

## Research Challenges in Digital Archiving and Long-term Preservation

Margaret Hedstrom, University of Michigan\*

Research in almost every discipline depends on well-managed, reliable, and readily accessible digital resources. Many digital resources (scientific databases, medical records, government statistics) are accumulated over long periods of time at considerable expense. Some have long-term value for monitoring changes in the natural environment, for analyzing the effectiveness of policy interventions, and for producing a record of scientific and artistic creativity. Many of the digital resources we are creating today will be re-purposed and re-used for reasons that we cannot imagine today.

Future research capabilities will be seriously compromised without significant investments in research and the development of digital archives. The current processes and technologies for digital archiving, which attempt to replicate traditional practices for appraising, acquiring, and managing archival materials, have not succeeded because they are labor intensive and based on assumptions that no longer hold in the digital environment. Digital technologies are shaping creation, management, preservation, and access in ways which are so profound that traditional methods no longer are effective. These changes will require a paradigm shift in research if it is to provide the innovations - - whether theoretical, methodological or technical -- necessary to underpin long-term access to digital resources.

### **What are the major research challenges?**

1. Digital collections are vast, heterogeneous, and growing at a rate that outpaces our ability to manage and preserve them.

Digital preservation concepts, methods, and tools cannot accommodate the complexity and fluidity of dynamic multi-media digital entities. There are no effective (or cost-effective) methods to preserve dynamic databases, complex web sites, analytical tools, or software for the long-term. Yet, increasingly, digital resources are impossible to interpret

or use without accompanying tools for analysis and presentation. Moreover, current methods rely on significant human intervention in for selection, organization, description and access. Human labor is the greatest cost factor in digital preservation – and the one that is likely to increase over time. New methodologies are needed to manage vast quantities of digital data with as little human intervention as possible.

## 2. Digital archiving requirements reflect concerns for *the long term*.

Digital preservation research shares many issues and goals with the broader field of digital library research, but one unique aspect of preservation is its concern with *the long term*, where long term may simply mean long enough to be concerned about the obsolescence of technology or it may mean decades or centuries. When long-term preservation spans several decades, generations, or centuries, the threat of interrupted management of digital objects becomes critical. Unlike many physical objects that can withstand some period of neglect without resulting in total loss, digital objects require constant maintenance and elaborate “life-support” systems to remain viable. We need research to develop systems that are as self-sustaining, self-monitoring, and self-repairing as possible. Redundancy, replication, and security against intentional attacks on archival systems and against technological failures are critical requirements for long-term preservation, as are issues of forward migration.

## 3. The challenges of maintaining digital archives over long periods of time are as much economic, social, and institutional as technological.

There are no formal economic or business models for digital preservation activities. Economic and policy research needs span a wide range of issues such as incentives for organizations to invest in digital archives and incentives for depositors to place content in repositories. These issues are deeply intertwined with questions of intellectual property rights, privacy, and trust. Metrics are needed to measure the performance of various storage systems over the long-term, assess the effectiveness and costs of different preservation strategies, estimate the value of or benefit from archiving services, and conduct market analysis of user demand. Evaluation of digital archiving is impossible without concrete measures of costs, benefits, and value of digital objects.

#### 4. Affordable and reliable digital preservation requires new tools and technologies.

Digital preservation will not scale without tools and technologies that automate many aspects of the preservation process and that support human decision-making. Decision models are needed to support selection, choice of preservation strategies (normalization, migration, emulation), and the costs and benefits of various levels of description and metadata. Digital preservation strategies are “metadata-intensive.” Therefore there is a critical need to develop tools that automatically supply core metadata, extract metadata from resources at ingest, and restructure and manage metadata over time. It is important to recognize that metadata, schemas, and ontologies are dynamic – subject to frequent extension and revision. It will be essential for future users of archived materials to recover and relate the metadata schema used when the entity was created. Managing schema evolution is a major research issue. Likewise, managing the identity of preserved digital objects over time is a challenge for digital archives because the identifiers assigned to digital objects can be changed easily and the technologies for naming and tracking digital objects evolve over time. Research issues in the area of naming and authorization include development of methods for unique and persistent naming of archived digital objects, tools for certification and authentication of preserved digital objects, methods for version control, and interoperability among naming mechanisms used by different content providers. Tools are also needed to automatically transform preserved digital objects from obsolescing to contemporary into the formats, standards, and data models and to document the effects of these transformations.

#### 5. Affordable, sustainable, and effective digital archiving requires infrastructure.

Digital archiving strategies to date have been either institution- or collection-specific. Research is needed on the requirements for a shared and scalable infrastructure to support digital archiving. This would involve the development of more generic processes and tools and creation of shared services across the digital archiving community. The digital archiving community would benefit from shared services such as a format registry with full documentation of all versions of all open and proprietary formats along with tools for

automatic format conversion. A metadata schema registry (or registries) is also needed. Emulation shows some promise as a digital preservation strategy, but its effectiveness and financial benefit depends on the availability of executable software. Developing a software repository (or a few software repositories) would be a major community undertaking which also requires research on software preservation. Finally, a significant part of the infrastructure is the facility for interoperability (especially search) across widely distributed and heterogeneous digital archives.

\*Many of the ideas in this paper were drawn from my participation in two NSF-funded projects: The NSF/LOC Workshop on Research Challenges in Digital Archiving: Towards a National Infrastructure for Long-Term Preservation of Digital Information, April 12-13, 2002; and the NSF-DELOS Working Group on Digital Preservation and Archiving (ongoing, almost completed).

The opinions, conclusions, and recommendations expressed in this paper are mine and do not necessarily reflect the views of the National Science Foundation, the Library of Congress, or the other participants in the workshop or working group.