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


Charles Beagrie

KEEPING RESEARCH DATA SAFE

A COST MODEL AND GUIDANCE FOR UK UNIVERSITIES

Neil Beagrie, Julia Chruszcz, and Brian Lavoie

with case studies contributed by the Universities of Cambridge, Southampton, King's College London, and the Archaeology Data Service University of York.

Final Report - April 2008

Prepared by:

Charles Beagrie Limited

www.beagrie.com

A study funded by

JISC

Copyright HEFCE 2008

The authors have asserted their moral rights in this work

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.

KRDS Factsheet

FS_v_8.docx - Microsoft Word

Home Insert Page Layout References Mailings Review View

Keeping Research Data Safe Factsheet

Cost issues in digital preservation of research data

This factsheet illustrates for institutions, researchers, and funders some of the key findings and recommendations from the JISC-funded Keeping Research Data Safe (KRDS1) and Keeping Research Data Safe 2 (KRDS2) projects. Further information on the research and findings can be found in the final reports.

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 lifecycle as a whole. (KRDS1, p.25; KRDS2, pp.31-52)

Activity Costs for the Archaeology Data Service

Outreach/ Acquisition/ Ingest	Archival Storage and Preservation	Access
c. 55%	c. 15%	c. 31%

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 curation and preservation. (KRDS2, p.83)

Impact of Fixed Costs

- The costs of long-term data curation/preservation are dominated by fixed costs that do not vary with the size of the collections;
- Staff are the major cost component overall and there is a minimum base-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 leveraging economies of scale.

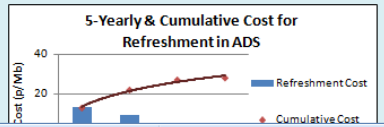
(KRDS2, pp.32-34, 79-80)

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 content can be added at increasing levels of efficiency and lower cost. (KRDS1, pp.77-78)

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 Archaeology Data Service (ADS) based on their experience of the first 10 years of operating the data service. (KRDS1, pp.4-6)



Recommendation to Funders and Institutions

The implications of these factors and projection for sustainability of data archives e.g. via archive charges to

Benefits from digital preservation of research data

Analysis of the costs of preserving research data sets is not enough to assess economic feasibility. Cost analysis should be accompanied by a framing of the anticipated benefits. As a first step in this process, KRDS has defined a few important dimensions and a taxonomy that illuminate the broad contours of the benefits digital preservation investments potentially generate. Users can sharpen these generic expressions of preservation benefits into more focused value propositions for specific cases. The KRDS taxonomy for categorising the benefits from long-term preservation of research data (KRDS2, p.55) is presented below. It is illustrated with examples from our studies.

Dimension 1	**Direct Benefits**	**Indirect Benefits (Costs Avoided)**
	-New research opportunities -Scholarly communication/access to data -Re-purposing and re-use of data -Increasing research productivity -Stimulating new networks/collaborations -Knowledge transfer to industry - Increasing skills base of researchers/students/staff -Increasing productivity/economic growth -Verification of research/research integrity -Fulfilling mandate(s)	-No re-creation of data -No loss of future research opportunities -Lower future preservation costs -Re-purposing data for new audiences -Re-purposing methodologies -Use by new audiences -Protecting returns on earlier investments
Dimension 2	**Near-Term Benefits**	**Long-Term Benefits**
	-Value to current researcher & students -No data lost from Post Doc turnover -Short-term re-use of well curated data -Secure storage for data intensive research -Availability of data underpinning journal articles	-Secures value to future researchers & students -Adds value over time as collection grows and develops critical mass -Planned management from an early stage in the research lifecycle is ultimately more cost-effective than late intervention (providing proper selection of what to keep is done)
Dimension 3	**Private Benefits**	**Public Benefits**
	-Benefits to sponsor/funder of research/archive -Benefits to researcher -Fulfil grant obligations -Increased visibility/citation -Commercialising research	-Input for future research -Motivating new research -Catalysing new companies and high skills employment

