Keeping Research Data Safe:
costs of research data preservation

Neil Beagrie (Charles Beagrie Ltd)
PASIG conference Oct 2012
Outline

• Background to KRDS
• Framing the Costs
• Findings on Costs (“rules of thumb”)
• Key Reminders
“Keeping Research Data Safe”

• Project sponsored and funded by UK Joint Information Systems Committee (JISC)
  – Charles Beagrie Consulting, +10 UK universities and data centres and OCLC Research

UK perspective and research data but wider applicability: a general model
What was Produced?

- DP cost activity model;
- Published cost data (8 case studies and 13 costs surveys – rare data);
- KRDS Factsheet (4 page summary of key findings – see print copies);
- KRDS User Guide;
- KRDS Benefits Tool kit (link to costs);
- Follow-on projects – impact & benefits.
Keeping Research Data Safe Factsheet

Cost issues in digital preservation of research data

What Costs Most?
Acquisition and ingest costs most. The costs of archival storage and preservation activities are consistently a very small proportion of the overall costs and significantly lower than the costs of acquisition/ingest or access activities for all our case studies. Note we believe early preservation action during ingest or pre-ingest produces lower costs over the life cycle as a whole. (KRDS1, p.23; KRDS2, pp.31-32)

Activity Costs for the Archivology Data Service
- Outreach/Acquisition/Ingest: c.55%
- Archival Storage and Access: c.35%
- Preservation: c.10%

Recommendation to Funders:
From our research, it is likely that the largest potential cost efficiencies will come from future tool development supporting automation of ingest and access activities for creation and preservation. (KRDS2, p.63)

Impact of Fixed Costs
- The long costs of creation/preservation are dominated by fixed costs that do not vary with the size of the collection.
- Staff has the major cost component overall and there is a minimum basic level of staff cover, skills and equipment required for any service.
- Activities characterised by significant fixed costs can reduce the per-unit cost of long-term preservation by averaging economies of scale. (KRDS2, pp.32-34, 78-80)

Declining Costs over Time
We found a trend of relatively high preservation costs in the early years reducing substantially over time for data collections. An example is the preservation costs projected for the Archivology Data Service (ADS) based on their experience of the first 10 years of operating the data service. (KRDS1, p.4-5)

Recommendation to Institutions:
Repositories should take advantage of economies of scale, using multi-institutional collaboration and outsourcing as appropriate. Once core capacity is in place additional costs can be justified at increasing levels of efficiency and lower cost. (KRDS1, pp.77-78)

Benefits from digital preservation of research data

Dimension 1
Direct Benefits
- Increased research productivity
- Enhanced access to data
- Improved research coordination
- Improved data discovery
- Improved data long-term accessibility

Indirect Benefits (Costs Avoided)
- Reduced research opportunities
- Lower future preservation costs
- Improved data visibility
- Improved data discoverability
- Improved data accessibility

Dimension 2
Near-Term Benefits
- Value to current researcher & students
- Avoids costs over time associated with long-term storage
- Saves for data intensive research
- Improves availability of research data

Long-Term Benefits
- Value to future researchers & students
- Avoids costs over time associated with long-term storage
- Saves for data intensive research
- Improves availability of research data

Dimension 3
Private Benefits
- Reputation
- Reduced access time
- Improved reliability

Public Benefits
- Access to research data for future research
- Improved reliability
- Improved access to research data

S-5 Yearly & Cumulative Cost for Refreshment in AIDS

The implications of these factors and projection for sustainability of data archives e.g., via archive changes to
KRDS Costs

• KRDS Approach to Costs
  – Implementation of a lifecycle costing approach to research data preservation
  – Method
    • detailed analysis of 4 models: LIFE digital lifecycle model & NASA Cost Estimation Toolkit in combination with OAIS and UK Transparent Approach to Costing (TRAC)
    • Plus literature review; interviews; detailed case studies.
Activity Model

• Enumeration of full range of activities required to support long-term preservation of research data
  – How are costs allocated across these activities?
• Three major categories:

  Pre-Archive Phase
  Activities related to the creation of research data for later transfer to the archive

  Archive Phase
  Activities which occur during period of archival retention

  Support Services
  Administrative and non-preservation technical services (e.g. campus computing, Finance, HR etc.)
Multiple levels of granularity

