Comparing Preservation Strategies and Practices for Electronic Records

Michèle V. Cloonan and Shelby Sanett, University of California, Los Angeles

This presentation reports on a study we conducted on behalf of the Preservation Task Force of the International Research on Permanent Authentic Records in Electronic Systems, known as the InterPARES Project. InterPARES is an international research initiative that involves national archives, university archives, and a team of academic researchers in archival science, preservation, and computer science to address issues related to the permanent preservation of authentic electronic records. The Project is investigating and developing theoretical frameworks, methodologies, and prototype systems required for the permanent preservation of authentic electronic records. In this paper we will focus on our study, and not on the InterPARES Project as a whole.

This is a report of phase one of a three-part study on preservation strategies for electronic records. We surveyed projects and institutions that are developing, evaluating and/or implementing digital preservation strategies. Our goal was to identify a variety of digital approaches, including migration, emulation, bundling, Universal Preservation Format (UPF), and robotics. We wanted to find out which methods had been selected, and why, as well as how effectively they are working. We selected 14 projects and programs based in Europe, Australia, and the United States; to date we have interviewed representatives from 12 of them.

We had originally planned to report only on the digital preservation strategies themselves, but it is not yet possible to offer substantive observations on some of them because they are still in the developmental stages. In one case, research stopped when the funding ran out. Sources and duration of funding are important issues that we will touch on below. With some of the other projects we investigated, research is moving more slowly than originally anticipated. We plan to conduct follow-up interviews in 2001 and 2002; this survey constitutes the initial phase.

We will report on responses to some of the questions that we included in our survey instrument. The questionnaire was divided into 14 sections:

A. Information about the Institution/Project and Respondent
B. Program and Policy
C. Specifics of Preservation Technique/Method/Strategy
D. Selection for Preservation
E. Cooperation
F. Staffing
G. Technical Questions
H. Costs
I. Preserving Records

J. Description/Documentation of Preservation Processes
K. Access to Preserved Records
L. Charges
M. Reproduction and Copyright
N. Preservation Policies

Not all sections or questions pertained to each project or program. However, we chose to be comprehensive in order to learn about as many aspects of each program as possible. We will revise the questionnaire in the 2001 iteration of the survey. In this iteration we tried to achieve breadth and scope; in the next version we will focus on a few of the above sections in depth.

Three questions in particular yielded stimulating and--we believe--pertinent observations: definitions of preservation (in the *Program and Policy* section), categories of costs (*Costs*), and preservation policies (*Preservation Policies*). Today we will focus on these three.

We have followed the procedures for Human Subjects research established by the University of California. Respondents were guaranteed anonymity unless they provided express written consent. Since this research is ongoing, we have not yet requested such consent and cannot yet reveal the names of individuals or institutions in this presentation. All quotations used here will be designated as Respondent 1, 2, etc., and references to projects as Institution 1, 2, etc.

Definitions of Preservation

In the section on *Program and Policy*, we asked: "How do you use the word ‘preservation’ at your institution?" In other words, what definition does your institution associate with the term ‘preservation’?" We included this question because we felt that such definitions might have a bearing on approaches that projects take to developing digital preservation strategies. For example, depending on your view of preservation you might select one approach over another. Since we interviewed archivists and librarians from 7 countries representing a dozen projects or institutions, we anticipated getting a range of responses. As we continue this study into the next phase of this research, we will try to determine whether or not the definitions of preservation continue to evolve.

Ten respondents defined ‘preservation’: The following key phrases emerged:

Respondent 1: "Preservation for paper records is a regime which tries to slow entropy and avoid degradation. For digital records, it is to preserve the document to perpetuity. Digital Preservation includes issues of authenticity."

Respondent 2: "Preservation means ensuring the object is accessible over the long-term. Access and preservation are separate."

Respondent 4: "Preservation covers all activities directed towards ensuring the ongoing accessibility to the information content of the records. Hence, we consider the ambient conditions in our repositories as a preservation issue, along with the specifications of the media on which recorded information is stored. Migration of digital objects is thus a preservation strategy."

Respondent 5: "The ability to discover, access, and present electronic records through arbitrary changes of technology. We can preserve things forever."

Respondent 7: "Forward migration or prospective preservation to whatever new technologies exist. [We are beyond] thinking about ‘x’ number of years of preservation."

Respondent 8: "Enabling long-term access to materials."

Respondent 9: "Ability to present the record unchanged repeatedly."

