Digital Library Curriculum Development
Module 2c-8d: File Formats, Transformation, and Migration

(Last Updated, 10/09/09)

1. Module name: File Formats, Transformation, and Migration

2. Scope

This module covers the general principles and application of the transformation and migration processes for the preservation of digital content. Key issues surrounding preservation strategies are highlighted.

3. Learning objectives

By the end of this module, the student will be able to:

a. Explain the standard process of migration projects, from recognizing the need for migration, initiating the preservation effort, selecting appropriate media, and maintaining the long-term usage of a collection of digitized materials.

b. Demonstrate an understanding of the critical issues and challenges of a preservation project.

c. Practice the implementation of a migration effort.

4. 5S Characteristics of the module

- Streams: Documents are represented in bitstreams.
- Structures: Documents and metadata are migrated for the preservation of collections. Metadata uses structures to describe digital objects, such as documents, which are streams with some structure(s) imposed.
- Spaces: Copies and replicas of digital objects are kept in different locations. Migration occurs across time and space. Preservation allows for interoperability over time.
- Scenarios: Migration strategy decisions are considered in response to a scenario change. In this context, migration itself is a scenario.
- Societies: Communities dictate what formats are widely-available and what must be preserved. They also make decisions and implement policies related to preservation.

5. Level of effort required

a. Class time: 1 hour

b. Student time outside class: 2.5 hours

- Reading before the class starts: 2 hours
- Homework assignment: 0.5 hours

6. Relationships with other modules

- 2-a: Text Resources, 2-b: Multimedia

  2-a and 2-b cover the nature, structure, composing factors, and formats of various types of digital objects (e.g., text, images, video, etc.).
- 3-b: Digitization

  3-b covers the process of digitization regardless of the object types, and discusses digital file formats.
- 4-b: Metadata, cataloging, metadata mark-up, metadata harvesting

  4-b covers uses of metadata and metadata standards related to the context of digital libraries in general.
- 8-a: Preservation

  8-a covers the related technology, standards, and policies concerning the preservation of digital objects.

7. **Prerequisite knowledge required**: None

8. **Introductory remedial instruction**: None

9. Body of knowledge

1. Definitions of key terms
   a. File Formats
      i. A file format refers to the layout of the data inside of the file and the organization of that data in terms of bits, since digital data can only be stored using a binary system in terms of 0s and 1s.
      ii. Packages of information can be stored as data files or transmitted as data streams (a.k.a. bitstreams or byte streams)
      iii. A format is a fixed, byte-serialized encoding of an information model
   b. Filename Extensions
      i. Extensions are suffixes to the filename, which give an indication as to the format of the content of the file.

   ii. Filename extensions have historically been a 3 character suffix. However, modern operating systems don't have such limitations anymore.

   iii. Reference: http://www.file-extension.com

c. Bitstream Copying

   i. Copying a stream of data into a duplicate stream.

   ii. It is commonly known as "backing up" your data.

   iii. Bitstream copying refers to the process of making an exact duplicate of a digital object.

d. Transformation

   i. Transformation is the process of altering the format of an object (destination format could be a digital file or output display).

   ii. Transformation means that a file is converted from one file format (e.g., .avi) to another file format (e.g., .mov).

e. Migration

   i. "*Transfer of a data object, as from one format to another, or from one medium to another, or between instances of the same type of storage medium" (Rosenthal et al., 2005)*

   ii. Migration refers to the movement of data from one media technology to another.

   iii. Switching from storing digital data in CDs to storing it in hard disks means migration from the CD to hard disks as media of storage.

f. Refreshing – copying content to new media periodically

   i. Refreshing is to copy digital information from one long-term storage medium to another of the same type, with no change whatsoever in the bitstream (e.g., from an older CD-RW to a new CD-RW). New media is of the same type as the old media. You copy data from an old CD to a new CD.

   ii. The goal of refreshing is to preserve the data from any bad effects that can be caused by damage to the media that hold the data.

   iii. Refreshing is key to the preservation of data.

g. Modified Refreshing

   i. Modified refreshing is the copying to another medium of a similar enough type that no change is made in the bit-pattern that is of concern to the application and operating system using the data.

   ii. For instance, you may copy from a 100 MB Zip disk to a 750 MB Zip disk.

