Digital Library Curriculum Development

Module: 8-a: Preservation

Draft, 12/13/09

1. Module name

Preservation.

2. Scope

This module covers the general ideas, strategies and challenges for the long-term preservation of the digital information.

3. Learning objectives

The student will be able to:

a. Describe the task of determining the properties of digital objects.

b. Explain the basic strategies and methods related to digital preservations.

c. Explain the economic issues involved in the digital preservation.

d. Explain the fundamental challenges and concerns associated with long-term digital preservation.

4. 5S Characteristics of the module

- Spaces: The physical storage for keeping the preserved objects. Vector/probability/feature spaces for content surrogates and processing.
- Streams: Preserving the digital objects, which could be text documents, images or multimedia, are the main topics of this module. Streams are present in the digital objects in the form of bit streams, streams of pixels, etc.
- Scenarios: collect digital objects like digital images, texts, audios, videos, etc… and ingest them into Digital Library with a systematic plan for preserving them over the long term. Policies related.

- Societies: Individuals, groups and organizations involved in setting policy and in carrying out digital preservation policies, the creators, publishers and users of the digital objects.
- Structures: organizational structures. Metadata (we follow some structure in building the metadata for digital object and databases (it depends on a formal structure for building their tables, entities, relations and attributes.

5. Level of effort required

a. Prior to class: 4 hours for reading/preparation of the required readings.

b. In class: 4 hours

1. 2 hours for understanding the digital preservation approaches, strategies and challenges.
2. 2 hours for participating in the learning activity.

6. Relationships with other modules

Close connections with:

- Should be taught before or in connection with 2-c (8-c) File formats, transformation, and migration; to understand the preserved digital objects and how to deal with it.
- Should be taught before or in connection with 8-b module; Web archiving. Web arching is dealing with storing the World Wide Web contents for future needs and then it's highly related to digital preservation for the digital objects because the World Wide Web is a kind of digital objects.
- Should be taught before or in connection with 9-f: module Cost/economic issues much of this Digital Preservation module addresses a subset of cost/economic issues, viz. those associated with digital preservation decisions.

7. Prerequisite knowledge required

- None required, though an understanding of the main components of a computer system (hardware and software) will allow students to better appreciate and understand the fundamental issues of digital preservation and distinguish between the available technical strategies.
- Understanding the fundamental ideas of long term storing, manipulating and managing digital objects.

8. Introductory remedial instruction

Ask the students to take notes for the key points in the class to help them prepare the Exercises/Learning activities.

9. Body of knowledge

1. Introduction

In the world of traditional collections, the principle obstacle to preservation is entropy. Physical materials suffer damage and decay: the acids present in paper damage its fibers, causing it to become brittle and discolored over time; color dyes in photographic films and prints continue to be chemically active, fading through exposure to light or high temperature. Such concerns also apply to digital objects: the physical storage media will degrade over time, or may become corrupted. However, digital preservation must also overcome a unique and much more significant challenge-that of technological obsolesce. Digital information is stored in the form of bits – ones and zeros which denote values in binary notation. Theses bits have no inherent meaning. But rather represent the encoding of information in accordance with some predefined scheme. Such information can't be directly interpreted by a user, but rather requires the mediation of software capable of translating that information into human-readable form. {Deegan, M., & Tanner, S. 2006}

The digital birth of cultural content and conversion of analogue originals into bits and bytes has opened new vistas and extended horizons in every direction, providing access and opportunities for new audiences, enlightenment, entertainment and education in ways unimaginable a mere 15 years ago. Digital libraries have a major function to enhance our appreciation of our engagement with culture and often lead the way in this new digital domain we find ourselves immersed within.

Digital libraries play an important role in preserving culture and in connecting people with their national and regional identities, this work is essential to support the very foundation of our civilization, which is based upon our ability to pass information and knowledge, whether technical or cultural, from one generation to the next.

