Digital Preservation Annotated Bibliography

Across the country and around the world, libraries, museums and archives are spending an increasing amount of their budgets to purchase and create digital content. What will happen to this investment in the next 5 to 10 years? Without a concerted effort, these digital resources could become obsolete. This annotated bibliography provides a very brief glimpse into the complicated world of digital preservation.

Gladney, H. (2004). Trustworthy 100-Year Digital Objects: Evidence After Every Witness Is Dead. ***ACM Transactions on Information Systems, 22(3),*** **406-436.**

Henry M. Gladney has a Ph.D. in Chemical Physics from Princeton. He worked at IBM as a researcher from 1963 until 2000 and designed the IBM Digital Library software product (available in 1993). Currently, he consults on management and technical issues surrounding digital content management and digital preservation. He is an ACM member and a Fellow of the American Physical Society. He is widely published and is responsible for publishing the Digital Document Quarterly.

This very long article is a proposal to build a system to encrypt and archive digital objects with their contextual metadata. Using existing infrastructure and theory, the author designed a workflow for data producers that would create a Trusted Digital Object (TDO). This object would be encoded to prove the authenticity of the underlying bit-stream. The validity of the data would need to be verified by humans.

The article mentions a number of very important concepts in digital libraries and digital preservation:

- the need for trust in the institutions that will archive digital objects
- the need to trust that the bit-stream has integrity
- end-user's basic requirements for digital objects
- the need for unique and persistent identifiers like URNs and URIs
- the Reference Model for an Open Archival Information System (OAIS) conceptual framework
- private/public key encryption
- Open Archives Initiative – Protocol for Metadata Harvesting (OAI/PMH)
- digital audit trail

But the article is provocative. It suggests a number of controversial practices and glosses over a great number of very difficult issues:

- storing the entire TDO vs. building it for delivery
- embedding metadata vs. binding metadata
- reusing identifiers for new editions
- centralized services for resolution and other important functions
- storing and migrating files
- hardware and software obsolescence

The author was overly ambitious in his attempt to develop both the theory and the practice of trustworthy digital objects.

Gilliland-Swetland, A., Eppard, P. (2000). Preserving the Authenticity of Contingent Digital Objects. *D-Lib Magazine, 6(7/8).* (http://www.dlib.org/dlib/july00/eppart/07eppard.html)

Dr. Anne Gilliland-Swetland has 8 articles lists in the DBLP bibliography server. She is an Associate Professor in the Department of Information Studies at UCLA. She has an M.A. from Trinity College Dublin, an M.S. from the University of Illinois at Urbana-Champaign, and a Ph.D. from the University of Michigan. Her areas of research are electronic records management, the design and evaluation of information systems containing primary sources, and archival education.

Since 1988, Philip B. Eppard has been on the faculty of the School of Information Science and Policy at the University of Albany, SUNY. He was dean of the school from 1995 to 2003. He has an M.A. from Andover Newton Theological School, a Ph.D. in American Civilization from Brown University, and an M.S. in Library and Information Science from Simmons College. He has been the editor of the *American Archivist*. Dr. Eppard is co-director of the U.S. research team participating in the InterPARES Project (International Research on Permanent Authentic Records in Electronic Systems).

This article is about preserving electronic records which, according to the authors, have the most stringent requirements for digital integrity, authenticity and preservation of any type of digital object. Electronic records have a complex context in which they need to be preserved – juridical-administrative, procedural, provenancial, documentary and technological. Electronic records will "need to be both fixed and mutable when accessed for different purposes."

The InterPARES project is a large research project designed to identify the problems of maintaining electronic records – for long-term preservation and guaranteeing authenticity. A multi-national, cross-industry effort, this project has four research domains: authenticity; preservation; appraisal; and policies strategies and standards. The theoretical framework for InterPARES is an 18th century European analytical approach to the determining the authenticity of ancient documents. Diplomatics "studies the genesis, forms, and transmission of archival documents; [and] their relations to the facts represented in them…" The goal of this analysis is to develop the requirements for the authenticity of the records. The project has developed a Template for Analysis – a model of an ideal record that contains all of the possible element types. The project intends to develop a predictive model to help archivists preserve their data. Using IDEF0, a modeling language described at http://www.microafrica.co.za/tass/idef0.htm, the teams are creating models for preservation and appraisal.

Rather than using statistical sampling, the project is using case studies to refine the Template for Analysis. Using a Template Element Data Gathering Instrument, they populate and refine elements in the Template for Analysis. The case studies are

"interpretive and are directed towards not only understanding the elements of form of electronic records but also the situatedness of those records within their various contexts…" Case studies used so far are large databases, geographic information systems, and web-based applications.

Unlike e-prints (which will be discussed later), electronic records present a huge challenge to the digital preservation community. The issues abound – technical, temporal, administrative, and contextual. Without a substantial, cross-domain, international effort, creating consensus for defining these issues, developing models, and proposing solutions would be impossible. This project could be an important part of that effort.

