods to ensure that the information will continue to exist, that we will be able to access this information using the available technology tools, or that any accessible information is authentic and reliable.

SLOUCHING TOWARD ENTROPY

To get a sense of what's at stake from a historical perspective, consider this cursory sampling of irretrievable information from the past 50 years: 50 percent (approximately 25,000) of the films produced in the 1940s, most TV interviews, the first e-mail sent in 1964, and many objects of intellectual and cultural heritage.

Failing to address the problems of preserving digital information is analogous to fostering cultural and intellectual poverty and squandering potential long-term gains that we should rightfully receive as a return on our professional, personal, and economic investments in information technology. This failure can incur substantial costs for recovery, even in the short run, as observed in the following cases:

- The Census Bureau saved the 1960 Census on Univac tapes.
- NASA/NSF/NOAA rescued valuable 20-year-long TOVS/AVHRR satellite data documenting global warming.
- The federal government spent more than $15 million to recover e-mail from the Executive Office of the President in the Reagan and Bush administrations.

0018-9162/01/$10.00 © 2001 IEEE

- Private companies facing discovery orders for their digital information in connection with lawsuits often find that recovering this information costs more than the computer system itself.

DIGITAL PRESERVATION PARADOX

Traditionally, preserving things meant keeping them unchanged; however, our digital environment has fundamentally changed our concept of preservation requirements. If we hold on to digital information without modifications, accessing the information will become increasingly more difficult, if not impossible. Even if we could find a physical medium to contain unaltered digital data permanently, formats for recording the information would change and the hardware and software needed to recover the information would become obsolete. As an example, few of us would consider accessing the Vietnam-era historical data if we had to learn 30-year-old technology to retrieve it.

This situation creates a fundamental paradox for digital preservation: On the one hand, we want to maintain digital information intact as it was created; on the other, we want to access this information dynamically and with the most advanced tools.

Finding ways to resolve the tension between the creation context and the use context constitutes an important research challenge. To evaluate the issues of preserving digital information, we need to analyze requirements about content, formats and styles, and context; storage media; systems technology; workflow process; and metadata policies. Because the interdependencies among these issues are complex and unexplored, I can only highlight the basic issues.

Digital preservation requirements

The creation context refers to digital information production—information that is "born digital" and information that exists in another form but is converted or captured in digital form. The creation context determines the basic requirements for preserving digital information in three ways. First, the creation context provides the information content that must be preserved. Second, it captures this information in specific formats and styles essential for maintaining the information's authenticity. Third, often we need to know the contextual circumstances in which the digital information is created—especially who created it, for what purposes, and how they organized and processed it—to understand it.

Content. The initial problem with digital preservation is the content itself. When preserving objects in the analog world (for example, books and music), we deal with static objects; grabbing all of their contents and storing them in some form is simple. Dealing with the digital content requires us to reconsider the meaning of preserving the object. For instance, Web sites have links that not only change but point to dynamically changing sites. As the object grows and changes over time, new questions emerge about what it means to preserve a digital object.

Formats and styles. Computer technology facilitates creating an increasing diversity of recorded information forms. Traditional forms—books, maps, photographs, and sound recordings—continue to be produced in digital formats. New formats have emerged, such as hypertext, dynamic pages, geographic information systems, multimedia, and interactive video. Each format or style poses distinct challenges (for example, encoding and compression) for digital preservation.

If we hold on to digital information without modifications, accessing the information will become increasingly more difficult, if not impossible.

Context. We must use basic research to help develop an understanding of the contextual circumstances. This research is necessary to help accomplish the following: define, develop, and evaluate techniques to preserve the content, formats and styles, and context of digital information across information-technology generations; develop criteria and methods for assessing and demonstrating the preserved information's authenticity; and articulate methods for assessing the long-term value of digital information, including methods for specifying what it says about the past or present and for exploring what it might say about the future.