2. File Formats

a. To preserve documents, a collection's file format must ensure the following:
    i. Save the bits so that somewhere a copy survives and that copy can be found.
    ii. Ensure that the bits can be interpreted later (file format retention).
    iii. Make the bits trustworthy by reliably associating sufficient metadata.
    iv. Include library content lists among the set of saved documents.
    v. Minimize the need for digital archeology (rescuing content from obsolete technology) through the ability to translate to other formats.

b. Which formats can preserve content or lead to a longer duration between transformations?
    i. International standards are preferred.
    ii. XML allows users to understand an object's structure and content without using specific machines or software.
    iii. Text documents (Word or other proprietary formats) depend on compatibility between software versions. XML with metadata tagging would lead to longer preservation.
    iv. PDF documents lead to PDF/A documents (archiving).
    v. Image options – TIFF, GIF, JPEG, JP2, Flashpix, ImagePac, PNG, PDF

c. Sustainability factors for file formats
    i. Adoption – used by the primary creators, disseminators, or users of information resources
    ii. Disclosure – complete specifications and tools for validating technical integrity exist and are accessible
    iii. External Dependencies – depends on particular hardware, operating system, or software
    iv. Impact of Patents – ability of archival institutions to sustain content in a format will be inhibited by patents
    v. Quality and functionality – ability of a format to represent the significant characteristics of a given content item required by current and future users
    vi. Self-documentation – contains basic descriptive, technical, and other administrative metadata
    vii. Transparency – digital representation is open to direct analysis with basic tools
    viii. Technical Protection Mechanism – implementation of mechanisms, such as encryption, that prevent the preservation of content by a trusted repository

d. File format selection is a first step in long-term preservation measures that is part of a larger information management strategy.

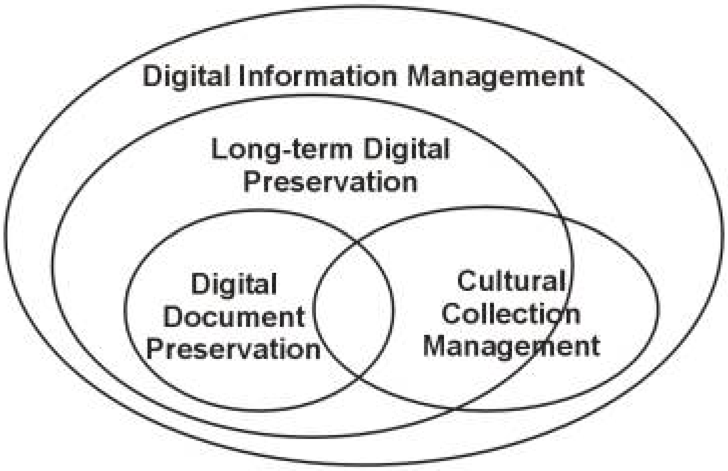

Digital Information Management
Long-term Digital Preservation
Digital Document Preservation
Cultural Collection Management

Figure 1: Relationships between digital preservation and management

Taken from: http://home.pacbell.net/hgladney/ddq_1_2.htm

e. Difficult items to store or move between formats

  i. mathematical symbols

  ii. chemical formulas

  iii. archaic scripts or ideographs, such as Egyptian or Mayan hieroglyphs

  iv. musical notations

3. Emulation

a. *"The essential idea behind emulation is to be able to access or run original data/software on a new/current platform, by running software on the new/current platform that emulates the original platform."(Granger, 2000)*

b. The digital object is kept in its original file format; the hardware and/or software needed to render that format are emulated.

c. Advantages and Disadvantages

  i. Emulation may retain the "look and feel, and interactivity" of a digital object.