As all of these project report articles, this paper gives broad strokes, big picture, high level outlines of the project with very little detail. The reader has no real basis for judging the quality of the work or deliverables. For this particular project, enough information was given on the methodology to raise some concern. While the reader can empathize with the concerns that statistical sampling of records would be difficult and perhaps even misleading, one worries that using only case studies will skew the work even more. InterPARES will need to use real datasets to test its theories.

Hart, P., Liu, Z. (2003). Trust in the Preservation of Digital Information. ***Communications of the ACM, 46(6),*** **93-97**.

Dr. Ziming Liu has 6 publications listed on DBLP Bibliography Server. After receiving a Ph.D. from UC Berkeley, Dr. Liu was a researcher scientist at Ricoh Silicon Valley, Inc. and taught at the University of Washington before becoming an Associate Professor at School of Library and Information Science, San Jose State University. Dr. Lui's research interests are "user behaviors in electronic environments, management of information organizations, social studies of information, international and multicultural issues of information services, and trans-border information flow."

Dr. Peter Hart earned an undergraduate electrical engineering degree from Rensselaer Polytechnic Institute and an M.S. and a Ph.D. degree from Stanford University. Currently, Dr. Hart is Chairman, President and Founder of Ricoh Innovations, Inc. His research interests are management of innovation, multimedia document analysis and communication and information appliances.

This short essay is an attempt to study the issues of trust in digital preservation and to begin the process of developing a model for inter-organizational support for digital preservation. To determine the level of trust people have regarding digital preservation, the authors surveyed 110 individuals "with extensive experience in handling electronic information, such as students, teachers, engineers, and office workers." Of the respondents, 86% state that they would keep paper copies of important digital documents. The authors have determined that there are 5 reasons that people do not trust their electronic documents:

- *Inaccessibility* – people want to be able to get to their data without barriers of electrical power, hardware reliability or reading devices.
- *Lack of tangibility* – people want a solid thing. "'Invisible electromagnetic bumps on the plastic disk'" do not feel real.
- *Fluidity* – people feel that the electronic copy can be easily altered while the paper copy is fixed in time and space.
- *Short preservation period* – people understand that the physical media of digital storage is short lived and that to keep their electronic documents viable they need to be active. Paper can live in a file drawer for a long time without any human intervention.
- *Privacy and security* – people think that electronic information can be more easily compromised than paper.

The authors developed a very thoughtful comparison of paper documents and monetary currency. Both paper documents and currency are stores of value and mediums of exchange. The paper document's store of value is its information. As a medium of exchange, a paper document can be consumed (i.e. read) and used (i.e. annotated). If people do not trust electronic documents, why do they trust electronic currency? The answer is institutional trust – people trust their bank's system. When they see a bank statement, either on paper or on an ATM screen, they do not need to see the actual cash. Institutional guarantees are the key. Institutional trust is also a key to increasing trust with digital media. The authors developed the following general characteristics that trust:

- *Increases with familiarity.*
- *Is linked to a given condition.*
- *Requires accountability and tangibility.*
- *Is often associated with scale.*

While the discussion of trust was excellent, the conclusion of this article has several significant weaknesses. The survey upon which many of the conclusions are based is not fully disclosed in the article. Only a few questions and a several summary statistic are enumerated. The rest are vaguely described in a sidebar. In the concluding paragraph, the authors casually state that "as of this writing, the marginal cost of disk space for storing a document page image is approaching 150 times less than the cost of the paper the document is printed on, a huge gap that continues to widen." The comparison is invalid – we cannot compare the underlying numbers because they are not equivalents. Both paper and digital storage need infrastructure. Paper needs file cabinets, floor space and people to handle it. Digital storage requires computers, disk management software, electricity, often special environmental conditions, people to manage the systems. True costs depend on conditions, scale and time frame. To toss a number without real meaning into the conclusion of such a thoughtful piece undermines its importance.

While the "150 times" number is aggravating, the real weakness with this article is its lack of definition of digital preservation. In their conclusion, the authors conflate institutional electronic records management with personal record keeping, archiving with business practice, preservation with data storage. While they state that "preservation of digital information is an extremely complex issue," they did not even attempt to define

different levels of preservation. The authors did not differentiate between established, highly regulated businesses with clear statutory requirements – banks and insurance companies – and new information industries like content producers, document companies and professional organizations. By the end of the conclusion, the reader is confused as to what problem the authors are trying to solve.

National Science Foundation, Library of Congress. (2003). *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation.* Final Report: Workshop on Research Challenges in Digital Archiving and Long-term Preservation.

Rather than an article published in a journal, this is a report published jointly by the National Science Foundation (NSF) and the Library of Congress (LC). Together, they sponsored a workshop to discuss and document the research challenges of digital preservation and archiving. With a gathering a number of leading thinkers in digital libraries and preservation, the workshop was able to produce an excellent report outlining, not solution, but questions for future study.