2. Nature of Digital Objects

Digital objects are sets of instructions for future interaction

- Digital objects are useless (& don't even exist) if no one can interact with them.
- Interactions depend on numerous technical components
- Only a small part of preservation work is about treating them like physical artifacts.
- Rothenberg takes this even farther, contending that all digital objects should be seen as programs.

Bit-Level Considerations

- Physical media should be stored in appropriate environmental conditions.
- Take care in handling of media.
- Maintain integrity of bit stream through security, checksums, periodic sampling & other validation.
- Bit rot & advantages of newer media both call for periodic refresh & reformatting.
- Ensuring the integrity of the bit stream in such transfers is extremely important.

New Conception of "Long-Term"

> "A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future." (OAIS)

Problems with digital data

- Data at risk because they are recorded on a transient medium, in a specified file format, and they need a transient coding scheme (a programming language) to interpret them.
- Digital data can be highly complex, and meaning derived from data can depend as much on how individual's data objects are linked as on what those objects are.
- With digital data, a machine needs to be interposed between the data and the human interpreter, which adds another layer of complications.

3. Digital Preservation Approaches and Strategies Figure 1

How is digital data to be preserved

There are two key issues for data preservation

- Preserving the physical media on which the bitstream is recorded.
- Preserving the means of interpreting, reading and utilizing the bitstream.

Given that the bitstream is merely a very long series of binary codes; the preservation of the physical media should maintain its integrity over time. However, being able to read, use or interpret that bitstream may become increasingly difficult as systems evolve, adapt and eventually become redundant, so presenting a fog through which the bitstream become unusable.

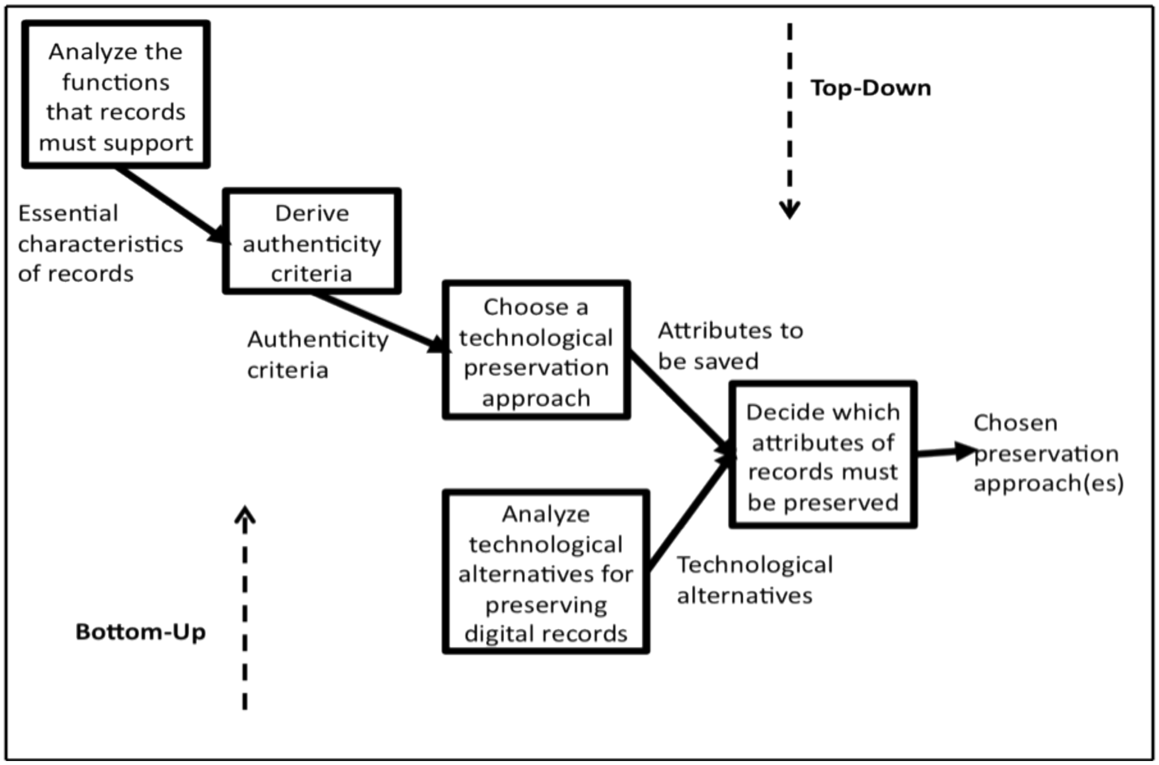
Analyze the functions that records must support
Essential characteristics of records
Derive authenticity criteria
Authenticity criteria
Choose a technological preservation approach
Attributes to be saved
Decide which attributes of records must be preserved
Chosen preservation approach(es)
Analyze technological alternatives for preserving digital records
Technological alternatives
Top-Down
Bottom-Up