Direct Benefits

Understanding costs as part of curation saves money. KCL and Southampton currently out-source archival storage to the Atlas Data Store a central repository maintained by the Science and Technology Facilities Council (STFC). Outsourcing to Atlas has allowed the NCS at Southampton to reduce costs for archival storage by 41% between when this was an in-house and staff-intensive and when this was outsourced and highly automated. (KRDS1, pp.70, 74)

Indirect Benefits (Costs Avoided)

The Digitale Bewaring Project in the Netherlands, which focused on government electronic records estimated it costs approximately 333 euros for the creation of a batch of 1,000 records in an appropriate manner at creation i.e. in the Pre-Archive phase. Conversely once 10 years have passed since creation it may cost 10,000 euros to 'repair' a batch of 1,000 records with badly created metadata. (KRDS1, p.25)

Near-Term Benefits

The constant turnover of post-doctoral researchers often results in lost data. Currently there are no established mechanisms to routinely collect and organise the data that post-doctoral researchers generate. In some cases, researchers that generated data several years ago could not make sense of them now as they had not kept

Long-Term Benefits

One advantage of archiving data over many years is that long time series of consistent data are built up. Richard Berthoud has analysed the General Household Survey between 1974 and 2005, to describe changing patterns of advantage and disadvantage in employment. The analysis was described by the civil servant responsible for commissioning the research as having made more

Page: 1 of 4 Words: 1,914 English (United Kingdom) 80%

14:48 08/11/2010

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

Pre-Archive Phase	Outreach
	Initiation
	Creation
Archive Phase	Acquisition
	Disposal
	Ingest
	Archival Storage
	Preservation Planning
	First Mover Innovation
	Data Management
	Access
Support Services	Administration
	Common Services
Estates	