Storage media

Digital preservation is plagued by the short media life, obsolete hardware and software, and slow read times of old media. Rapid technological advances do not solve the problem; instead, we need to migrate digital materials from one technology generation to another every few years. For digital records, the preservation issues extend beyond media life considerations. Devices for reading these media rapidly become obsolete; the various formats for digital documents and images introduce additional complications. Using research to develop policies, procedures, standards, and protocols based on solid frameworks provides accurate concepts and essential attributes of preservation in the digital information life cycle.

Systems technology

The current communication infrastructure—the Internet, personal computers, cable television, and data storage—has changed the digital information preservation landscape. Considering preservation issues at every life-cycle stage is essential. Although the solution will not be purely technological, understanding the long-term research issues of digital information preservation is critical.

The PITAC Report[1] described four key information technology research areas—software, scalable information infrastructure, high-end computing, and socio-

economic impact. These research areas are all relevant to data archiving and information preservation issues. Software must be reusable and capable of preserving information. The information infrastructure must be scalable in time to preserve digital information. High-end computing generates scientific and nationally critical data, which must be preserved over the long term. Preservation issues play a central role in the "socio-economic" research area. We recommend that digital preservation play a central role in all PITAC activities.

Workflow process

The proliferation of computers makes producing and disseminating information easier. However, it also raises the bar with regard to our ability to use the information efficiently. We can access information stored anywhere in the world over the Internet, but we often have no idea who provides this information, whether the data is valid, or whether it actually appears as it was intended, let alone what it really means.

These problems confront us during the infancy phase of the digital age, but their significance and impact will be compounded in the case of transmitting documentary materials across generations. Workflow process, a concept used in office automation and electronic commerce, will become essential to delineate information values, operational systems versus archives, working document versus archived record, and other information access issues.

The possibility of rapidly accessing information stored anywhere in the world, and the development of increasingly powerful, flexible, retrieval tools raise both the demand for access and the level of service expected in response. Increasing demand invalidates the established model of archives and libraries as places where people go to get information. In the digital environment, the repositories must deliver the preserved information to users anywhere in the world. The solutions, which facilitate migrating digital records across generations of both technology and people, must be capable of interfacing with the digital access systems shown in Figure 1. Interfaces—between preservation systems and access systems and between access systems and end users—must protect the integrity of the recorded information that is delivered through them, yet remain interoperable at all times.

Metadata policies

The disparity between the creation and future use contexts requires increasing quantities of metadata associated with preserved digital records. We should continuously update some metadata, such as copyrights and workflow data.

Despite the feasibility of storing data, the costs incurred in providing and managing adequate metadata will be high. How much metadata is needed to preserve the unknown category effectively? We need policies and automatic tools to translate those policies into metadata decisions. For each preserved record, metadata represents the following: standards, languages, data structures, linkages to logically or physically remote ancillaries, presenters' and interpreters' software, terminology (document type definition, reference, and ontology), and contexts. Short media life and obsolete hardware and software are beyond our control, but sound metadata policies will alleviate these critical issues.

The metrics for archival decisions (for example, metadata) depend on economic and social models, storage and software costs, and human resource costs. Although more semantics in metadata will increase

Current Activities

The research opportunities for preserving digital information are abundant, including the study of specific future requirements and the development of augmented digital technologies. We need near-term research in the following areas:

- interoperability of archiving and digital library systems;
- standards for archiving and preservation;
- workflow process and information preservation;
- metadata for archiving and preservation;
- multimedia container software;
- archival metrics; and
- social and economical models of archives and digital libraries.

To address the digital preservation problem, the National Science Foundation sponsored a Workshop on Data Archiving and Digital Preservation (http://cecssrv1.cecs.missouri.edu/NSFWorkshop/). The workshop's goal was to identify research trends concerning the preservation of archival and information data. The recent NSF ITR Initiative lists digital preservation as a research area (NSF homepage, http://www.nsf.gov/).