Respondent 11: "Everything you have to do to guarantee you can deliver records [and] respecting the sanctity of the original order."

Respondent 12: [The technical and managerial processes that protect the integrity and longevity of materials – regardless

of genre.]

The respondents' key phrases cover three components of preservation: 1. preservation processes; 2. length of time for retention; and, 3. preservation outcomes.

KEY PHRASES FROM RESPONDENTS: DIGITAL PRESERVATION DEFINITIONS

1. Preservation Processes

- all activities directed towards ongoing accessibility
- specifications of the media on which information is stored
- employing digital preservation strategies such as technology preservation, migration, and/or emulation
- discovering/accessing/presenting electronic records
- technical and managerial decision-making

2. Length of Time for Retention

- ongoing accessibility
- long-term
- perpetuity
- forever
- "we are beyond thinking about 'x' number of years."

3. Preservation Outcomes

- ensuring the object is accessible
- enabling long-term access
- maintaining the ability to preserve the record unchanged, repeatedly
- protecting the integrity and longevity of materials
- preserving to perpetuity

Overall, the responses demonstrate a shift taking place from defining preservation as a once-and-forever approach for paper-based materials, to an all-the-time approach for digital materials, that may begin even before a record has been created. (Of course, paper-based materials also require all-the-time care, we are merely drawing a distinction.)

To contextualize the respondents' definitions, we offer published definitions of preservation culled from the archives and the library fields.

The Paper-Based Perspective

SAA, 1974
A. The basic responsibility to provide adequate facilities for the protection, care, and maintenance of archives, records, and manuscripts. B. Specific measures, individual and collective, undertaken for the repair, maintenance, restoration, or protection of documents.

Ratcliffe Report, 1984
Strictly, all the steps taken to protect materials, that is including conservation and restoration, but often used in reference to the treatment of materials on first entering the library; it is preventive rather than remedial.

Transition to Digital

IFLA, 1986
Includes all managerial and financial considerations including storage and accommodation provisions, staffing levels, policies, techniques and methods involved in preserving library and archive materials and the information contained in them.

Feather, Matthews, and Eden, 1996

The managerial, financial and technical issues involved in preserving library materials in all formats--and/or their information content--*so as to maximize life*. [Italics ours]

Digital

SAA, 1997
Preservation of digital information is not so much about protecting physical objects as about specifying the creation and maintenance of intangible electronic files whose intellectual integrity is their primary characteristic. Preservation goes beyond saving such media as optical disks or magnetic tape; the access system itself must be preserved.

Kenney and Rieger, 2000
Digital Preservation means retaining digital image collections in a usable and interpretable form for the long term. While "long-term" suggests an indefinite future, David Bearman interprets it more usefully as ‘retention for a period of continuing value.’

Archival and library definitions have shifted from the physical care and protection of materials to retaining them in usable form for an indefinite amount of time. In the paper-based information world, librarians and archivists sought to preserve books and documents for 500 years or more. As is apparent from both the study respondents, and the professional literature, professionals now think about maximizing "useful life" or preserving digital documents "forever" through emulation or forward migration, but without the emphasis on a specific number of years.

Further analysis of our data indicates that archivists and librarians view preservation through different lenses. This reflects a fundamental difference in the archival and library professions. Librarians tend to be custodians of printed materials that are not unique. Librarianship carries custodial responsibilities, but--with the exception of special collections--missing or damaged items can *usually* be replaced. Therefore, librarians often view their materials in terms of immediate utility. In the archival arena, when a record is gone, it is really gone and cannot be replaced--whether that is due to an accident or a disposal schedule. Archivists have responsibility for one-of-a-kind records, which are lodged in a repository. In current practice, the repository and the object cannot be divorced. This relationship differs from libraries and printed materials. In archives, long-term accessibility to the records may be mandated by legal warrant and business processes, and more broadly, by societal memory. The impact of electronic records may have an effect on the requirements that the repository and the object remain together in archives. In the digital environment *both* librarians and archivists have responsibility for documents and records that are born digital. These digital assets are susceptible to obsolescence and incompatibility. Therefore, the integrity and authenticity of digital objects is of mutual concern to both professions. As librarians and archivists work closely on digital preservation strategies, the definition of preservation may shift to accommodate both professional perspectives.

Costs

In the section on Costs, we asked: "What do you estimate are the costs to preserve the records?" Responses included staff, equipment, space, energy and other related costs.