Figure 1: The preservation strategy

Approaches to Preserving Layers of Meaning

- Make information useful
- Policies & procedures (periodically revisited & audited)
- Promoting awareness of issues among creators, managers and users of digital resources
- System development
- System administration
- Ongoing maintenance - copying, converting, reformatting, emulating, normalizing, migrating.

Measures for Promoting Interoperability

Abstraction and Virtualization

- Virtual machines
- Application programming interfaces
- Middleware

Open Standards

- Supporting interoperability over time by adopting conventions that are widely adopted by others, formally approved, maintained and documented (e.g. ISO standards).
- Openness is a matter of degree – few specifications are completely closed or completely open.

Representation Information & Format Registries

- Concept of representation information: "The information that maps a Data Object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e., a Data Object) is mapped into a symbol." (OAIS)
- How representation information is often expressed through file formats
- Role of registries of representation/format information.

Spectrum of Technical Digital Preservation Strategies Figure 2

Traditional Dichotomy: Emulation vs. Transformation/Migration

- Emulation – Use of software to imitate obsolete computer equipment on new computer equipment, i.e. trick files and applications into thinking they're still running in their original environment.
- Transformation/Migration – Digital object that depends on obsolete computer equipment is changed in order to run directly on new equipment.
- Advocates of emulation contend that it better supports notion of preserving an "original," along with its "look and feel," and it can be more cost-effective than repeated transformations of digital objects.

Emulation

Emulation - "To reproduce the action of or behave like (a different type of computer) with the aid of hardware or software designed to affect this; to run (a program, etc., written for another type of computer) by this means."

Open Issues Related to Emulation

- What level to emulate
- When to create the emulator - now vs. later, once vs. periodically
- How to develop emulators - what language, what platform
- Intellectual property rights

Migration

- Periodic transformation of the bits/bytes to run directly on newer platforms.
- Used widely as an approach to actively managing legacy systems.
- Work can be expensive and introduce errors of translation.

- Since the resulting objects can run directly on newer platforms, layers of technology can be minimized.

Not Just "Emulation vs. Migration"

- All strategies use standards in some way
- General consensus to keep original bits
- Transformation can be minor or extensive
- Transformation/Emulation can take place now or at time of access

Significant Properties

- "Whoever takes the decision that a particular digital object should be preserved will have to decide what properties are to be regarded as significant.
- "Properties of digital objects that affect their quality, usability, rendering, and behavior".
- Defining Significant Properties can Serve a Variety of Purposes
    - Writing specific provisions into submission agreements.
    - Developing criteria & empirical tools for evaluating preservation approaches.
    - Documentation of preservation decisions in terms of specific properties.
        - Allowing archivists to revisit previous decisions.
        - Indicating to researchers what properties have not been retained?
- Examples: font, line breaks, precision of numeric values, color space.