Phase

Activity

Sub-activity

Archive

Acquisition

Selection

Negotiate submission agreement

Outreach and support

Ingest

Receive submission

Quality assurance

Generate information package for Archive

Generate administrative metadata

Generate descriptive metadata and user documentation

Coordinate updates

Reference linking

Costs can be allocated at any level
# KRDS High-level Activity Model

<table>
<thead>
<tr>
<th>Pre-Archive Phase</th>
<th>Outreach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initiation</td>
</tr>
<tr>
<td></td>
<td>Creation</td>
</tr>
<tr>
<td>Archive Phase</td>
<td>Acquisition</td>
</tr>
<tr>
<td></td>
<td>Disposal</td>
</tr>
<tr>
<td></td>
<td>Ingest</td>
</tr>
<tr>
<td></td>
<td>Archival Storage</td>
</tr>
<tr>
<td></td>
<td>Preservation Planning</td>
</tr>
<tr>
<td></td>
<td>First Mover Innovation</td>
</tr>
<tr>
<td></td>
<td>Data Management</td>
</tr>
<tr>
<td></td>
<td>Access</td>
</tr>
<tr>
<td>Support Services</td>
<td>Administration</td>
</tr>
<tr>
<td></td>
<td>Common Services</td>
</tr>
<tr>
<td>Estates</td>
<td></td>
</tr>
</tbody>
</table>
Cost variables

- Key variables that shape cost of preserving research data
- **Service Adjustments**: “adjustable” aspects of the preservation process that impact costs
  - i.e., choices; preservation goals
  - Examples: number of acceptable file formats; volume and frequency of deposits; richness of metadata description …
- **Economic Adjustments**: spreading costs over time
  - Rate of inflation/deflation: recurring costs subject to changes in prices
  - Rate of depreciation: upfront expenditures for resources that are consumed gradually over time
Resource Template

- Need to align KRDS Cost Framework with existing costing systems in UK universities
- Transparent Approach to Costing (TRAC) Model
  - Endorsed by UK HE, government, research funders
  - Express Full Economic Cost: “the total costs to an institution of undertaking a project or activity in a sustainable manner”
  - Cost categories (resources): Staff, Equipment, Travel, Consumables, Estate costs, Indirect costs
- Resource Template: organizes TRAC cost categories according to Activity Model, in a form closely aligned to TRAC methodology.
Putting it all together

Identifies cost allocations across preservation process

Activity Model

Cost Variables

Resource Template

Pulls all of it together into TRAC-friendly costing model

Service adjustments: adjust costs to specific requirements
Economic adjustments: spread costs over time
## Cost Findings

<table>
<thead>
<tr>
<th>Section</th>
<th>Staff</th>
<th>Equipment (capital depreciated over 3 years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institutional Repository (e-publications): Annual recurrent costs</td>
<td>1 FTE</td>
<td>£1,300 pa ($2096 pa)</td>
</tr>
<tr>
<td>Federated Institutional Repository (data): Annual recurrent costs</td>
<td>4 FTE</td>
<td>£58,764 pa ($97,750)</td>
</tr>
<tr>
<td>Cambridge</td>
<td>2.5 FTE</td>
<td>£27,546 pa ($44,415)</td>
</tr>
</tbody>
</table>
KRDS Cost Findings

- National subject repositories costs (UKDA)

<table>
<thead>
<tr>
<th></th>
<th>Acquition and Ingest</th>
<th>Archival Storage and Preservation</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>c. 42%</td>
<td>c. 23%</td>
<td>c. 35%</td>
<td></td>
</tr>
</tbody>
</table>

“getting stuff in and out” costs more than “keeping it (bit preservation + migration)”; staff costs c70% of total costs.
KRDS Cost Findings

- ADS projection of long-term preservation costs
- Decrease sharply over time

- Includes preservation interventions (file format migrations)
- Declining long-term storage costs
- Assumptions of archive growth (economies of scale)
- Assumptions on “first mover innovation”/technical development
Dryad Costs per Paper: effects of curation levels and volume
Some Key Reminders

- Staff costs most significant factor (c 70%)
- Accession/access cost more than preservation
- Costs of preservation found to decline over time
- Costs depend on the service adjustments like NSB Data Collection levels (key cost variables)
- Economies of scale important
- Like restaurant meals – final bill and unit costs depend on the choices and volume
Further Information

KRDS webpage:
www.beagrie.com/krds.php