In the InterPARES (International Research on Permanent Authentic Records in Electronic Systems) Project, the world's archival community is joining together to tackle the "electronic records problem" in a comprehensive, collaborative manner (http://www.interpares.org/). Canada's Social Science and Humanities Research Council, the US National Historical Publications and Records Commission, and the Italian National Research Council have provided funding for this project. Other participating countries include The Netherlands, Germany, France, Spain, Portugal, Sweden, England, Ireland, Australia, China, and Hong Kong.

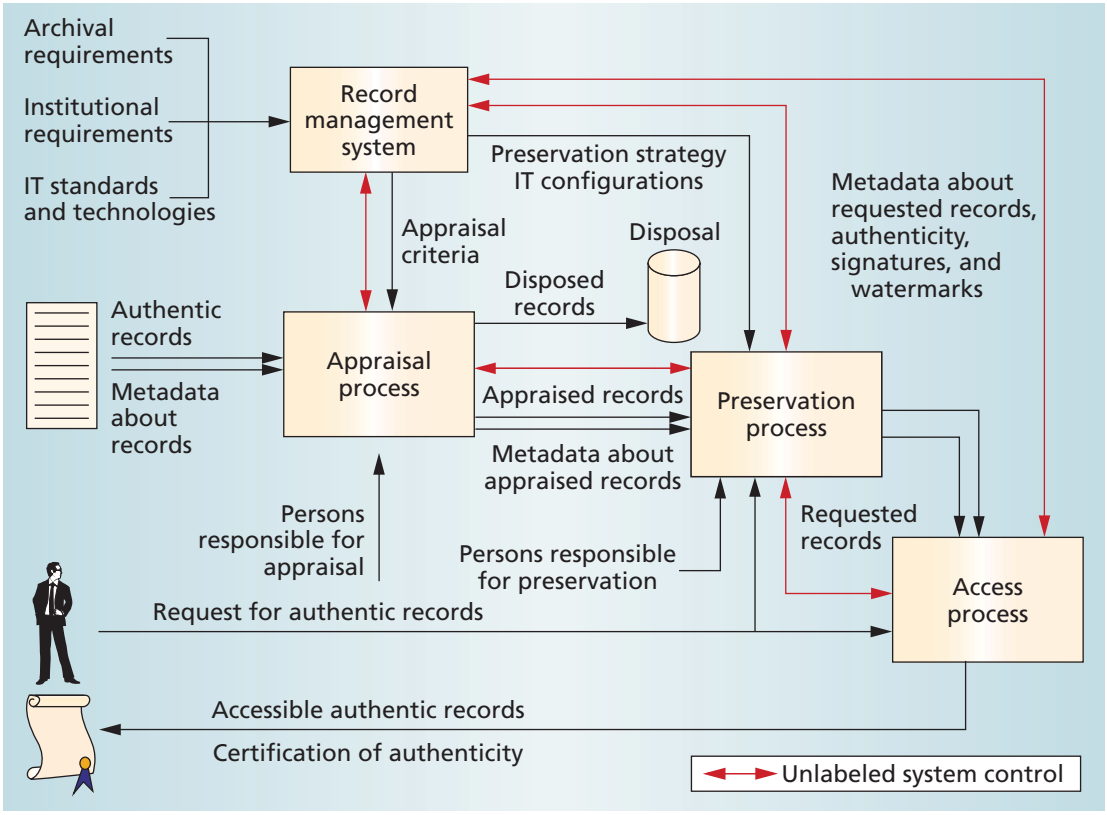

Figure 1. A digital appraisal, preservation, and access model. Rapidly accessing information stored anywhere and developing powerful, flexible retrieval tools are critical. The proposed model permits migrating digital recordings across generations of technology and people.

costs, it will minimize human intervention in accessing data; seamless support, transition of stewardship, and lifetime maintenance will improve.

Preservation is not an exact science, only probabilistic. The information content company, Corbus, experienced a loss rate of 150 of 1 million images in storage systems around 1999. What will be the maximum acceptable loss rates? Quantitative metrics should include a cost/benefit analysis in particular domains.