Passive Preservation

Passive preservation is concerned with the secure storage of digital objects, and the prevention of accidental or unauthorized damage or loss. As such, passive preservation needs to encompass the following functions: {Brown, A. 2006}

- Security and access control
- Integrity
- Storage management
- Content management
- Disaster recovery

Active preservation

Active preservation generates new technical manifestations of objects through processes such as format migration or emulation, to ensure their continued accessibility within changing technological environments. It comprises three basic functions, operating in a continuous cycle, potentially supported by the services of a technical registry. {Brown, A. 2006}

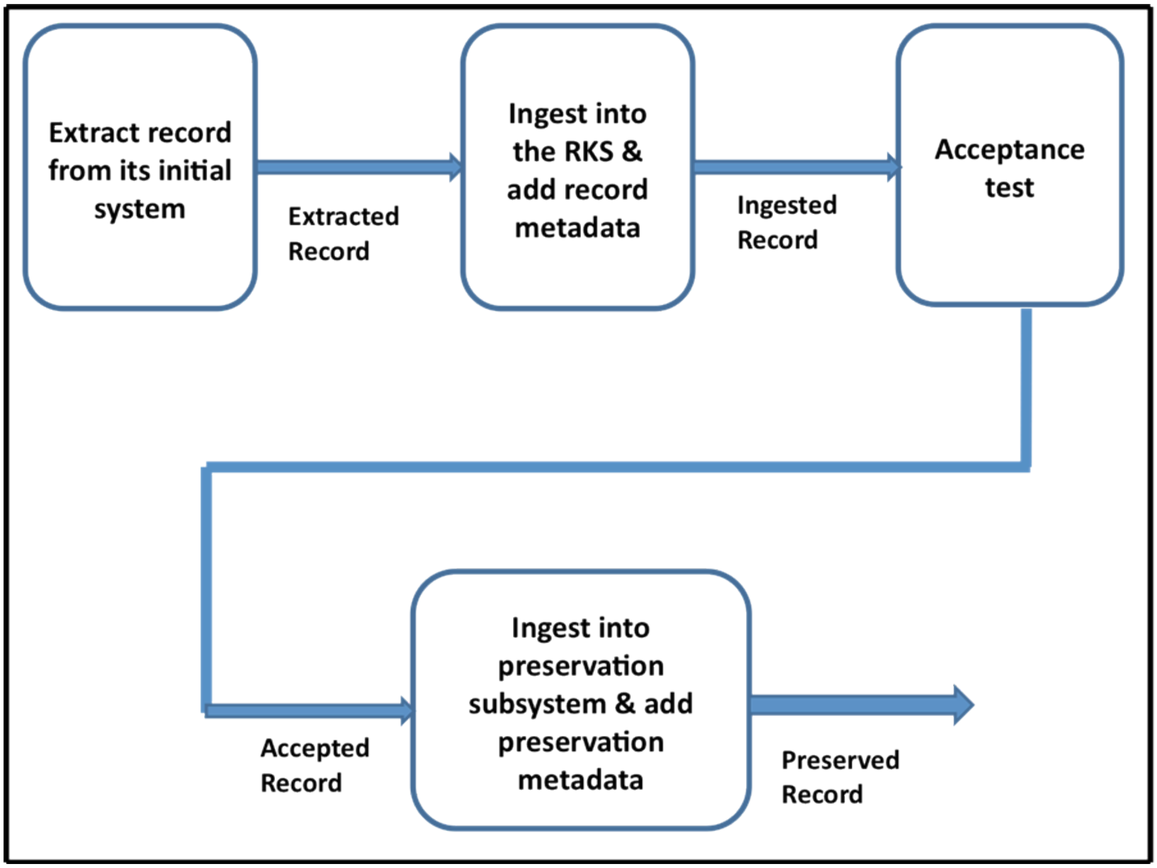
Extract record from its initial system
Extracted Record
Ingest into the RKS & add record metadata
Ingested Record
Acceptance test
Accepted Record
Ingest into preservation subsystem & add preservation metadata
Preserved Record

