Future-Proofing the Web: What We Can Do Today

16 September 2005

John Kunze, California Digital Library

Conclusion

- *Dessication* now; migration/emulation maybe
- Modest preservation budgets, competing organizational priorities, and diminishing expert format knowledge may make it worth while to save original formats along with simple, low-tech, *dessicated* formats having an excellent preservation outlook just in case
  - the original format should fail and
  - we never get funding to touch the objects again

California Digital Library (CDL)

- A university library with no books, students, or faculty
- Central services for the 10 campus libraries of the University of California
- Content hosting: electronic texts, datasets, finding aids, etc.
- New preservation challenge for CDL: capture and long-term retention of material found on the web

What’s digital preservation?

- The activity of storing objects that remain usable and faithful to the creators’ original intention
- How? By safeguarding information’s …
  - *Viability* (intact bit streams)
  - *Renderability* (by machines)
  - *Understandability* (by humans)
- *Viability* not in scope here

Migration and Emulation

- Migration problems
  - Unknown costs, human review, format errors
- Emulation problems
  - Unknown costs, human review, software IP
- Both try to keep up with or preserve an object’s technical context
- An approach to reduce that context…

The Lesson from Paper

- As a recording and display device
  - Can last for 1000 years
- Why this astonishing performance?
  - No technical intermediation required
- What trick can we borrow?
  - The simplest technologies to maintain and understand today are the simplest to carry forward and to recreate in the future

Low-Tech Dependencies

- Semantic technology
  - Loss inevitable due to linguistic shifts
- Substitute light source
  - Fire, the lowest tech invention
- Microfilm
  - Light source plus lens, 500-year old technology

Dessicated Data

- Remarkable lesson from the longest-lived online digital format
  - Plain text archives of IETF internet RFCs
  - High in value, low in features
- Preservation through “dessication”
  - No fonts, graphics, colors, diacritics, etc.
  - But essential cultural value retained

Hedging our Bets

- Always save the original format
- In addition, derive dessicated formats in case the original format ever fails
- Extra storage cost may be incurred anyway if your access system requires a plain text derivative for search indexing
- Question: what about Latin-1 support
- Question: surfacing hidden features

Next Lowest-Tech Technology

- Raster image as alternate dessicated format
  - Rectangular grid of picture elements
  - Technical impact of pressure to compress
    - Open run-length encoding or wavelet?
- Rendering tools will never be better than at peak of format's popularity
  - Very common malformed format instances
- Additional fall back format in case the original and plain text versions fail
- Question: surfacing hidden data

Beyond Text and Images

- No attempt yet to formulate dessicated data versions of audio, video, or multi-media
- General lesson: technology will clearly be part of digital preservation, but the greater the technological dependence, the greater the risk

Summary

- Save the original *and* dessicated versions as fall backs in case of failure
  - Few features, much value
  - Low cost, done at peak of tool sophistication
- Web archiving has no preservation metadata
- We may never have the money to touch most of our objects again
- Dessication is something we can do today