On the one hand, domain specialists can initiate specific future requirements based on known and understood requirements. These requirements will impact the system design for archival purposes and preservation. On the other hand, as digital technologies proliferate, individuals have personal archiving and preservation needs (for example, digital camera images). Sample research issues include central versus decentralized systems, multilevel control, scalability, reliability, redundancy, access linkages, and enforcement techniques of standards.

In organizations, workflow processes continuously create documents and records. Starting from creation and ingestion, we should integrate the workflow process with the preservation process: appraisal, verification, maintenance, and, eventually, retirement. Augmenting digital technologies to develop automated tools enables archivists to select and attach adequate metadata to a digital record, access information at many service levels (professionals, public interest, and individuals), and interchange documents beyond what is now available.

The software industry's container idea (for example, Apple's Bento) is useful for bundling metadata with records for aggregation, interchange, and emulation.[2] The Reference Model for an Open Archival Information System is an ongoing NASA standard effort for digital preservation.[3] The UPF (Universal Preservation Format) project is an excellent example of container standards for multimedia content, including text, images, voice, video, and their interlinkages.[4]

ESTABLISHING A RESEARCH AGENDA

Social, legal, and ethical issues are important in establishing a research agenda for digital preservation. For each issue, we must address important questions, such as what should be preserved for the future.

Businesses, governments, and individuals increasingly use digital technology to conduct their affairs. Terabytes of new content are added to Web sites every month, and the growth rate is escalating. Satellites steadily transmit vast quantities of data to Earth. Although we can use technology to preserve such vast quantities of data, technology cannot determine what data we should preserve. Making this decision is, in essence, a value judgment involving complex factors, rather than a scientific or objective determination.

We may decide that we should preserve specific information because of what it says about the present or the past. From this perspective, the future value of stored information depends on its inherent characteristics: what it is about and how accurate, complete, reliable, and comprehensible it is. The value of information also depends on what it has to offer the future. Identifying and measuring this value are not simple procedures. Information technology advances make such value judgments more difficult. Thanks to today's technology, we can now mine data and create information previously considered impossible. Future technological advances will expand the horizons for generating knowledge.

Decisions about preserving information should consider the costs. We can use current technology to determine the costs of retaining information; however, both expenditures and technology will evolve. Whereas we can project the costs for basic elements of technology—such as digital media per unit volume of information and unit processing by computers—there are no proven techniques for estimating the costs of long-term digital information preservation.

Now that we can make information available to communities worldwide via the Internet, we face the challenge of preserving digital information with its paradox of short media life, obsolete hardware and software, slow read times of old media, and defunct Web sites. Despite the wealth of accumulated, technology-generated information, we currently lack proven methods for preserving this information or for using optimal technology tools to access it and determine its authenticity.

Failure to address these digital preservation problems is analogous to squandering potential professional, personal, and economic gains, contributing to cultural and intellectual poverty, and resulting in exorbitant costs for recovery. We are compelled to meet the research challenge to resolve the conflict between the creation context and the use context to facilitate digital information preservation. ✸

Acknowledgments

I thank the NSF Workshop on Data Archival and Information Preservation participants and the InterPARES Project members for providing valuable information.

References

1. "Information Technology Research: Investing in Our Future, PITAC (President's Information Technology Advisory Committee) Report," Feb. 1999, National Coordination Office for CIC, http://www.ccic.gov/ac/.
2. J. Rotherberg, "Ensuring the Longevity of Digital Information," *Scientific Am.*, Jan. 1995, pp. 42-47.
3. "Reference Model for an Open Archival Information System," http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html.
4. UPF Home, http://info.wgbh.org/upf/.

Su-Shing Chen *is a professor of computer engineering and computer science at the University of Missouri-Columbia. His research interests include digital libraries, Web technology, and bioinformatics. Chen received a PhD in mathematics from the University of Maryland at College Park. Contact him at schen@cecs.missouri.edu.*