They developed a framework for long-term digital preservation:

- Digital objects require constant and perpetual maintenance.
- Accelerating rate of data collection and creation are escalating problems of scale.
- As Digital objects become more complex, the problems increase.
- Libraries, archives, and museums need solutions to these challenges to meet their missions. Governmental agencies and individuals are also vested in finding solutions.
- People are an essential part of the process.
- Funding is a huge issue.
- The problems of long-term preservation are as much social as technical.

There are four main areas of research:

- Technical architectures for archival repositories
    - Capacity and scale
    - Relationship of objects – how to manage the complexity
    - Effective validation of content
    - Interoperability
    - Rights management
    - Schema, ontology and metadata management
- Attributes of archival collections
    - Appraisal and selection criteria within complex objects
    - Definitions of acceptable levels of information loss
    - Develop collection-level metadata schemas that allow for inheritance of metadata
    - Emulation methodologies
    - Metrics for measuring the quality and fidelity of preserved digital objects
- Digital archiving tools and technologies

    - Tools to automatically transform disparate types of objects into standard file types for ease of management
    - Tools for persistent object identifiers and authorization
    - Object version control
    - Standard and stable representations of objects (text, sound, moving images)
- Organization, economic and policy issues
    - Develop metrics for cost, effectiveness, performance for preservation methods, hardware platforms and functionality of systems
    - Funding issues – revenue streams, tax incentives, market analysis for systems and services

Not only does this report suggest research topics, it proposed a number of research scenarios:

- Theory-building – on digital objects, authentication, economics
- Exploratory – test alternative architectures and methods
- Simulations/experiments – policy, tools
- Observational – user studies with content in alternative formats
- Testbeds – develop metrics from existing databases

This report accomplishes its goal of laying out a ten year agenda for research for digital archiving and long-term preservation. Any Ph.D. student needing a thesis topic would do well to consult this report.

Pinfield, S., James, H. (2003). The Digital Preservation of e-Prints. ***D-Lib Magazine, 9(9).*** **(http://www.dlib.org/dlib/september03/pinfield/09pinfield.html)**

Stephen Pinfield is Assistant Director of Information Services at the University of Nottingham, United Kingdom. He is an active researcher and developer in the field of scholarly communications. He is a member of the Consortium of University Research Libraries Task Force on Scholarly Communication and the Society of College, National and University Libraries Advisory Committee on Scholarly Communications.
He is the director of a multi-institutional project, known as UK SHERPA, to set up e-print repositories in the UK.

Hamish James is the collections manager for the Arts and Humanities Data Service at King's College London. He has written several articles on digital libraries, preservation and data standards.

Should e-prints be preserved? This is the question. This article provides an interesting and unique perspective to digital preservation in general and e-prints specifically. The authors define "e-prints" as "electronic versions of research papers or similar research output," either pre-prints or post-prints. A number of open-access online repositories are often referred to as "archives" (sometimes because of the misnamed Open Archives

Initiative Protocol for Metadata Harvesting), but this term does not "necessarily imply a curation or long-term preservation function."

The case against worrying about long-term preservation of e-prints (a summarization of the opinions of Stevan Harnad):

- e-Prints are duplicates of published literature.
- e-Prints repositories were created to provide immediate, free access.
- The focus should be on filling repositories.
- With the focus on volume, preservation will be a distraction
- Preservation will be a barrier because of all of the submission requirements that authors will have to deal with.
- Preservation efforts should be focused on the conventionally published materials that are truly endangered.

The case for preserving e-prints (a summarization of the opinions of Peter Hirtle):

- Preservation is necessary to continue to provide open access.
- Preservation is necessary to ensure citations will continue to be valid.
- e-Prints have additional information that will be lost if e-prints are not preserved.
- Preservation is important when a repository has a coherent collection of e-prints.
- Guarantees of preservation may attract authors.

The authors propose that both are possible. Filling repositories is important, and efforts to preserve the data can be accomplished as well. Describing their project, UK SHERPA, the authors assert that libraries are the "natural home for [preservation] since they have a tradition of managing access to information resources and preserving them into the future." But libraries cannot do this important work without the support of the entire academic community.

Preserving e-prints is not as technically difficult as other digital objects. Because they are "paper documents made electronic," they normally contain only text and still images. It will be important for repository managers to know versions of HTML and types of images. Over time, e-prints will evolve to include moving images, sound and more external links. These issues will need to be resolved. But rather than focusing on the technical issues, the authors addressed many of the social and organizational issues. Institutional commitment is vital as the scholarly communications process shifts. Funding needs to become a stable stream, and costs need to be better understood. Preservation appraisal is important because digital preservation, just like physical preservation, is not an "all or nothing" proposition.

This article presents a refreshing pragmatic approach to digital preservation. Understanding the tradeoffs between access and preservation, costs and benefits, "good enough" and perfection is the only way to make appropriate decisions. UK SHERPA has potential to provide a path through the preservation wilderness.