

Why does it cost so much?

Decisions and choices in preservation of digital content

New England Archivists Fall 2008 Meeting
Boston, Massachusetts

Amy Friedlander, Ph.D
Council on Library and Information Resources
November 15, 2008

Council on Library and Information Resources: Introduction

- Not-for-profit organization that undertakes activities at the intersection of higher education, advanced research, and libraries
- Interests in preservation, digital archiving, and scholarship and the infrastructure, including libraries, that supports and fosters research and education.
- Sponsorship from the Mellon Foundation and individual academic and research libraries and organizations.

This talk will:

- Describe some of what has been learned during the work of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF-SPDA)
- Outline several of CLIR's projects that complement and extend aspects of the work of the Task Force
- **Not** represent the work or consensus views of the Task Force
- Acknowledgements: Charles Henry, Amy Lucko, Fran Berman, Sayeed Choudhury, Clifford Lynch, Brian Lavoie, Lorraine Eakin

Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF-SDPA)

- Two-year effort engaging 19 experts from economics, computer sciences, library and information science
- Addresses the “data deluge” in science as well as more generally (Gantz 2008)
- Support from NSF, Library of Congress, the Mellon Foundation, CLIR, JISC and others
- Deliverables:
 - Year 1 report that establishes the conceptual framework
 - Year 2 (final) report that describes the model(s)



BRTF-SDPA: Objectives

- General **cost framework**: key cost categories of digital preservation
- Set of **economic models** which provide alternative ways of addressing sustainable digital preservation
 - Pros, cons, costs, trade-offs of each
 - List real world conditions for which each model is best suited.
- **Actionable recommendations**: “If your digital preservation context is X, you should consider using model Y for sustainable digital access and preservation.”

Source: F. Berman, “Research and Data,” ARL/CNI Workshop, October 2008. Used with permission.

Digital preservation is visible.

- Preservation was once a dimension of technical services in libraries.
- Digital preservation requires active management.
- Preservation v. Curation: Total process of management
 - Acquisition
 - Management of the content
 - Re-purposing and re-use of the material

What are “costs”?

- Acquisition v. total cost of ownership
- Operations
- Maintenance
- Infrastructure, threshold investment
- Value proposition
 - We can quantify some of the costs.
 - We have trouble quantifying the value of the collections *and services*.
 - So the cost-benefit analysis is undefined.
 - Costly relative to what?

Components of Costs in a Nutshell

- Labor, especially metadata creation
- Format
- Scale and heterogeneity
- Granularity
 - Collection or item
 - Resolution
 - Tagging/mark-up
- Environmental factors
 - Heating and cooling
 - Power consumption to operate
 - Regulatory framework
- Time



Prior studies

- What are the assumptions?
- What are they measuring or estimating?
- How does the study map to your context?

Note: This is based on the excellent work by Lorraine Eakin; background paper to be posted to the BRTF-SPDA website in December 2008.

Roquade Project / Dekker et al. (2001): Published literature

- Personnel costs of assigning metadata: approximately 10 euros
- Processing SIP's: approximately 10 euros per information item
- 5,000 items per year added: 6 PC's with a network card and AV facilities: 1500 euros each + professional server: \$5000 euros
- Total hardware costs: approximately 32,000 euros, depreciated over 4 years
- Software and licensing fees: 15,000 euros per year using proprietary software
- Maintenance support costs: 2,000 euros per year
- Technical support: 0.2 FEs = 9,000 euros per year
- Data refresh every 5 years @ 1 euro per MB; if DIPs are kept for 20 years and DIP is about 500 kB, cost - about 2 euros per information item, that is, 10,000 euros per year for all information items
- **Total per information item costs: 29 euros per item**


Chapman (2003):storage

- Excludes ingest and access
- Based on billable square feet
- **\$0.08** per 332-page (microfilm) volume per year in the standard vault
- **\$0.19** per 332-page (microfilm) volume per year in the film vault
- **\$0.31** per 332-page (book) volume in the standard vault



OCLC/Chapman (2003): cost/GB

- Excludes ingest and access
- Based on GB of data deposited
- **\$0.01-0.06** per 332-page ASCII text
- **\$0.47/\$1.01/\$1.89** per 332-page 600-dpi 1-bit page image (variable rate, based upon total amount of data deposited per account)



Digital Preservation Testbed, Nationaal / Testbed Digitale Bewaring, Archief of the Netherlands (2005): e-mail per yr.

