Charles Beagrie

Keeping Research Data Safe

JISC Research Data Digital Preservation Costs Study

MPG Workshop Gottingen

June 2008

Overview

- Aim – investigate costs, develop model and recommendations
- Project team – Neil Beagrie, Julia Chruszcz, Brian Lavoie (OCLC), Cambridge, KCL, Southampton
- Method – detailed analysis of 2 cost models (LIFE & NASA CET) in combination with OAIS and TRAC; literature review;12 interviews; 4 case studies.
- 4 month study
- Final report and Exec Summ at http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx

UK Background

- Dual Support System
- Funded as a pre-cursor to a UK Research Data Service Feasibility Study
- Focus on universities
- Research costs linked to research funding
- Sustainability of research – UK universities move to Full Economic Costs (FEC) –
- Data management can be charged as direct or indirect costs to research grants
- Implications of research consolidation/excellence

What have we Produced?

- A cost framework consisting of:
  - activity model in 3 parts: pre-archive, archive, support services
  - Key cost variables divided into economic adjustments and service adjustments
  - Resources template for TRAC
  - Used in combination to generate cost/charging models
- 4 detailed case studies (ADS, Cambridge, KCI, Southampton)
- Data from other services.

Findings

Institutional Repository (e-publications):	Staff	Equipment (capital depreciated over 3 years)
Annual recurrent costs	1 FTE	£1,300 pa

Federated Institutional Repository (data): Annual recurrent costs	Staff	Equipment (capital depreciated over 3 years)
Cambridge	4 FTE	£58,764 pa
KCL	2.5 FTE	£27,546 pa

Findings

- Timing. costs c. 333 euros for the creation of a batch of 1000 records. Once 10 years have passed since creation it may cost 10,000 euros to ‘repair’ a batch of 1000 records with badly created metadata (Digitale Bewaring Project)
- Efficiency Curve effects – start-up to operational
- Economy of scale effects – Accession rates of 10 or 60 collections - 600% increase in accessions will only increase costs by 325% (ULCC)

Findings

- Unit costs – examples in Case studies for Archaeology, Chemistry, Humanities
- However costs depend on the adjustments (key cost variables)
- Like restaurant meals – final bill and unit costs depend on the choices and volume

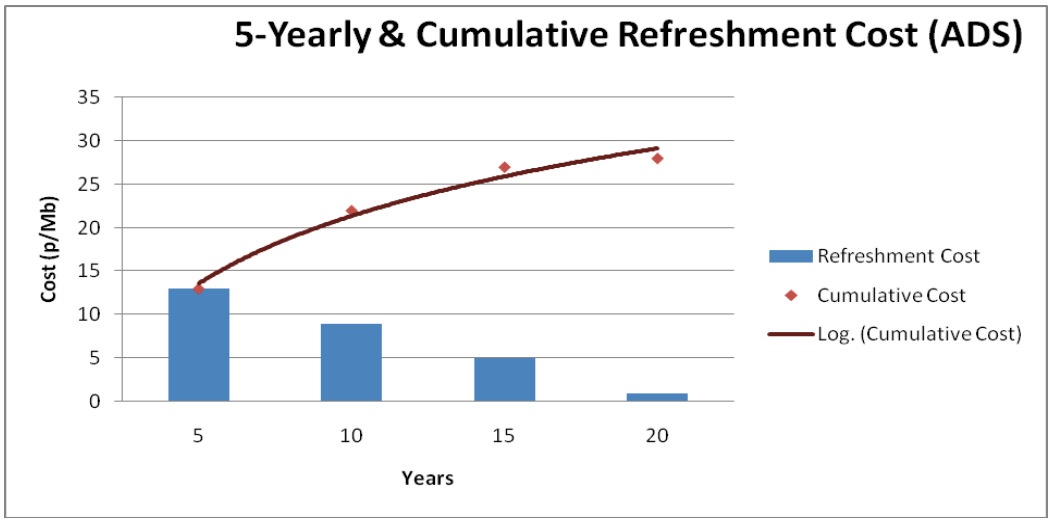
Findings

- National subject repositories costs (UKDA)

Acquisition and Ingest	Archival Storage and Preservation	Access
c. 42%	c. 23%	c. 35%

Findings

- ADS project of long-term preservation costs
- Implications for sustainability via project charges