  ii. Emulation avoids information loss.

  iii. Emulation requires full knowledge of the original system and context.

  iv. Emulation software needs to be upgraded to work with current systems.

d. Technical Issues

        i. How many layers of emulation?
        ii. Are there standards and open specifications that can facilitate emulation?
    e. Legal issues
        i. proprietary systems
        ii. reverse engineering

4. Transformation and Migration
    a. Migration is a fundamental digital preservation strategy.
    b. Goals
        i. Preserve content and functionality of a digital object.
        ii. Ensure continued access to the digital object.
        iii. Minimize physical and intellectual information loss.
    c. Possible Approaches / Strategies to Migration (Hedstrom)
        i. Transfer to paper or microfilm store in "software-independent" format.
        ii. Retain in the native software environment.
        iii. Migrate to a system that is compliant with open standards.
        iv. Store in more than one format.
        v. Create surrogates.
        vi. Save the software needed for access and retrieval (see Emulation section).
        vii. Develop software and hardware emulators (see Emulation section).
    d. Types of Migration
        i. To a newer version of the file format
        ii. To a different file format
            1. To a standard file format (Normalization)
        iii. To a different hardware/software environment
    e. Frequency of Migration
        i. Automated
            1. Migrations are handled by the system, without intervention from the content creators or system administrators.
        ii. On Request
            1. A content creator or administrator initiates a migration.
            2. (May also apply to when a DL stores all digital objects in a "master" file format and converts them to other formats only at the request of an end user.)
    f. Issues with Migration
        i. Conflicting goals for a particular migration (e.g., better access, better preservation)
        ii. Is there value in holding multiple versions of a work?
        iii. Tradeoffs between information loss and reduced complexity and cost of operation

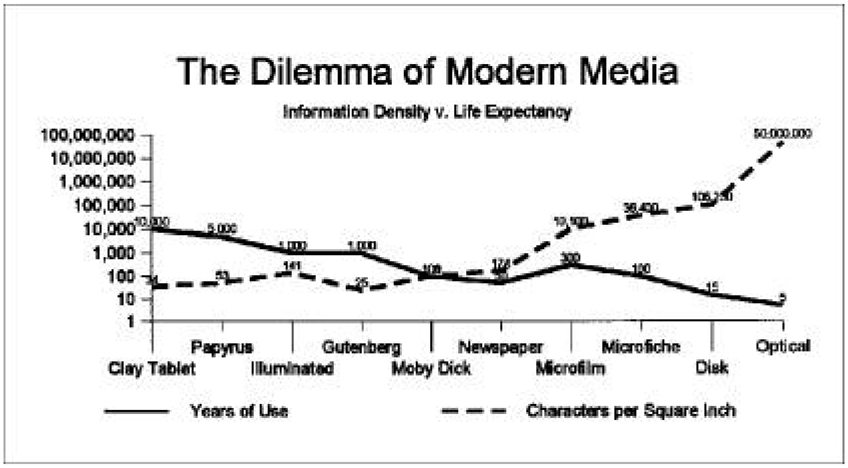


Figure 2: Media Comparison
http://clir.org/pubs/reports/conway2/index.html

g. Example object lifecycle (see Figure 3)
   i. Content exists as an "information package", which contains the content object, in analog or digital form, and perhaps metadata.
   ii. Transformation occurs to digitize the object to an archival master.
   iii. An archival master digital object exists.
   iv. Migration action on a digital object changes the master object to create a derivative object.
   v. The archival master is migrated to new formats through transformations.