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

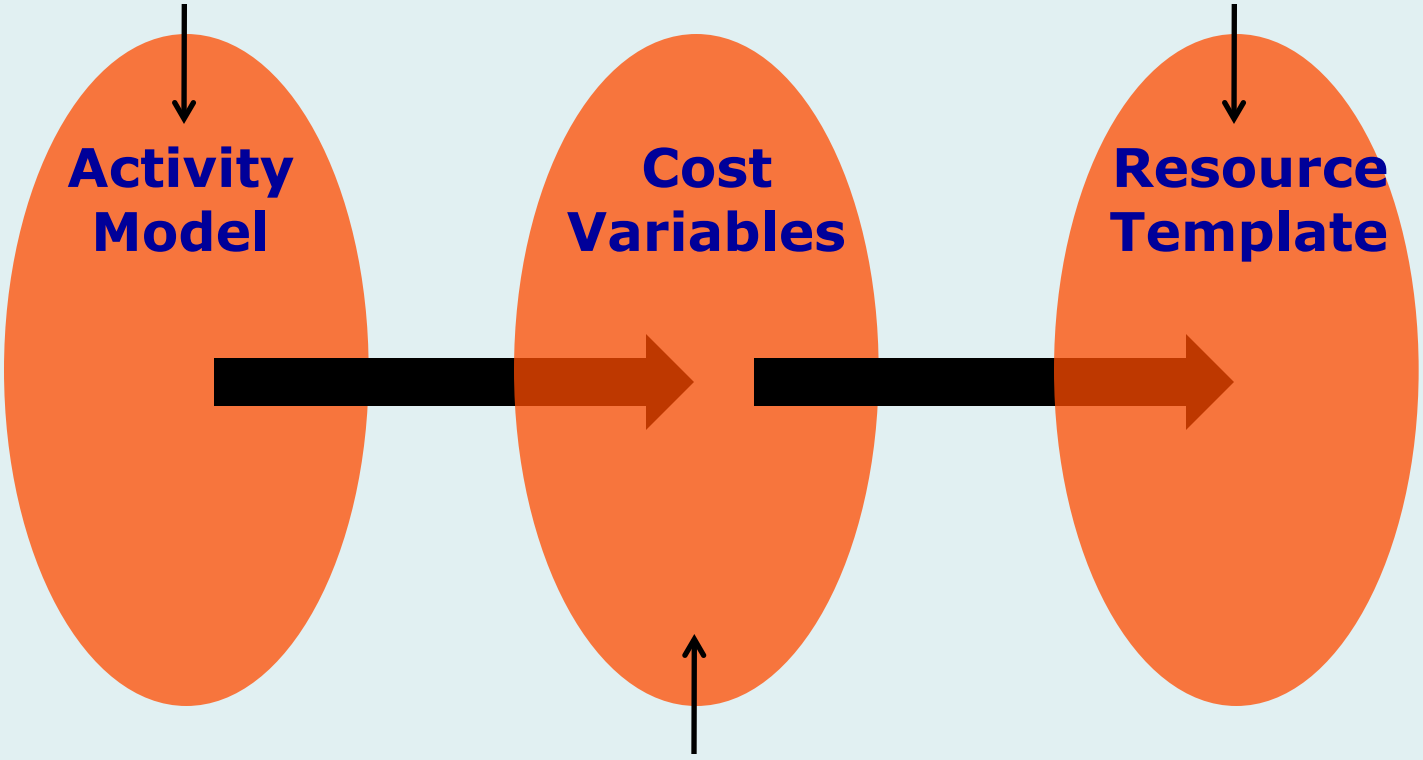
Resource Template

- 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

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

Institutional Repository (e-publications):	Staff	Equipment (capital depreciated over 3 years)
Annual recurrent costs	1 FTE	£1,300 pa ($2096 pa)

Federated Institutional Repository (data): Annual recurrent costs	Staff	Equipment (capital depreciated over 3 years)
Cambridge	4 FTE	£58,764 pa ($97,750)
KCL	2.5 FTE	£27,546 pa ($44,415)

KRDS Cost Findings

- National subject repositories costs (UKDA)

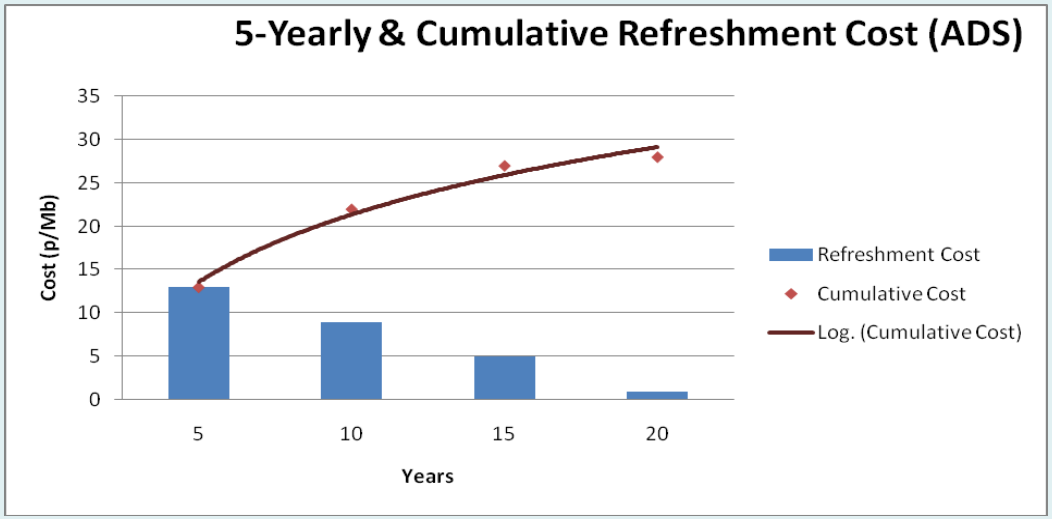
Acquisition and Ingest	Archival Storage and Preservation	Access
c. 42%	c. 23%	c. 35%

"getting stuff in and out" costs more than "keeping it (bit preservation + migration)";

staff costs c70% of total costs.