- Preservation interventions (file format migrations)
- Long-term storage costs
- Assumptions of archive growth (economies of scale)
- Assumptions on “first mover innovation”

Findings

- NSB Long-lived data collections identifies 3 research data collection types with different preservation, access, and cost requirements:
  - **Research collections** – used by research team only, often limited retention and preservation needs;
  - **Community collections** – used by a discipline, data validation and community standards, preservation medium to long-term
  - **Reference collections** – used by many disciplines, major use of standards , quality control, long-term preservation;
- Data collections can move between levels (normally with substantial additional investment if upgraded)
- Triage – need to prioritise and restrain costs

UKRDS Feasibility Study

- Survey covered 4 universities: Bristol, Leeds, Leicester, Oxford
- Diverse Research Sizes: Oxford (£248m); Bristol (£81m); Leeds(£90m); Leicester(£37 m)
- Online Questionnaire completed by Bristol, Leeds, Leicester – Oxford 40 separate interviews with their research groups
- Interim Results –not weighted or checked
- 179 responses covering 500 researchers plus additional 200 researchers in Oxford

Data Storage

- Data life/usefulness after project
  - c 49% of data has a useful life of under 10 years
  - c 22% 10-50 years
  - c 27% is seen as having indefinite value.

Data Access

- most researchers share data – only c12% do not make their data available. Informal peer exchange/networks within research teams and with collaborators pre-dominant. Only c.19% share data via a data centre.
- access to other researchers data -In contrast c.43% access other researchers' data via a data centre.

Data Access

- Data use/users. The majority of the data is seen as useful/used by a small number of users. Only 24% have 20-100+ external users.
- In US National Science Board data collection levels terms, the majority of data surveyed would be “research” data collections with only 24% likely to be in a community or reference level data collection.

What's New?

- FEC based – not in or partial in other models but
  - Requirement for HEIs
  - Absence of FEC (a) distorts business cases eg for automation (b) cannot accurately compare in-house or out-source costs
- Not just DIY – application neutral – can cost for in-house archive, full or partial shared service(s), national/subject data centre archive charges
- Preservation: archival storage, preservation planning, data management, "first mover innovation"
- Tailored for research data: different collection levels, documentation+ metadata, products from data, etc

Recommendations

- **Recommendation 1**: The outcomes of this study should be considered and utilised by the forthcoming JISC Data Audit Framework study.
- **Recommendation 2: Departments and Central Services within HEIs should utilise recurrent data audits to inform both their initial appraisal and development of data policies and future capacity planning for services.**
- **Recommendation 3**: **HEIs should consider utilising the US National Science Board (the governing body for the National Science Foundation) long-lived data collection levels to aid understanding and categorisation of user requirements and costs over time.**
- **Recommendation 4: HEIs should consider federated structures for local data storage within their institution comprising data stores at the departmental level and additional storage and services at the institutional level. These should be mixed with external shared services or national provision as required. HEIs should work with and utilise national and international disciplinary data archives where these exist. The hierarchy of data stores should reflect the detailed nature of the content, services required, and the changing nature of its importance over time.**
- **Recommendation 5**: We recommend consideration of the study and further work on development and implementation of relevant cost models and tools to HEIs, research funders, and service providers.

Recommendations

- **Recommendation 6:** JISC should produce a short briefing paper or summary of this report and its findings aimed at senior managers including university academics, administrators and research support services.
- **Recommendation 7**: **JISC should consider developing project costing tools to build on and implement work within this study. These tools may be valuable for some of JISC's own projects and may also be of interest to other research funders and have potential for joint funding and development.**
- **Recommendation 8: JISC should consider undertaking additional work to examine how the cost components and variables defined in our framework can be further quantified, and what additional data and data collection mechanisms are needed to support them.**
- **Recommendation 9**: **JISC should consider further detailed study of longitudinal data for digital preservation costs and cost variables to extend the work of this study. Possibly this could be part of a UK based taskforce to feed into its joint international work on digital preservation costs.**
- **Recommendation 10**: JISC and/or other funders should consider funding further work on quantifying the benefits of research data preservation.

Cost Observations for Repositories

- Not just formula of function costs
- Can illustrate effect of some choices on costs
- Sustainable project archive funding model?
- Start-up v running costs
- bleeding-edge costs – “first mover innovation”
- Audit/capacity planning
- Not last word on costs....