Archives and Museum Informatics **11:** 287–292, 1997.

Research Issues in Migration and Long-Term Preservation

MARGARET HEDSTROM
University of Michigan, Ann Arbor, MI, USA

Key words: digital preservation, migration strategies, research issues, software dependencies

Long-term preservation and migration are concerns not only of the recordkeeping community, but also of other professions and institutions with stewardship or custodial responsibilities for information assets. Unanswered problems with preservation and access to information over time, including records in the archival sense, are acknowledged as a limitation to the development of digital libraries, electronic archives, electronic patient records, legal documentation, scientific databases, and administrative support systems. Properly framed research questions could draw support from a variety of sources to address the conceptual, process, and technical aspects of long-term preservation using multi-disciplinary methods and expertise.

In 1996, a task force commissioned by the Commission on Preservation and Access and the Research Libraries Group issued a report on preserving digital information.[1] Although the report has stimulated considerable discussion and follow-on activities in the library community and among preservationists, experts in electronic records management have been critical of the report's conceptual basis and recommendations.[2] As a member of the Task Force, I was an active contributor

[1] Commission on Preservation and Access and the Research Libraries Group, Task Force on Archiving of Digital Information, *Preserving Digital Information* (Washington, D.C.: Commission on Preservation and Access, May 1, 1996).

[2] Bearman, D., "Preserving Digital Information: A Review," *Archives and Museum Informatics* **10**(2) (1996): 148–153; and Cunningham, A., Reviews Section, *Archives and Manuscripts* **24**(2) (1996): 378–380.

to the discussions and the final report, particularly in defining the concept of migration as an alternative to the simpler concept of "refreshing" digital information by copying it periodically to new media. The report itself and the responses to it serve as a useful point of departure for defining a set of research issues around long-term preservation because the report has achieved a high degree of visibility and, at the very least, raised awareness of long-term preservation issues.

The underlying assumptions behind the research questions posed below are that exact replication of digital objects is rarely feasible or cost-effective and that archivists must accept some loss of information when migrating digital information from one generation of technology to the next. These assumptions raise a series of questions about the characteristics of information loss and the specific conditions under which various types of information loss are acceptable.

Systematic research should be conducted to define acceptable levels of information loss during migration and to identify a set of minimal record attributes, which if not retained, would make investments in preservation pointless. These requirements will vary by format and by the circumstances of creation and use of the digital objects. Although current discussions of migration strategies acknowledge the wide variety of formats of digital objects, most migration strategies still address text or bit-mapped images, and there are no proven methods for migrating many formats of digital information. One research strategy is to define a taxonomy of document types which share common attributes in terms of their logical representations and hence are amenable to common migration strategies. At a very crude level, for example, data, text, images (still and moving), sound, maps, relations, hyperlinks, and executables all present format-specific challenges to migration. Migration will involve varying degrees of degradation, some of which may be readily acceptable, while others may destroy the meaning of the document.

Systematic research on migration should also examine the physical attributes of digital objects. Migration of digital objects with color encoding, compression, and encryption are examples of extremely complicated processes that require complex algorithms, software routines, and in some cases, specific hardware. We have little concrete research that can guide decisions about the possible loss of meaning and integrity of a digital objects if their physical structure is altered during the migration process. Some of the specific questions that this raises include:

- Under what circumstances is lossy compression an acceptable storage format?
- When is it necessary to retain the color encoding scheme of a digital object?
- If color is an essential attribute of the document, must the exact color scheme be retained or are degrees of degradation acceptable?
- Is it necessary to retain voice annotations in their original format or is a computer-generated transcript of the voice annotation an acceptable alternative?

These are a few examples of the format-specific issues that the archival community will confront in preserving contemporary electronic records.

It is important to recognize that there are a series of options for migration which have different implications for information loss, loss of functionality, and costs. The draft section on Migration Strategies distributed for this conference identified eight possible approaches to migration of digital information:

- transfer to paper or microfilm store in "software-independent" format
- retain in the native software environment
- migrate to a system that is compliant with open standards
- store in more than one format
- create surrogates
- save the software needed for access and retrieval
- develop software and hardware emulators

These strategies are not mutually exclusive and undoubtedly there are other possible approaches. The challenge is not to settle on any one strategy as the ideal, but to analyze the feasibility, costs, and risks associated with each strategy so that appropriate and cost-effective methods are selected with full knowledge of their impact on the integrity and quality of the information.

A related question concerns the degree of functionality that is needed or desirable when preserving digital information. Most of the computer science research on migration addresses the migration of legacy systems with as much functionality as possible from obsolete to new systems. Brodie and Stonebraker propose a model of decomposition to simplify migration processes and to focus decision makers on the essential functionality that needs to be migrated.[3] The goals of migrating legacy systems are not entirely consistent with archival objectives because organizations often want to move an entire information processing environment from an aging system architecture to a new one. Generally organizations want to migrate not only data (or records) but also to retain the processing capabilities of the system.