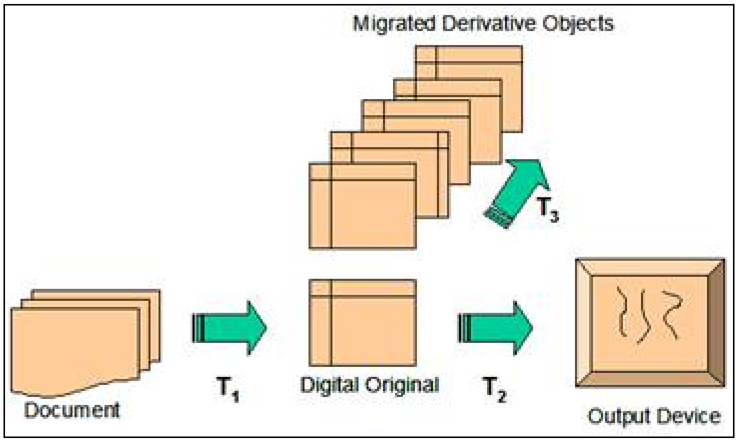


Figure 3: View of an object undergoing three transformations and multiple migrations
http://www.dlib.org/dlib/june05/jantz/06jantz.html

h. The OAIS Reference Model provides (among other things) a detailed framework for defining transformation and migration functions in a digital library system.

5. Emulation versus migration

i. Selection of a file type depends on whether the system used to manage the content will be emulated, or if file type will be converted to something else at a later time

ii. Trade-off between the speed of processing, cost of the emulating or migration software, and the accuracy and quality of content

iii. Vision

1. automatically handle content from creation to preservation
2. a self managed system that takes care of emulation and migration
3. abstract the preservation process

iv. Universal Virtual Computer

1. middle of the emulation vs. migration spectrum
2. *"It uses elements of both migration and emulation which allows digital objects to be reconstituted in their original form. The UVC concept consists of the UVC itself, a logical data scheme with type description, the UVC program (format decoder) and the logical data viewer." (PADI, Universal Virtual Computer Papers.)*

6. Preservation Issues

a. Preservation requires

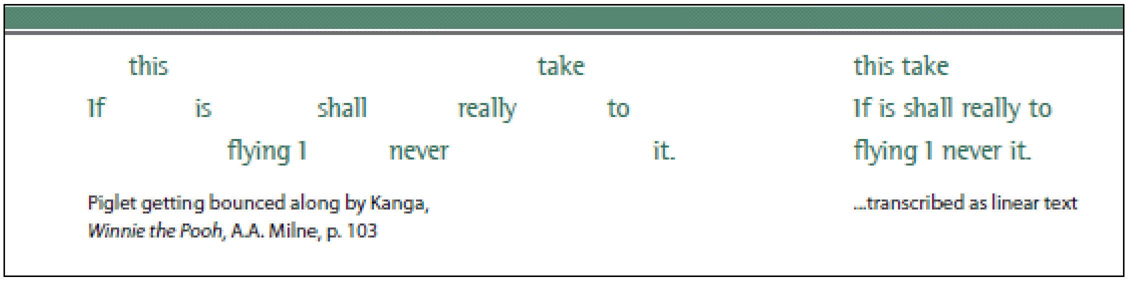
i. Protection of an original item

ii. Preserve the technology used to digitize the material, and the software needed to retrieve and render these digital objects. This is an overhead.

iii. Digital archives must support the new file formats.

iv. Maintenance of digital objects from digit corruption or destruction. Who will take the responsibility for long-term preservation?

v. Best practices to preserve formatting so that the context is not lost to future generations



this take
If is shall really to
flying I never it.

Piglet getting bounced along by Kanga,
Winnie the Pooh, A.A. Milne, p. 103

this take
If is shall really to
flying I never it.