In essence, we were asking, what is it going to cost the institution to preserve, maintain, and provide access to electronic records? We thought this was an important question because for many institutions and projects, knowing what the bottom line is, is THE major factor which influences decision-making, and determines goals and objectives, as well as the strategies to meet them. The majority of the managers we interviewed are gathering financial data now and plan to report costs as part of their projects’ results. Only a few projects are far enough along to have developed cost figures. The interviewees ranged from large national archives, to projects developing testbeds. The range of the costs for electronic record preservation is from $10,000 - $2.6 million per year. Cost categories include staff, consultants, facilities, equipment, storage system monitoring, staff access and research and development.

Most of the projects are currently funded through initial allocations, and some of these figures reflect the impact of early research and development costs, which could also account for the wide range of costs. In fact, as one respondent said, the costs for his project might be reduced by as much as half during the following year. This question will be followed up as part of the second phase of the research interviews. It will be interesting to see what the forecast figures for preservation, storage and staffing actually turn out to be, especially when the initial costs of research and development are reduced over time.

At the time of these interviews, none of the respondents had yet gathered enough information to determine the categories of preservation costs or cost modeling protocols.

Sources of funding include various government agencies, EU (European Union), NSF (National Science Foundation), NPACI (National Partnership for Advanced Computational Infrastructure), NEH (National Endowment for the Humanities), NHPRC (National Historical Publications and Records Commission), NARA (National Archives and Records Administration), and JISC (Joint Information System Committee). As always, the question remains as to what extent the source(s) of funding have shaped the research agenda and from there, the future.

The follow-up study will gather data on the further development of a preservation cost model. So far, cost modeling for digital projects has received scant attention. The present focus appears to be on budgeting for digital conversions rather than preserving authentic electronic records. In addition, the literature is quite skimpy in the area of cost models for born digital electronic records. Two exceptions are studies by Tony Hendley and by Kelly Russell and Ellis Weinberger. Tony Hendley, in his report on the *Comparison of Methods & Costs of Digital Preservation*, provides a "Table of Digital Preservation Cost Elements" which was compiled by Neil Beagrie, Daniel Greenstein, and the Arts and Humanities Data Service. In it, the cost elements involved in developing and preserving digital collections are keyed to the life cycle stages of a digital resource. In their article, *Cost Elements of Digital Preservation*, Kelly Russell and Ellis Weinberger posit that the ongoing costs of digital preservation span a more extended timeframe than traditional preservation and will therefore require resource commitments of a different nature. Different strategies may necessitate different costing timeframes and schedules. Russell and Weinberger state that current cost models have yet to reflect this more complex environment. They further state that, "The creation of a digital object is the true starting point for digital preservation."

To estimate a budget for image acquisition, Anne Kenney and Oya Rieger refer to the "RLG Worksheet for Estimating Digital Reformatting Costs" in their book, *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. The Worksheet in combination with an assessment of costs derived by Cornell's Department of Preservation identified costs for image acquisition in six cost categories. These costs include personnel, equipment, cataloging, supplies, contingency and overhead/indirects.

A number of these categories may be adapted for the model development in our follow up study, and if so, would include the addition of the following categories: providing access costs to the materials; and costs related to long-term creation and maintenance of digital materials, production of metadata, administration and, research and development.

One respondent provided information about plans to form a consortium of institutions to form a National Preservation Center. This idea should be explored not only because of its potential for cost-effectiveness of preservation, but also for the opportunities to enrich the library, archival and museum professions, which may occur as a result of providing a forum for communication across institutional settings and domains.

In a speech for directors of the Association of Research Libraries, Clifford Lynch stated,

> "The fundamentally hard things about managing bits into the future mostly aren't technical; they're economic and organizational. Bits need care and feeding. They don't do well with benign neglect. This means that we need to come up with financial models to keep these bits cared for and healthy as they are migrated into the future. We don't lose a lot of bits to technical failures in a well-managed environment, but we lose a lot due to financial or organizational failures to maintain that well-managed, caring environment on a continual basis. "

We include this quote to emphasize that technical processes cannot be separated from economic issues. The library and archival professions have not fully grappled with the economic influences on preservation decisions. It is necessary to identify concepts and approaches for evaluating the full economic impact of digital preservation. Institutional, national, and multi-national policies must be put in place to assure preservation in perpetuity.

Preservation Policy

We concluded the survey with the following three questions about policy:

Do you have a general preservation policy that includes records in electronic form?

If not, do you have a policy for reformatting, refreshing, migrating, emulating, or bundling data to newer technological platforms? Please describe any policies you might have that relate to preservation of electronic records.

