

DIGITAL PRESERVATION & ARCHIVING SUMMARY

OF A FORTHCOMING EU(DELOS)-US WORKGROUP¹ REPORT²



H.M. Gladney

HMG Consulting
20044 Glen Brae Drive
Saratoga, CA 95070

21 December 2002

SUMMARY

The need for digital preservation touches all our lives, whether we are workers in commercial or public sector institutions, engage in e-commerce, participate in e-government, or merely collect digital photographs. We create, use, and trust e-content, and might expect that this content will remain accessible for whenever we want it. This assumption is not supported by the facts; absent precautions that are seldom taken, many saved documents will become inaccessible in a few decades.

This Workgroup considered current practice for preserving digital information to begin its inquiry into prudent action. After examining the economic consequences of information loss, changing technology, library practices, and on-going research, it identified more than a dozen areas for research and development needed to achieve efficient digital archiving and long-term preservation.

The group sought questions whose answers promise rapid development of tools and methodologies for preserving a significant fraction of the digital information being created. For timely mitigation of looming information loss, the research would have to lead to tools that require minimal human work, that can be used by librarians and information producers with minimal training, and that require minimal disturbance to existing business and institutional practices and organization. Many possibilities among those the group identified can lead quickly to methods for protecting information at risk and avoiding losses that will otherwise occur.

The current article summarizes the forthcoming group report's most important research questions.

RESEARCH OBJECTIVES

Replicating traditional mechanisms for appraising, documenting, and managing archival materials based mostly on paper would be impractical for digital content. This is because **the number of digital entity candidates for preservation is orders of magnitude larger than the**

number saved on paper. However, traditional archival principles and values can be achieved with methods that exploit the advantageous properties of digital encoding.

The full group report will communicate **four research objectives:**

- **Ensuring that our descendants can understand and use any information we care to preserve.**
- **Ensuring that anybody can decide whether saved data can be trusted for his applications.**
- **Replacing human effort for preservation by automatic procedures whenever this is feasible.**
- **Empowering each information producer to create and package metadata and content to minimize what professional archivists must do.**

Research in digital archiving should: (a) focus on specific domains, (b) focus on tangible deliverables, (c) emphasize engineering and computing science more than it has recently, and (d) publicize the value of digital entities for industries and citizens of the 21st century—as intellectual capital.

RESEARCH AGENDA

Some research (§2 and §3 below) will produce benefits quickly. In view of published concerns about imminent content loss, these deserve urgent attention.

The final WG report also identifies policy, organizational, educational and other activities that would benefit from research, but that lie outside the research and development areas traditionally funded by the European Commission and the National Science Foundation. Other funding agencies should support these.

The Workgroup felt that **three research areas are likely to have the greatest impact:**

- **objects that carry their own descriptions,**
- **metadata and evolution of ontologies, and**
- **preservation of complex and dynamic objects.**

§1. Emerging Research Possibilities

1A: Repositories: How can repositories or museums that collect imminently obsolete hardware and software best help with digital preservation? What engineering, software design, and formal testing would make such repositories economical? What is the design of generic connections to enable legacy devices for communication with current computers?

1B: Archival Media: How can current methods for extracting bitstreams from deteriorating media be improved and extended? What inexpensive, long-lived media (and machines to access their contents) might be created and used for archival content?

1C: Salvage and Rescue: How can “digital archaeology” be made practical? How can analytical tools be extended, characterized, and packaged so that expert users can most effectively recover information from files that were not conditioned for preservation? How can such tools be made easy enough for ordinary users?

1E: Documentation of Functionality and Behavior: What approaches to functionality abstraction and representation can help us describe systems sufficiently for reconstruction? How can such representations be used to establish benchmarks for consistency of results across migrations or emulations? What tests might ensure automatic verification that system behavior is not inadvertently changed by preservation action?

1F: Self-Aware Digital Entities: How can existing research into agents and self-awareness among digital entities be extended to support preservation objectives and context sensitivity?

1G: Accelerated Aging: Further work on aging of storage media and systems could inform the development of preservation processes.

§2. Re-engineering Preservation Processes.

In view of continuing exponential growth in the numbers of digital entities, it is critical to minimize human time and expense for preserving any entity.

2A: Staging of Intervention: To what extent and how can preservation functionality be built into digital entities when they are created or into the systems that manage them? What do we really mean by “preservation functionality”? How can we communicate this to persuade system developers to include it?

2B: Automation of Processes: How can institutional processes be automated to accomplish preservation with minimal human intervention? Which processes can be automated and which will forever depend on human value judgment and opinion?

2C: Detecting Trustworthiness and Information Quality: What must be provided when digital entities are saved so that future users can decide whether archived entities are sufficiently trustworthy for their applications?

2D: Scalability: How can we enable an archive to manage 10,000,000,000 digital entities? What metrics can best assess digital repository scalability?

2E: Collection Completeness and Anomaly Detection: What tools can we provide to help end users and archive custodians assess the completeness of collections? Is it possible to differentiate between anomalies and inherent knowledge that is not clearly expressed?

2F: Distributed Storage: Are there ways not yet described to reduce data loss risks (including risks of political suppression) by distribution of content across networks? What impact does distributed storage have on ingest, documentation, and delivery of digital entities? How can distribution enhance discovery and delivery of saved content?

§3. Preservation of Systems and Technology

3A: Formats of Digital Entities: What representation properties enhance or detract from preservation quality? How can implicit redundancy be exploited? What metrics of preservability might help archivists?

3B: Managing Dynamic Digital Entities: What is meant by “dynamic entity” relative to preservation? How are dynamic properties best preserved?

3C: Automated Metadata Creation: How can we package preservation support into tools convenient for authors and editors of digital content?

3D: Long-term Metadata Viability: How can the drift of word meanings and context, or of social and scholarly interests, be reflected in metadata and ontologies?

3E: Multilingual Entities and Technology: How can multilingual interests best be supported by preservation methodology?

3F: Acceptable Loss: How can one estimate risks of loss? What cost-saving measures can be adopted when loss risks are accepted?

3G: Re-purposing: What might be done for optimal eventual information re-purposing?

¹ The workgroup members are Kevin Ashley, Birte Christensen-Dalsgaard, Wendy Duff, Henry Gladney, Margaret Hedstrom, Claude Huc, Anne Kenney, Reagan Moore, Erich Neuhold, and Seamus Ross.

² This summary shortens and adapts a draft prepared by Seamus Ross. Neither draft has been approved yet by the full workgroup.