

The State of the Art in Digital Preservation:

A perspective from the Digital Library Federation

Donald Waters
Director, Digital Library Federation
October 22, 1998

Digital preservation

- Preserving Digital Information: The Report of the Task Force on Archiving of Digital Information (<http://www.rlg.org/ArchTF>)
- Recent developments in the state of the art

10/22/98: 2

The Task Force report

- The limits of digital technology
- A note on definitions
- Preserving object integrity and ensuring persistence
- Key findings and recommendations

10/22/98: 3

The Limits of Digital Technology

Rapid changes in the means of recording information, in formats for storage, in operating systems, and in application technologies threaten to make the life of information in the digital age “nasty, brutish, and short”

- **Question:** In the face of these limits, why preserve digital information?
- **TF Answer:** We are in danger of losing our cultural memory
- Is this a sufficient and compelling answer?

10/22/98: 4

A note on definitions

Digital libraries are “organizations that provide the resources, including specialized staff, to select, structure, offer intellectual access to, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities” (DLF Working Definition)

- The definition of digital libraries is often extended by focusing on one or more of the defining features and ignoring or downplaying others
- The Task Force distinguished *digital libraries* from *digital archives* to draw attention to the preservation function

10/22/98: 5

"Preserve integrity and ensure persistence"

The Task Force on Archiving of Digital Information argued that these are linked functions

- Persistence is defined in terms of the integrity of objects
- Preservation of object integrity is a necessary but not sufficient condition of persistence
- Other factors need to be present to ensure persistence: organizational will, financial means, and legal rights

10/22/98: 6

Integrity of information objects

Defined in terms of these features:

- Content: a continuum of abstraction
- Fixity: information in discrete objects
- Reference: ability to locate an object
- Provenance: origin and chain of custody
- Context: conditions of interaction

10/22/98: 7

The political economy of "ensuring persistence"

A functional analysis of organization

- Manage costs and finance, which are not trivial
- Create an operating environment
- Manage migration

10/22/98: 8

Create an operating environment

- Appraisal and selection: can't keep everything, so what principles govern retention?
- Accession: preparation of objects for archiving, including description, authentication, and security
- Storage: on-line, near-line, off-line; distinguish high-quality archival copy from various transformations for access and use
- Access: facilitate discovery, retrieval and use, and manage intellectual property rights
- Systems engineering to manage migration

10/22/98: 9

Manage migration

A set of organized tasks designed to achieve the periodic transfer of digital material from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation.

- Migration includes refreshing as a means of preservation
- Refreshing is the copying of digital information from one medium to another
- However, it is not always possible to make an exact digital copy of an object as hardware and software change and still maintain the compatibility of the object with new technology

10/22/98: 10

Migration strategies

- Internal
 - Build emulators
 - Change, or “refresh,” the media
 - Change formats
- External
 - Creators must incorporate standards
 - Systems designers must build migration paths: interchange formats to transfer data to the future
 - Use processing centers

10/22/98: 11

Key findings and recommendations

- The first line of defense rests with creators/providers/owners
- A deep infrastructure is needed for the long term
 - Trusted organizations capable of storing, migrating, and providing access to digital collections
 - A process of certification to establish a climate of trust
 - A fail-safe mechanism: certified archives have a right and duty to exercise aggressive rescue
- Recommendations: pilot projects, support structures, development of best practice

10/22/98: 12

Recent developments in the state of the art

- Say prayers?
- Broad discussion in the press
- Digital library service model
- Institutional justifications
- Development of standards: certification and fail-safe
- Enhance the repertoire of preservation methods

10/22/98: 13

Preservation state of the art: say prayers?

- Daioh Temple of Rinzai Zen Buddhism:
“There are many 'living' documents and softwares that are thoughtlessly discarded or erased without even a second thought. It is this thoughtlessness that has drawn the concern and attention of Head Priest [Shokyu Ishiko](#). Head Priest Ishiko hopes that through holding an "**Information Service**" and by teaching the words of Buddha, that this 'information void' will cease to exist”
(<http://www.thezen.or.jp/jomoh/kuyo.html>).
- We need to help him!

10/22/98: 14

Broad discussion in the press

- Rothenberg and Scientific American
- LA Times and Time and Bits conference
- "Into the Future:" U.S. News and World Report, New York Times, Business Week
- James Gleick, New York Times Magazine, "The Digital Attic An Archive of Everything."
 - "Anyone wandering through the Internet might begin to feel that memory loss isn't the problem. Archivists are everywhere, in fact -- official and self-made."
 - BUT, the average life of a web page is 70+ days??

