Digital Preservation

Cost Models

David Wang

secure
sba-research.org

Outline

- Introduction
  - Motivation
- Characteristics of Cost Models
  - Cost structure and variables
- Existing Cost Models
- ULCC National Archive of Datasets (NDAD) Cost Analysis

Introduction (1)

- Motivation
  - Digital preservation requires **resources**
  - **Sustainability** of digital information assets
    - Key issue for business records, research data, cultural heritage collections
  - Ensure **funding**
  - Improvement of **cost effectiveness**
  - Basis for comparable solutions, supporting **decision-making**

Introduction (2)

- Difficulties for cost calculations
  - Digital preservation consists of interrelated activities
  - Activities can be **implemented differently** (meet different quality requirements)
  - Cost calculations require detailed cost information
  - Changes (systems/procedures) require updating of cost calculations
  - Rapid growth of assets causes increased budget pressure
  - Difficult to assess return of investment (ROI), effects of digital preservation not available at once

Characteristics (1)

- Cost structure
  1. Activity
  2. Resource
  3. Time

1. Activity

   OAIS often used as point of reference

   Activities are structured in categories with sub categories

Characteristics (2)

2. Resource
   - **Capital costs**
     Building space (server space, office space),
     equipment (server, network),
     energy,
     materials (storage media),
     et cetera.
   - **Labour costs**
     Differentiated by level of education and/or job function (developer, metadata officer, et cetera)

Characteristics (3)

2. Resource (continued)
   - **Direct costs**
     Associated with resources used for activities, e.g. acquisition of storage media and drives or staff costs for adding metadata (if spent resources directly measurable)
   - **Indirect costs**
     Incurred by usage of shared resources, e.g. general management, administration, facilities, systems
     Indirect costs cannot be distributed to specific activities due to lack of detailed measures

Characteristics (4)

2. Resource (continued)
   - **Variable costs**
     Depend directly on the amount of production, normally equal to direct costs
   - **Fixed costs**
     Do not vary with the amount of production, normally equal to indirect costs

Characteristics (5)

3. Time

Costs are often divided by accounting periods

Past costs (ex post) used in accounting

Future costs (ex ante) used in budgeting

– **One-time costs**
  Capital or investment cost
  (such as acquisition of equipment)

– **Recurring costs**
  Operating cost
  (such as consumption of media, energy and labour)

Characteristics (6)

3. Time (continued)
   - **Depreciation**
     Costs can be expressed as the depreciation of assets, e.g. the time a server becomes obsolete may be 5 years. With a 5-year period the cost of using is simply its acquisition cost. With a 1-year period the cost would be the depreciated acquisition cost (linear, exponential or other). Depreciation is a mechanism for distributing capital costs over estimated **useful lifetime**.

Characteristics (7)

3. Time (continued)
   Other important time aspects:
   General price increases (**inflation**), economic growth and cost of capital (**interest rates**), individual price changes related to specific resources, e.g. storage media, energy, office space

Existing Cost Models

- Test bed Cost Model for Digital Preservation, NL
- NASA Cost Estimation Tool, US
- LIFE^3 Costing Model, UK
- Keeping Research Data Safe, UK
- Cost Model for Digital Archiving, NL
- Cost Model for Digital Preservation, DK
- DP4lib Cost Model, DE
- PrestoPRIME Cost Model for Digital Storage, PrestoPRIME project
- Total Cost of Preservation, US
- Economic Model of Long-Term Storage, US

Test bed Cost Model for Digital Preservation

- Developed by the **National Archives of the Netherlands**

- Latest version of the model and the computational spreadsheet from 2005
- Purpose of estimating costs of long-term preservation and comparing the costs of applying different preservation approaches
- Information assets: Texts, emails, spreadsheets, databases

Test bed Cost Model for Digital Preservation

- List of **cost indicators** influencing total costs
  - Cost of a digital repository and preservation system
  - Personnel costs
  - Cost of developing or acquiring software and strategies
  - Cost of performing preservation action, and
  - Other costs

- Model based in OAIS terminology

Test bed Cost Model for Digital Preservation

- Cost model straightforward and easy to use
- Out of date
- Breakdown of activities not sufficiently detailed

NASA Cost Estimating Tool

- Developed for the **National Aeronautics and Space Administration (NASA)**

- First version published 2004, latest version September 2008
- Estimating **lifecycle costs** for ground data centre activities to improve budgets for NASA missions
- Information assets: Space data, multidimensional data sets

NASA Cost Estimating Tool

- Designed to generate life-cycle cost estimates
- Estimations based on **analogy approach**, using data about existing data activities for user-defined activities
- Work efforts expressed in **Full Time Equivalents** (FTE), denotes the time of a single person working full time for a year
- Model uses a data activity reference model (based on OAIS but preservation planning and migration not explicitly included)