- Creation of a batch of 1000 records (assuming 50kb per email, 100 kb per text document, 250 kb per spreadsheet, and 2 Mb per database): **333 euros**
- "Repair" of a batch of 1000 records (assuming 50kb per email, 100 kb per text document, 250 kb per spreadsheet, and 2 Mb per database): **10,000 euros**
- Acquisition and input of metadata for "normal" email: **1.41 euros**
- Acquisition and input of metadata for XML email: **0.06 euros**

Riksarkivet/National Archives of Sweden / Palm (2006)

- Looked at:
 - Cost per year per 1 Gb stored;
 - Total costs per year
- 1 Hierarchical Storage Management System (i.e., HSM) (2003 price + 3% interest per year): 449,694 euros over five years
- Storage medium for additional 40 Tb/year: 43648 euros over five years
- Staff
 - Staff operations costs: 132240 euros over five years (0.6 FTE)
 - Staff ongoing data input: 88160 euros over five years (0.4 FTE)
- Total annual input cost: 131808 euros over five years (staff & storage medium included)
- Facilities ("Premises") (100 square meters): 66228 euros over five years
- Service/support: 138300 euros over five years
- Digitization of paper materials (1-bit 600 dpi files in A4 format): 0.10 euro per file, with 5 million images scanned annually
- Scanning of large-format drawings and maps (8-bit grey-scale at 297 dpi, in manually fed scanners): 0.61 euro per file, with 1,321,000 image files created annually
- Production costs for 1 Gb 1-bit digitized information: approximately **17 euros per Gb**
- Production costs for 1 Gb 8-bit digitized information: approximately **30 euros per Gb**
- Production costs for Audiovisual information: approximately **11 euros per Gb**


Academy of Motion Picture Arts and Sciences/ AMPAS Science and Technology Council (2007)


- "All film" production generating no digital assets, annual storage costs for archival master: **\$1059**
- A film-captured, digital finished production at 4K, annual storage costs for archival master: **\$12,514**
- Digitally captured, digitally finished production using HDCAM SR videotape as the capture medium at 1920 x 1080, annual storage costs for archival master: **\$1,830**
- Digital captured, digitally finished production using an uncompressed digital data capture system at 2K, annual storage costs for archival master: **\$1,955**
- Digitally captured, digital finished production using an uncompressed digital data capture system at 4K, annual storage costs for archival master: **\$12,514**

Time has several senses.

- Technology changes.
 - Migrate formats
 - Refresh data
 - Respond to changes in hardware and software
 - → Learning curves that do not always show up in “the numbers”
- Technology may help – automatic capture of metadata element.
- Preservation/curation has a life cycle.
- Perpetuity means open-ended.

LIFE²: Life Cycle Model


$$L_T = C + Aq_T + I_T + BP_T + CP_T + Ac_T$$



Lt: Life cycle
C: Creation or purchase
Aqt: Acquisition
It: Ingest
BPt: Bitstream Preservation
Cpt: Content Preservation
Act: Access



Source: The LIFE2 Final Project Report (August 22, 2008), p. 16, Figures 3 and 4.



LIFE² Estimates, total cost per year

- Several different projects, yielded ranges
- Year 1: £15.00 - £31.50
- Year 5: £16.50 - £32.00
- Year 10: £16.70 - £32.20

What might minimize costs?