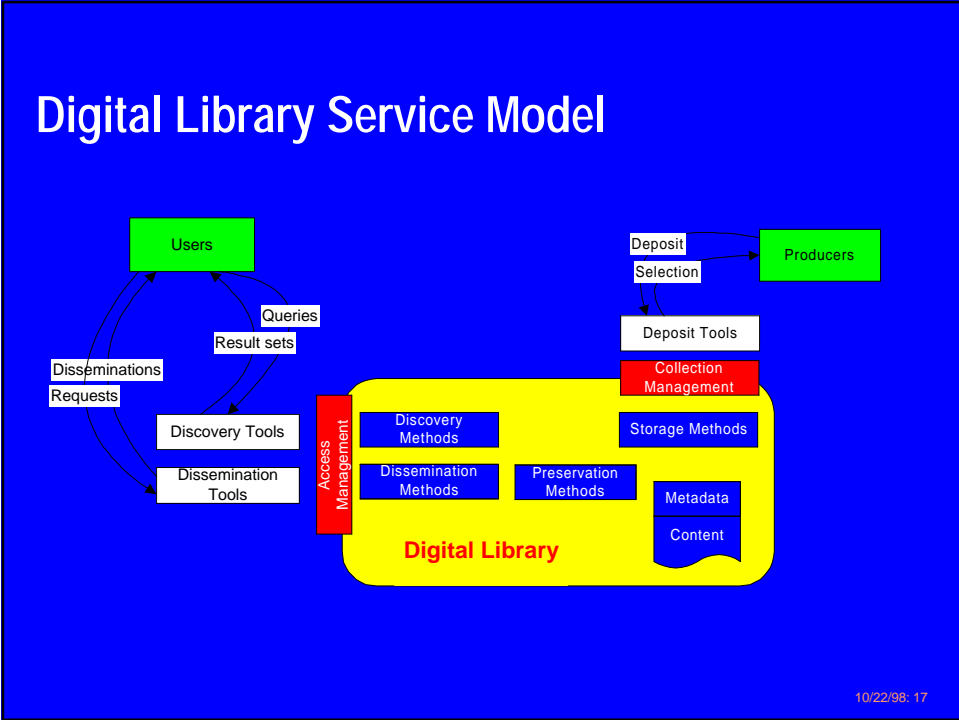
10/22/98: 15

Institutional justifications of preservation

Digital libraries are thought to make a difference in the pursuit of the following goals:

- Organizing, providing access to, and preserving knowledge that is born digitally
- Leveraging the management of intellectual works in support of efforts to redesign the scholarly communication process.
- Providing an accessible and durable knowledge base that improves the quality and lowers the costs of education.
- Providing access to information that is needed to extend the reach of the scholarly enterprise to new audiences

10/22/98: 16



- ### Development of preservation standards
- Need for strategy to develop incentives and protections for fail-safe archives
 - Reference Model for an Open Archival Information System begins to address the issue of certification
 - Developed by the Consultative Committee for Space Data Systems (CCSDS), an international standards development organization at the direction of ISO TC20/SC13
 - <http://ssdoo.gsfc.nasa.gov/nost/isoas/us12/call.html>
 - In the service model, what metadata and methods are needed for preservation that are not already provided for other purposes?
- 10/22/98: 18

Metadata for preservation

- Descriptive
- Administrative
- Structural

10/22/98: 19

Descriptive metadata:

- Data about a resource that relieve potential users of having full access to the resource in order to know its existence or characteristics in relation to a particular information need
- Reference: Critical importance of naming

10/22/98: 20

Administrative metadata:

- Data about a resource that facilitate preservation, collection management, and access management
- Context, provenance, fixity (object authentication), user authentication and authorization

10/22/98: 21

Structural metadata

- Data about a resource that describes its internal structure and serves to organize its delivery
- Need for parsimony of genres

10/22/98: 22

The repertoire of preservation methods

Approaches

- Better media
- Refreshing bits (copying)
- Migration of content
- Emulation
- Archaeology

10/22/98: 23

Enhancing the repertoire

DLF activities:

- Migration in *Making of America, Part III*
- Cornell study of risk management and migration
- Internet-based file conversion service for migration
- Jeff Rothenberg and emulation

10/22/98: 24

Internet-based file conversion service

- Research and prototype developed at CMU by John Ockerbloom (<http://www.cs.cmu.edu/~spok/thesis.html>)
- Includes:
 - An object model for describing the structure of formats and relations among them
 - A distributed network of mediator agents for conversion
- Many-to-many type design facilitates systematic migration and on-demand conversion
- CMU and DLF working to move the research into a production service

10/22/98: 25

Rothenberg and emulation

- Migration is "an approach based on wishful thinking:" unpredictable, labor-intensive, and ultimately unreliable
- Research needed to develop emulation as a means of using an object's original hardware and software environment
 - Develop ways of encapsulating objects and software
 - High level emulation specifications for hardware
 - Annotation mechanisms about the encapsulated objects and software

10/22/98: 26