Only three of the projects/institutions indicated having policies in place; two others are revising existing policies to include electronic records; and one is currently developing a policy that includes multiple media. Two of the research projects indicated that policy development would be an outcome of their research.

This research will try to ascertain whether or not international concern about the longevity of digital information is being followed up in policymaking arenas. We suspect that policy is lagging far behind the development of standards, because the development of good public policy requires the appropriate political climate as well as the cooperation of numerous stakeholders. Further, there must be a legal environment that enables the preservation of digital information. Yet laws may vary. For example, the Berne Convention and US copyright law have significant differences between them. These types of discrepancies may impede the development of consistent, rational public policy.

Conclusion

To sum up, at the present time the interviews indicate three large themes. First, that the perception of what preservation is goes beyond library and archival practice to the media being preserved. Because electronic material is unique and the timeframe involved to preserve and provide access to this material extends into perpetuity, we expect that traditional definitions of preservation may not apply. Indeed, a shift is already apparent.

Second, the rush to develop the technological processes necessary to preserve authentic electronic records appears to be at the expense of directly addressing cost and policy issues at the start of projects. One respondent who is fully funded by his government, put it succinctly when he said, "We haven't yet been asked to measure costs! We don't need to justify costs. Fixed costs are unknown." Another respondent said, "The result will be cost determinations." And a third answered, "[Costs] should be a result of the testbed project." We feel that the problems posed by preserving authentic electronic records permanently, requires the development of a cost model, which will be unique and not a hybrid of existing digital conversion cost models. We agree with Hendley, Russell and Weinberger that preservation begins at the creation of the electronic material. A cost model for preserving authentic electronic records will need to reflect this perspective, which differs from traditional preservation.

We found that staff and equipment costs are the most consistent hard figures available so far, and of course, those will vary over time, which will ultimately connect with developing forecasting strategies. Many of the projects are nascent, and we suspect that for them, answering the survey questions was somewhat of a theoretical exercise. As the institutions and projects continue their progress, we expect to be able to gather hard data during phases 2 and 3 of the survey. Where Phase 1 of the survey has been largely exploratory and covered scope and breadth, we expect in the subsequent phases to explore the depth of the respondents' practices in several areas. We hope that by the conclusion of phase 3, we will have a substantial amount of hard data about institutions and projects that will have been active for at least 3 years. From this, we hope to develop a cost model for preserving authentic electronic records, which can be applied to libraries and archives and perhaps to other communities of practice.

Last, the lack of preservation policies in place is a distinct gap in the research design of many of the projects and possibly reflects a lack of commitment among the stakeholders in institutions. What is the reasoning behind developing policy as an end result of a project, instead of concurrently with its progress? We suspect that meeting the technological challenges of preserving electronic records is more of a priority within these institutions than developing policy and wonder whether as a result, the overall progress in this new arena will be more uneven than is necessary. Several institutions that responded to our survey have had active programs for a long time and we note that often policy evolved, rather than being strategically planned. In the subsequent phases of our survey, we plan not only to explore the why behind the positioning of policy development within the institution, but also the development of its content. We want to determine the role of the stakeholders and the influence of the legal and political environments, which provide the context in which policy is formed.

As a result of the information we will gather over the next two years about evolving preservation practice, we expect to strengthen the foundation underlying the development of the preservation function model, particularly those aspects which concern preservation, storage, and access to authentic electronic records over time. We also hope to provide

information, which will contextualize the work of projects and institutions around the world, and which will ultimately provide a pool of knowledge, which will be of benefit to us all.

A recently completed report on conservation published by the Getty Conservation Institute, raises many issues of parallel concern to digital preservation:

> Broadly, we lack any conceptual or theoretical overviews for modeling or mapping the interplay of economic, cultural, political, and other social contexts in which conservation is situated. Pragmatically, this kind of synthetic overview or framework would make clear how different disciplines can contribute to conservation research. Likewise, it would provide a context for and help to integrate the varied spheres of conservation work, with the ultimate aim of elucidating how conservation can be more effective in serving society.

International conferences such as this one are ideal forums for addressing these issues. We wish to thank the respondents to our survey who were willing to discuss their evolving programs with us.

Introduction **Information Infrastructures Papers** **Preservation 2000 Papers**

Web page maintained by Robin Dale of RLG.
Last updated on 18 December 2000

RLG, 1200 Villa Street Mountain View, CA 94041-1100 USA