Figure 2: Preservation Process

4. Digital Preservation Policy

At the present time, the policies for ensuring long-term storage, maintenance, migration and access to digital materials, whether at the local or national level, are not frequently present both in the private and in the public sectors. Moreover, the policies publicly available via web are mainly developed by cultural heritage institutions and have been elaborated very recently. Frequently, the confusion about the most appropriate practices and methods, the lack of a consensus, the difficulty in engaging the interest for these themes and the shortage of good models for digital preservation can be some of the difficulties that institutions meet in developing their policies, even if the need for defining policies is increasing at the same degree of the growth of the digital heritage. The primary aims of a policy are to provide guidance and authorization on the preservation of digital materials and to ensure the authenticity, reliability and long-term accessibility of them. Moreover, a policy should explain how digital preservation can serve major needs of an institution and state some principles and rules on specific aspects which then laid the basis of implementation. {Erpa guidance 2003}

General principles

Some general principles should be followed for qualifying this activity:

- A policy needs to convey the very philosophy of an organization concerning digital preservation; it should induce a common understanding of the objectives, of whether each collection item should be preserved with maximum effort possibly applying multiple preservation paths, or whether a certain pragmatism should be pursued;
- A digital policy should facilitate the sustainability of an institution's present and future digital holdings;
- A digital preservation policy has to demonstrate its benefits, its effectiveness;
- A digital policy should be connected and integrated with a risk assessment document;
- Every policy should be practicable, not definitive, capable of being put into practice by institutions with varying resources and needs, and, especially, flexible to adapt itself to changing administrative and technological circumstances;
- Any policy should be characterized by clarity, adequacy, transparency, efficiently, effectiveness and logical organization of contents;
- A digital preservation policy should be written in a simple and suitable language, without redundancies and, at the same time, without lowering the level of quality contained in its contents;
- once a digital preservation policy is operative, it should be re-though, reviewed or newly conceived on a regular basis to take into account changes in the organizational, legal and technical environment and to make rules and guidelines more precise and explicit where there is any ambiguity about implementation;
- A digital policy should offer achievable solutions, provide for the management training and, finally, be maintained through time.

Benefits

- To develop a digital preservation strategy
- To plan coherent digital preservation programs
- To ensure and reinforce accountability
- To demonstrate that such funds can and will be used responsibly and consistently
- To ensure digital materials available for current and future use
- To define the significant properties that need to be preserved for particular classes resources

- To assist agencies in designing digitization programs
- To provide a comprehensive statement on the digital preservation
- To provide security measures that ensure the protection of digital materials during use

Requirements

- Legal requirements
- Financial requirements
- Business requirements
- Technical requirements
  - Maintenance procedures
  - Preservation strategies
  - Technology forecasting
- Historical value

5. Cost-Benefit Analysis of Preservation Approaches

- Users derive value from digital objects by performing various high-level functions.
  - Properties that facilitate those functions have instrumental value.
- Cost = sum of all resources one must commit in order to carry it out.
- Benefits = value one can derive from the digital objects that have been preserved based on that approach.
- Opportunity costs = failure to derive benefits that one could have had by choosing a different approach.
- Make best guesses about benefits based on: significant properties + information about file formats and available technological strategies.
- Impossible to directly measure now the value of future use, so we must guess as to their expected value.
- All decisions should be well-documented & revisited periodically.

Cost

- Technical infrastructure
  - Equipment purchases, maintenance and upgrades
  - Software/hardware obsolescence monitoring/review
  - Network connectivity
- Financial plan
  - Strategy and methods
  - Commitment to long-term funding
- Staffing infrastructure
  - Hiring training
  - Ongoing training
- Outsourcing

6. Digital Preservation Challenges

Digital preservation requires the management of object over time, using techniques that may result in frequent and profound changes to the technical representation of that record. Any preservation strategy must therefore be underpinned be rigorous logical framework which supports the concept of multiple technical representation of an object. And the process of change through which they arise. {Brown, A. 2006}

Origins

- Documentation of activities allows us to know about those activities without having to be there.
- Historically, contributions to this process have included oral communication, physically fixed artifacts and now digital systems.