NASA Cost Estimating Tool

- Covers **detailed** in-project activity information

- Steep learning curve, **high complexity** of the cost estimating tool

LIFE^3 Costing Model

- Life Cycle Information for E-Literature (LIFE) developed by **University College London and British Library**

- Project ended in 2010
- Improve planning and management of preservation projects by producing a predictive costing tool. The model covers all **lifecycle** stages
- Information assets: websites, e-journals, digitized newspapers, sound, word processing documents, small databases

LIFE^3 Costing Model

- Was developed in the context of libraries and Higher Education/Universities
- LIFE^3 estimates costs based on lifecycle stages:

  Creation, acquisition, ingest, metadata creation, bit-stream preservation and content preservation and access
- Tool provides default cost values which can be/need to be refined

LIFE^3 Costing Model

- Covers most archival aspects well

- Pre-ingest needs a wider array of activities

Keeping Research Data Safe (KRDS)

- KRDS was funded by **JISC** and conducted by Charles Beagrie Ltd, OCLC Research, the UK Data Archive, the Archaeology Data Service, the University of London Computer Centre, and the universities of Cambridge, King's College London, Oxford and Southampton
- Ended 2010, latest documentation from 2011

Keeping Research Data Safe (KRDS)

- KRDS aims to improve understanding of long-term preservation costs and its benefits to justify and sustain investments in digital repositories
- Information assets: research data
- KRDS cost model is
  - activity based
  - based on OAIS
  - a **framework** (no cost predicting tool included)

Keeping Research Data Safe (KRDS)

- Includes a Benefits Analysis Toolkit
  - Collection of detailed guides and worksheets
  - Provides examples of generic benefits and potential metrics
- Provides analysis of preservation costs of the ULCC National Digital Archive of Datasets (NDAD)

Keeping Research Data Safe (KRDS)

- Activity model of KRDS covers lifecycle of assets

- Calculations need to be developed by cost model users according to the user guide

NDAD Cost Analysis

- National Digital Archive of Datasets (NDAD) contains **UK government databases**
- Designated for **permanent** preservation as public records
- Cost data based on real costs of 2006
- ULCC estimated costs for 3-5 years in the future
  - Repository ingesting 36 data sets/year
  - 5 GB/data set
  - Total 180 GB/year

NDAD Cost Analysis

- 900 GB over 5 years
- **£5,282.93/GB (£5.28/MB)** for first year
- Ingest staff costs: £3,936.82/GB (£3.94/MB)
  - ¾ of overall costs
- Staff costs (ingest, development, management, publicity, and reporting) constitute **90% of overall annual cost**s
- Curation of research data is highly **labour intense**

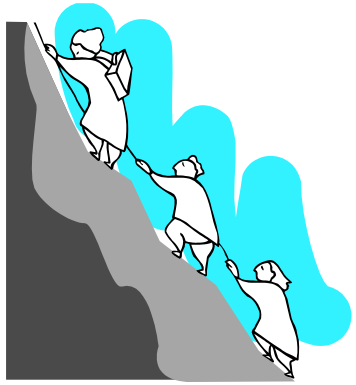
NDAD Cost Analysis

- Costs of simple bit preservation are low
  - Bit storage on tape (incl. multiple copies, multiple sites, periodic re-reading and checking, periodic migration): £0.0040/MB
  - Accessible copies on disk: £0.0038/MB
  - Administration and depreciation server costs: £0.0561/MB
  - Total simple bit storage: £0.0639/MB
    only 1% of overall costs (£5.28/MB)

NDAD Cost Analysis

- Key observations
  - Heavy predominance of **staff costs**, data curation highly **labour-intensive** activity
  - Cost of **simple bit storage** constitutes a **small proportion** of overall curation costs
  - **Ingest** (receiving data, preparing for long-term storage, incorporating into digital archive) receive **largest allocation** of resources

Challenges



- Cost models require a **breakdown** of the costs of digital curation **activities**
  - Curation costs are currently often not separated from other business activities
- No common agreed structures and definitions for modelling costs of digital information assets
  - Activity models often based on OAIS reference model but (local) activity structure of the organisations only comply to a limited degree

References

- Test bed Cost Model for Digital Preservation, https://web.archive.org/web/20061010043226/http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185&categorie=6
- NASA Cost Estimation Tool, http://opensource.gsfc.nasa.gov/projects/CET/
- LIFE3 Costing Model, http://www.life.ac.uk/3/
- Keeping Research Data Safe 2, http://www.beagrie.com/jisc/
- National Digital Archives of Datasets, http://www.nationalarchives.gov.uk/webarchive/archiving-datasets.htm
- 4C Project, http://4cproject.eu/

Thank you for your attention.

dwang@sba-research.org