The recordkeeping community has not defined which functionality is necessary to retain and which should be disabled in an archival system. There is a consensus that it is desirable, if not essential, to retain the functionality of retrieval and display. The concept of manipulability, however, is more problematic. It goes without saying that the holdings of an electronic archive should not be manipulable by its users. But there are many circumstances when the archives should be able to deliver to its users copies of documents which they can redact, reprocess, combine, and manipulate in a variety of ways. What remains ill defined is any clear set of criteria for manipulability linked to the original purpose of the record, its physical and logical format, or its potential uses.

Another area of research would address the requirements for retaining the relationships among digital objects. Strategies that only address migrations and transformations of discrete and bounded objects will not fully satisfy archival requirements because the relationships among archival documents is an essential

[3] Brodie, M. L. and Stonebraker, M., *Migrating Legacy Systems: Gateways, Interfaces & the Incremental Approach* (San Francisco: Morgan Kaufmann Publishers, 1995).

element of preservation. This raises a series of questions about how to retain hyperlinks between and within documents, relations with relational data structures, and links between digital objects and executables. Addressing the technical aspects of these questions will require expertise in computer science and systems engineering. The archival community, nevertheless, has a responsibility to determine when it is necessary to retain the dynamic nature of electronic records and for proposing alternative strategies when relational structures cannot be migrated across incompatible system architectures.

Retaining the integrity and archival qualities of digital records will depend as much on how well the migration process is documented as on the specific migration strategies that are chosen. Robust protocols for documenting migration are needed so that subsequent users of the records can determine specifically which characteristics of the document were lost in each transformation, why a specific migration strategy was chosen, and under whose authority it was carried out. Documenting information loss during migration also offers an alternative to preserving an exact replica of a document and all of its associated functionality. For example, an alternative to retaining a set of external hyperlinks and all of the associated documents might be to generate a list of hyperlinked documents without the functionality of linking to the documents. If archivists can develop documentation standards for migration processes, then future users will have a basis for assessing how the document they are viewing at one point in time deviates from its first and subsequent instantiations.

These recommendations are all aimed toward a larger goal of developing records that are "self-migrating" and that are stored in systems which can be managed by artificial agents or other advanced tools. Most of the migration that has been carried out to date has required intensive human intervention to analyze the structure of records, assess the quality of associated documentation, and to develop customized routines for reformatting. Such processes are too costly and too labor intensive to scale up to a level that can address the volume and complexity of contemporary electronic records. An alternative strategy is to encourage computer scientists and software engineers to design "archiving agents" or "migration agents" that can detect records in endangered formats, select the appropriate migration strategy, make the necessary transformations, and document the changes. Not being an expert in computer science or systems engineering, I am uncertain of the feasibility of this proposal. Nevertheless, until the archival community specifies what such agents should do, we will not be able to explore this approach.

We also need research and development of cost models for the various approaches to migration. Decision making and preservation planning are compromised by the absence of cost data that addresses multiple formats and multiple options for migration. Such research should relate the losses of information and functionality to the costs of different migration strategies. Data on the costs incurred with various migration strategies would support a variety of policy and adminis-

trative decisions including appraisal and selection, choice of migration strategies, and the organization of the archiving function.

In this discussion, I have assiduously avoided drawing a sharp distinction between long-term maintenance of electronic records and preservation of other types of digital information. Both Bearman and Cunningham in their reviews of the digital archiving task force report stressed the need to differentiate records from other types of digital information.[4] While this point is well taken, it does not really move the agenda forward unless the archival community can be more specific about why such a distinction is important and where and how the requirements for preserving electronic records differ from the requirements for other types of digital information. This raises a series of specific questions about the relationships between electronic records and other digital objects.

- Are the requirements for long-term preservation of records fundamentally different from the requirements for other types of digital information? If so, how are they different?
- Do electronic records require fundamentally different preservation strategies from other types of digital objects, or are they a subset digital information that can be maintained in common systems provided that a few additional requirements are met?
- Is it feasible, useful, and cost-effective to apply the requirements for preserving electronic records (metadata standards, process controls, etc.) to other types of digital objects?

These questions merit serious investigation for several reasons. First, the conceptual distinctions that archivists make between electronic records and other types of digital information are not well understood or widely accepted outside the archival community. Second, the computer and network architectures and the software systems used to generate, transmit, and store digital information are becoming increasingly integrated and inter-operable. Third, economies of scale might be achieved by developing preservation systems that can handle heterogeneous digital objects, of which electronic records are a subset. Finally, there is considerable anecdotal evidence that users want a single interface to vast stores of digital information. Failure to consider possible preservation strategies that can handle both electronic records and other forms of digital information could marginalize digital archives and place the entire archival enterprise in jeopardy.

In this discussion, I have concentrated on a series of requirements, specifications, options, and possible tools for migration. This is only a small subset of the research issues that are pertinent to long-term maintenance of electronic records. This research agenda is motivated, however, by a strong sense that the archival community has oriented its approaches to migration toward the problems and techniques of the generation of technology we are migrating from rather than taking advantage of the possibilities of the generation of technologies we are migrating

[4] Bearman, "Preserving Digital Information: A Review," p. 151; and Cunningham, Reviews Section, pp. 379–380.

to. An effective research agenda should explore how archivists can use emerging technologies to resolve the preservation problems caused by previous generations of technology.