- Automation
 - Metadata capture
 - Cataloging
- Time – Initial processing reduces costs
 - Standards process
 - Planning
 - Collaboration



Hidden Collections

- Generous grant from the Andrew W. Mellon Foundation to run a competition to catalog unprocessed materials held in the special collections of libraries, archives, museums, and historical societies
- ***Two known problems:***
 - *Small, distributed collections of materials of potential value to scholars either individually or in concert with others*
 - *Labor-intensive cataloging of manuscripts*
- First year with renewals for a total of five years

Hidden Collections: Eligibility

- Demonstrated value to scholars
- Web-accessible catalog with records that can be “discoverable” and hence compliant with current protocols and standards
- Long term responsibility for the maintaining the records (sustainability)
- Collections owned or held in the USA (Y1)
- Applicant a not-for-profit organization
- Digitization or format conversation not in scope

Hidden Collections: What will we learn?

- What constitutes an important collection? To whom? And how do individual collections relate to each other? → value proposition
- How do organizations build a shared infrastructure?
- How can description and cataloging be streamlined?

What increases ambiguity?

- Unknowns: risks and liabilities
- Risks
 - Random events
 - Natural disasters
- Liabilities
 - Intellectual property
 - Evolving expectations and perceptions
 - What is professionally appropriate?
 - How are research, confidentiality and personal privacy reconciled?

Note: These ideas owe much to Clifford Lynch.

Decisions

- Collection development and management
 - Legacy collections: Do you digitize? Can you digitize?
 - Native digital: Do you want to collect these materials?
- Context:
 - Network of similar institutions
 - Infrastructure
 - Resources – staff, volunteers, training, budget
 - Users: Who are they? How do they work?
- Access
 - Nature of the materials
 - Expectations of users
- Not much that deviates from standard practice in archives, libraries and museums. It's all about mission.

And Choices

- Predominantly analog collections; digital catalogs and finding aides
- Hybrid collections, based on collecting policies
- Digital collections for purposes of access; is there a tipping point?
- All digital – what do you do with the originals? And what parts of your collections are managed according to which policies?
- Does conversion mean preservation?
It depends.

Why does it cost so much?

- Does it cost so much? What is the value proposition?
- Because so much is unknown.
- We can reasonably expect:
 - Technology will become more stable.
 - Technological system solutions will appear, and these will be modular.
 - Costs of energy will rise, affecting heating and cooling as well as operations.
 - Organizational systems will offer alternatives. And will challenge institutional identity.
 - The learning curve will work with us as we simply become more accustomed to the medium and its challenges.

Surrogate for fear?

- Costs are necessarily open ended and hence unknown.
- Preservation is inherently an act of hope.
- Time will reduce some, if not many, sources of ambiguity.



Thank you.



References [1]:

- Ayris, P., R. Mcleod, et al. (2006). Lifecycle Information for E-literature: Full Report from the LIFE Project. JISC. London, UK, University College London and the British Library.
- Ayris, P., R. Mcleod, et al. (2008). LIFE² Final Project Report. JISC. London, UK, University College London and the British Library
- Beagrie, N., J. Chruszcz, et al. (2008). Keeping Research Data Safe: A Cost Model and Guidance for UK Universities. London, JISC.
- Berman F. (October 2008), "Research and Data," ARL/CNI Workshop, Arlington, VA.
- Chapman, Stephen. (2003) "Counting the Costs of Digital Preservation: Is Repository Storage Affordable?" Journal of Digital Information 4.2.
- Gantz, J. (January 2008). The Exploding Digital Universe: Implications for the Enterprise and Data Preservation, Presentation for the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Washington, D.C.

References [2]

- Dekker, R. E., Dürr, M Slabbertje, M. and K. van der Mee. An Electronic Archive for Academic Communities.(2001) ICEIS 2001/NDDL Workshop. April, 2002.
- Gantz, J. F. (2008). The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011, International Data Corporation (IDC).
- Palm, J.(2006). The Digital Black Hole. Stockholm, Sweden: Riksarkivet/National Archives.
- Science and Technology Council (2007). The Digital Dilemma: Strategic Issues in Archiving and Accessing Digital Motion Picture Materials, Academy of Motion Picture Arts and Sciences (A.M.P.A.S.): 74.
- Testbed Digitale Bewaring (2005). Costs of Digital Preservation. The Hague, Netherlands, Nationaal Archief of the Netherlands: 23.