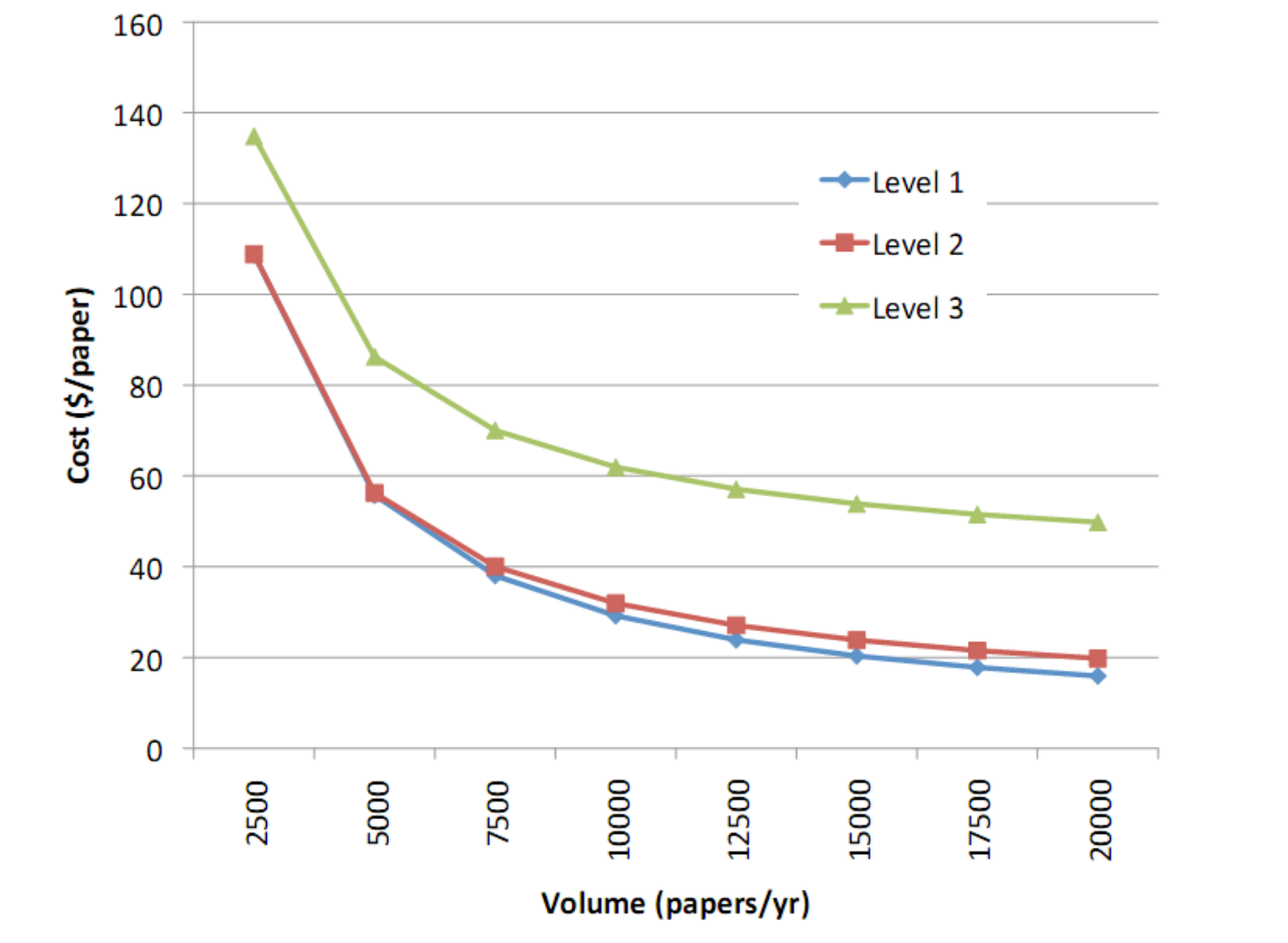
KRDS Cost Findings

- 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