eArchiving: the Digital Black Hole

Marilyn Deegan
Digital Resources Director
Refugee Studies Centre
University of Oxford

Two key issues in eArchiving

- Digital preservation
  - which means ensuring full access and continued usability of data
- Preservation through digitization
  - which allows for greater security of physical analogue materials

Digitization vs microfilm for preservation: microfilm

- Microfilm advantages
  - good microfilm is predicted to last 500 years
  - microfilm is self-explanatory
    - even without the technology, it can be understood
  - microfilm preservation is well-understood
  - microfilm is relatively cheap to store
  - it is a stable technology

Digitization vs microfilm for preservation: microfilm

- Microfilm disadvantages
  - microfilm is difficult to access
  - there is degradation between masters and copies
  - it is easily damaged in use
  - it has to be used in situ in the library as microfilm readers are not common outside libraries

Digitization vs microfilm for preservation: digitization

- Digitization advantages
  - digital files are easy to access and search
  - there is no degradation in copying
    - the first, tenth and millionth copies are all exactly the same
  - use does not damage
  - the technology is ubiquitous

Digitization vs microfilm for preservation: digitization

- Digitization disadvantages
  - the stability of the medium is questionable (tapes, CDs etc)
    - frequent refreshing needed
  - speed of hardware and software change means that data can be unreadable after a few years
  - formats change frequently
  - metadata systems are developing rapidly

Reborn and born digital materials

- Reborn digital
  - scanned from analogue materials
    - books, journals, manuscripts, photographs, etc etc
  - In most cases, we still have the originals

Reborn and born digital materials

- Born digital
  - created originally in digital form
    - may then be printed: books, journals, images, etc
  - may be unprintable because of the complexity
    - multimedia
    - reference works based on databases
    - web sites
  - the digital data is the original

Preservation and the digital black hole

- The scale of the problem
  - much vitally important data is now created digitally
  - some of it is unprintable
  - this data is part of the cultural memory of the late twentieth and early twenty-first centuries
  - some has undoubtedly already been lost
  - much more is in danger
- There is potentially a black hole in the record of our culture
- Action is urgently needed

Archiving the web

- The average life of a piece of information on the web is 44 days
- Web sites change almost daily
  - some change every few minutes
    - news sites
    - share prices
    - dynamic data
  - much web data is ‘hidden’
  - the ‘deep web’

Digital preservation initiatives

- The Library of Congress
  - in November 2000, had to act very quickly to archive web sites relating to Clinton as the were being dismantled very fast
  - is now planning to spend $100 million dollars on preserving federal digital information
- Europe
  - NEDLIB project
    - European deposit libraries who are trying to find ways of archiving digital legal deposit materials
- UK
  - British Library
    - £20 million on a digital store

Methods of digital preservation

- Technology preservation
- Refreshing
- Migration and reformatting
- Emulation
- Data archaeology
- Output to analogue media

Technology preservation

- Keeping all the hardware and software to run the data
- Means keeping a whole range of computers in operation
  - which is expensive
- A large number of operating systems and software packages would need to be supported
- Support would be very expensive

Refreshing

- Electronic media can become corrupt or be superceded
  - floppy discs, hard drives, tapes, CDs, etc
- It is necessary to move data periodically to new media
- There is no change in the configuration of the data
- Regular refreshing needs to be done even if other preservation strategies are adopted

Migration and reformatting

- As data formats change, data streams will need to be moved to new formats
- This process will change the actual configuration of the data, and some contextual information might be lost
- Data is sometimes reformatted when it is accessioned in order that it is easier to preserve it in the long term
    - text may be converted from a proprietary format to SGML or XML
    - images may be converted from a proprietary format to TIFFs

Migration and reformatting

- This is a relatively expensive process as all data has to be converted whether it is eventually needed or not
- It is a just-in-time method
- This is the method of preservation which has received most attention up to now

Emulation

- This is the process of building hardware or software which will mimic the functions of other hardware and software in order that programs will run
- An example of this is the emulators that can make Macintosh computers run Windows software
- Emulation would require
  - data to be stored in its original formats
  - software to be stored with full documentation
  - hardware to be built to emulate the original machines

Emulation

- Emulation is a just-in-case method
  - software is only emulated when the data is needed
- Costs are unknown
- Long-term implications unknown

Data archaeology

- Sometimes lost data can be recovered
- But it takes painstaking and intensive work on data archaeology
- May need to track electrical impulses with recording devices
- Not a strategy for general preservation
- But could be useful for rescue work

Output to analogue media

- Printing out documents
- Computer output to microform
  - automatic outputting to film or fiche
- For key documents, this may be a good insurance policy

Conclusions

- eArchiving is the most crucial issue in the digital libraries world
- It is of vital importance to the whole of society as it concerns our cultural memory
- It is expensive
  - and we are not really sure of the costs
- It requires international effort
  - lots of activity in USA, Australia, Europe