The Hermeneutic Gap

- All conveying of stored information within a new context runs into a hermeneutic gap.
- Context is never captured or perpetuated completely.
- We use current understanding & place in the world to fill in gaps of previous contexts in order to make sense of memories.
- This is one of our greatest strengths as humans but also raises many issues related to concepts we cherish (e.g. truth, tradition, accuracy, accountability).

Professional Activities to Partially Bridge the Gap

- Archivists & other record keepers work to bridge the gap through:
- Adding metadata into the system (filing cabinets, policies) & at point of creation (naming, filing, genre conventions).
- Retention scheduling
- Disposition actions
- Transfer of custody to trusted third parties
- Labor-intensive arrangement & description
- Controlled custody environments

Resources are Limited, Meaning is Expensive

- Always true, but increasingly important in a digital environment.
- Two often competing demands:
  - More heterogeneous access (any type of client can access any type of object).
  - More functionality (each object becomes increasingly complex, thus carrying more dependencies).

Technology Obsolescence

New Conception of "Long-Term"

"A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future."

Copyright

Copyright applies to work that recorded in some way. Rights exist for musical and dramatic work as well as films, sound recordings and literary, artistic or typographic arrangements. It gives the author/creator specific rights in relation to the work, prohibits the unauthorized actions (mainly copying or broadcasting), and allows the author to take legal action against such infringements.

Copy right issues in digital preservations

- Under what circumstances dose the preserving organization have the right or permission to ingest the content into the preservation system or storage environment? This activity may be deemed illegal copying under copyright laws unless the permissions to store and make available have been clearly agreed, with associated written evidence.
- Especially for moving image and sound recordings there may be many creators and copyright holders.
- There may be restrictions on the separation of elements of a work or their independent use, for example, removing a sound track from a visual track.
- Under what circumstances may the content being preserved be made accessible as this may be defined as publication, performance, or broadcasting? The digital domain creates this new problem because of the naturally one-to-many relationship of digital content and networked access.
- Metadata to record and track copyright will be required to enable digital preservation and eventual use. Sufficient metadata fields and associated records must be available to record all the rights holders and their relationships with each other and the wider collection.

10. Resources

a. Required reading for Faculty

- Brown, A. (2006). Archiving websites : a practical guide for information management professionals (Vol. Chapter 6, pages 82-126). London: Facet Pub. http://www.amazon.com/Archiving-Websites-Information-Management-Professionals/dp/1856045536/ref=sr_1_10?ie=UTF8&s=books&qid=1255127311

&sr=8-10.

- Campbell, D. (2007). Identifying the identifiers. [S.I.]; Singapore; Dublin Core Metadata Initiative; National Library Board; c2007. http://www.dcmipubs.org/ojs/index.php/pubs/article/viewFile/34/16.
- Deegan, M., & Tanner, S. (2006). Digital preservation. London: Facet. (chapter 1&2), http://www.amazon.com/Digital-Preservation-Futures-MarilynDeegan/dp/1856044858/ref=sr_1_1?ie=UTF8&s=books&qid=1255127456&sr=1-1.
- Hedstrom, M., & Lee, C. (2002). Significant properties of digital objects: definitions, applications, implications. http://www.ils.unc.edu/callee/sigprops_dlm2002.pdf.
- Rothenberg, J. (1998). Ensuring the Longevity of Digital Information. International journal of legal information : IJLI : the official publication of the International Association of Law Libraries., 26(1), 1. http://www.clir.org/pubs/archives/ensuring.pdf.
- Erpa guidance, Digital Preservation Policy Tool, ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK en , Information Technology Society, September 2003. http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf

b. Required reading for Students

- Brown, A. (2006). Archiving websites : a practical guide for information management professionals (Vol. Chapter 6, pages 82-126). London: Facet Pub.http://www.amazon.com/Archiving-Websites-Information-Management Professionals/dp/1856045536/ref=sr_1_10?ie=UTF8&s=books&qid=1255127311&sr=8-10
- Deegan, M., & Tanner, S. (2006). Digital preservation. London: Facet. (chapter 1&2), http://www.amazon.com/Digital-Preservation-Futures-MarilynDeegan/dp/1856044858/ref=sr_1_1?ie=UTF8&s=books&qid=1255127456&sr=1-1
- Hedstrom, M., & Lee, C. (2002). Significant properties of digital objects: definitions, applications, implications. http://www.ils.unc.edu/callee/sigprops_dlm2002.pdf.
- Erpa guidance, Digital Preservation Policy Tool, ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK (**e**$^{n)}$, Information Technology Society, September 2003. http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf

c. Additional support reading for faculty and students

- Abrams, S. L. (2004). The role of format in digital preservation. VINE.(135), 49.
- Brodie, M. L., & Stonebraker, M. (1995). Migrating legacy systems : gateways, interfaces & the incremental approach. San Francisco, Calif.; [S.l.]: Morgan Kaufmann Publishers ; IT/Information Technology.

- Dollar, C. M. (2000). Authentic electronic records : strategies for long-term access. Chicago: Cohasset Associates, Inc.
- Rothenberg, J., & Bikson, T. (1999). Digital preservation : carrying authentic, understandable and usable digital records through time. Den Haag: Programma Digitale Duurzaamheid.
- van der Hoeven, J., van Diessen, R., & van der Meer, K. (2005). Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects. Journal of Information Science, 31(3), 196-208.

11. Exercises / Learning activities

1. In class, break students into groups of 3-4.
    - Ask them to discuss the content of this module and summarize the key points and prepare a power point slides, a Cmap or a poster for each group.
    - Then ask each group to login in the second life and present what they prepared.
2. Each student can try to search for preserved Web sites using the Internet Archive "Way back machine" http://www.archive.org/web/web.php.

12. Evaluation of learning objective achievement

In their answers to the learning activities, students demonstrate an understanding of:

a. The basic strategies related to digital preservations.

b. The economic issues and policies involved in the digital preservation.

c. The fundamental challenges and concerns associated with long-term digital preservation.

13. Glossary

- Digital Object: An item as stored in a digital library, consisting of data, metadata, and an identifier.
- Long Term: A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future.(OAIS)
- Representation Information: The information that maps a Data Object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e., a Data Object) is mapped into a symbol.(OAIS)
- For definitions of many other specific terms, see:

Pearce-Moses, Richard. *A Glossary of Archival and Records Terminology*, Archival Fundamentals Series. II. Chicago, IL: Society of American Archivists, 2005. http://www.archivists.org/glossary/.

14. Additional useful links

- Digital Preservation Coalition - http://www.dpconline.org.
- Preserving Access to Digital Information (PADI) - http://www.nla.gov.au/padi.
- International Internet Preservation Consortium. http://www.netpreserve.org.
- Internet Archive http://www.archive.org/index.php.
- Relevant mailing lists:
    - Digital-Preservation.http://www.jiscmail.ac.uk/lists/digital-preservation.html.
    - Digipres. http://lists.ala.org/wws/info/digipres.

15. Concept map

See VTech Concept Map server under “Dlcurric” folder under “8a-Preservation” subfolder.

16. Contributors

- **Developers:**
    - Dr. Edward Fox
    - Tarek Kanan
- **Reviewers:**
- Seungwon Yan
- John Ewers
- Venkatasubramaniam Ganesan
- Nagarajan Kuppuswami
- Ashwin Khandeparker
- Ashwin P