...transcribed as linear text

vi. Will the digital archives be accessible perpetually in future?

vii. Quality of digitization – will it stand the test of time?

viii. Digital Objects would be a backup or the original materials will be the ones to be preserved?

ix. Single or Multiple repositories?

x. How much to preserve? Which material gets top priority to be digitized and preserved?

xi. Is it legally permissible for a library to rescan originals to replace unusable and corrupted digital objects?

xii. What are the copyright implications of transforming a digital object from TIFF to JPEG?

b. The ICA *Guide to Managing Electronic Records* sets out seven criteria for selecting media used for preserving digital records:

i. Open standards for digital recording on the medium
ii. Robust methods for preventing, detecting, and reporting errors
iii. Sufficient market penetration
iv. Known longevity
v. Known susceptibility to degradation or deterioration
vi. A favorable cost/benefit ratio
vii. Availability of methods for recovering from loss

c. Five broad categories of prevention strategies:

i. preserving the original technology used to create or store the records
ii. emulating the original technology on new platforms
iii. migrating the software necessary to retrieve, deliver, and use the records
iv. migrating records to up-to-date formats
v. converting records to standard forms

d. Preservation strategy includes

i. defining minimum digital preservation requirements necessary to ensure the persistence of digital materials and associated metadata files to facilitate shared storage and registry initiatives
ii. working with IT groups within cultural institutions (such as theory centers, central IT units, academic technologies, computer science departments) to develop and manage shared large-scale storage systems
iii. making data-redundancy arrangements among libraries for backup, or implementing other distributed and collaborative strategies such as LOCKSS
iv. developing storage metrics to share configuration and cost information in standardized ways
v. supporting standards for storage-management interoperability
vi. sharing open-source preservation applications and collaborating to develop access and preservation services as flexible and scalable

components to be added to repository models supporting preservation activities

vii. exploring usage trends created by the online availability of materials to assess how the 80/20 rule applies in the digital world and to consider how usage statistics can inform preservation decisions in support of priority setting and risk taking

viii. exploring how to incorporate risk assessment strategies in making and implementing preservation decisions, being sure to consider how preserving the analog books might affect the risk assessment strategies for the digital versions, and vice versa.

ix. creating a wiki (or a similar collaboration tool) to systematically distribute up-to-date information about preservation strategies implemented by different libraries

x. offering consultancies, workshops, and training sessions

10. Resources

a. Required readings for students

Arms, Caroline R. and Carl Fleischhauer. *Sustainability of Digital Formats: Planning for Library of Congress Collections*. May 21, 2007. http://www.digitalpreservation.gov/formats/index.shtml

Rieger, O.Y. *Preservation in the Age of Large-Scale Digitization*. Library of Congress white paper. CLIR pub 141, 52 pp. February 2008.

Van Wijk, Caroline. *Starting Point for Migration Research*. Migration Research Project, Koninkliije Bibliotheek. July 2006. http://www.kb.nl/hrd/dd/dd_projecten/Starting_Point_Migration_Research.pdf

b. Required readings for instructors

CCSDS. *Reference Model for an Open Archival Information System* (OAIS), Blue Book. January 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf

Rothenberg, Jeff. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. A Report to the Council on Library and Information Services. January 1999. http://www.clir.org/pubs/reports/rothenberg/contents.html

c. Recommended readings for students

File formats

- Chowdhury, G.G., & Chowdhury, S. (2003). *Introduction to Digital Libraries*. Chapter 6, Digitization (pp. 103-119). London: Facet Publishing.
- Gladney, H.M. (2002). *Perspectives on Trustworthy Information*. In Digital Document Quarterly V1: No 2. <http://home.pacbell.net/hgladney/ddq_1_2.htm>

- Library of Congress. *Sustainability of Digital Formats Planning for Library of Congress Collections*. May 21, 2007. <http://www.digitalpreservation.gov/formats/>

Preservation

- Beagrie, N. *National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and of Related International Activity*. Council on Library and Information Resources and the Library of Congress. April 2003.
- Conway, P. *Preservation in the Digital World*. March 2006. <http://clir.org/pubs/reports/conway2/index.html>
- ICPSR, University of Michigan. *Digital Preservation Strategies Tutorial*. 2003. <http://www.icpsr.umich.edu/dpm/dpm-eng/terminology/strategies.html>
- Council on Library and Information Resources. *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. National Digital Information Infrastructure and Preservation Program, Library of Congress. 99 pp. April 2002. <http://www.clir.org/pubs/reports/pub106/pub106.pdf >
- Granger, Stewart. *Emulation as a Digital Preservation Strategy*. D-Lib Magazine, V. 6 No. 10, October 2000. <http://www.dlib.org/dlib/october00/granger/10granger.html>
- Hedstrom M. *Research Issues in Migration and Long-Term Preservation.* Archives and Museum Informatics, Volume 11, Numbers 3-4, 1997, pp. 287-292(6).
- PADI. *Universal Virtual Computer Papers*. National Library of Australia. <http://www.nla.gov.au/padi/topics/492.html>
- Rosenthal, David S.H., et al. *Requirements for Digital Preservation Systems: A Bottom-Up Approach*. D-Lib Magazine, V. 11 No. 11, November 2005. <http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html>
- Thibodeau, K. *Preservation and Migration of Electronic Records: The State of the Issue*. The U.S. National Archives and Records Administration. <http://www.archives.gov/era/papers/preservation.html>
- Wijngaarden, H. and E. van en Oltmans. *Digital Preservation and Permanent Access: the UVC for Images*. 2004. <http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/uvc-ist.pdf>

Domestic and International Project Examples

- Content transformation at HP Labs. <http://www.hpl.hp.com/research/content.html>
- Council on Library Resources – long term preservation <http://palimpsest.stanford.edu/byorg/abbey/an/an08/an08-5/an08-504.html>
- HATII Planets Project. <http://www.hatii.arts.gla.ac.uk/research/planets.html>

11. Concept map (created by students)

12. Exercises / Learning activities

Homework assignment:

Multimedia, conversion software, and people's goals react in certain ways depending on the need for long-term preservation. In small groups of 2-3, discuss the following issues with respect to three settings where the priorities of preservation vary from important to possibly not necessary. The three settings are the Library of Congress's content on the early US governments, a university department's most notable publications, and your personal multimedia content.

Compression:

- Under what circumstances is lossy compression an acceptable migration strategy?
- Under what circumstances must content be kept in a raw format when migration has been chosen over emulation?
- Does the reduction of space justify compression, un-compression, and recompression of data when content will have to be migrated to a new format at a later time?

Image & Video:

- 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 small degrees of degradation acceptable?
- How should the continued development and wide-spread acceptance of new image formats be managed? Is this different for emulation and migration strategies?

Annotation, Audio:

- Is it necessary to retain voice annotations in audio files in their original format or is a computer-generated transcript of the voice annotation an acceptable alternative?

Preservation:

- How do we know we have kept enough metadata for digital materials to allow their migration for future purposes? For example, is information about an image's scanner and its own physical properties enough?

13. Evaluation of learning achievement

- Students may perform the assignment described in section 12 individually or in small groups, though small groups are recommended.
- The homework assignment in section 12 assumes that the course in which this module is taught requires students to build, design, or evaluate a digital library system.

- In this exercise, students should be evaluated based on the comprehensiveness of their description of the issues related to preservation of digital data with regards to the three settings described above.
- After the exercise, a group representative should summarize the discussion that went on among the group members to address the issues within each context.

14. Glossary

a. Bitstream Copying (Section 9.1.c)(Please change next ones similarly. Also add: OAIS, QVC)

b. Digital Archeology (Section 2)

c. File Format (Section 2)

d. Filename Extensions (Section 1)

e. Migration (Section 5)

f. Modified Refreshing (Section 1)

g. Refreshing (Section 1)

h. Transformation (Section 5)

15. Additional useful links

a. SunSITE. *Preservation Resources*. 2007. <sunsite.berkeley.edu/Preservation/>

b. The Library of Congress Preservation. <http://www.loc.gov/preserv/>

16. Contributors

a. Initial authors:

- Jonathan Leidig
- AJ Alon
- Amine Chigani
- Mahima Gopalakrishnan
- Sung Hee Park

b. Editor/Reviewer:

- Edward A. Fox