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 and on the KRDS website.

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.

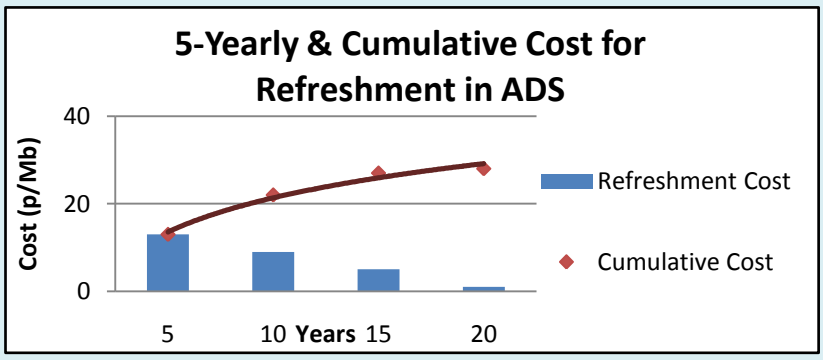
(KRDS2, pp.32-34, 79-80)

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)



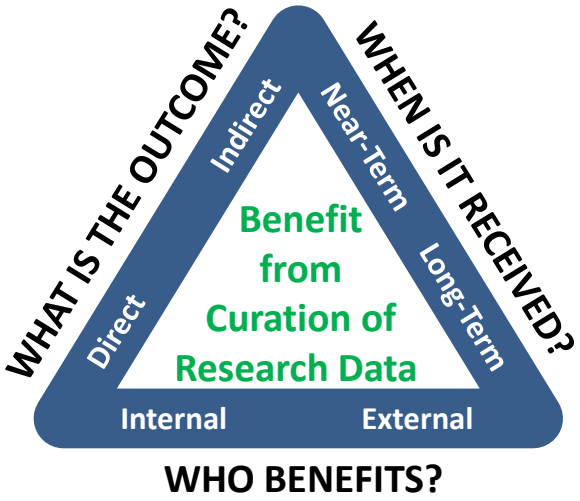
Costs for archival storage and preservation ("refreshment") decline to a minimal level over 20 years

Recommendation to Funders and Institutions

The implications of these factors and projection for sustainability of data archives e.g. via archive charges to project budgets, are notable and worthy of more extensive study and testing. (KRDS1, pp.5-6)

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 Benefits Framework and a toolkit which includes a list of common generic benefits. Users can sharpen these short generic expressions of preservation benefits into more focused value propositions for specific cases. The KRDS Framework for categorising the benefits from long-term curation/preservation of research data is presented to the right. It is illustrated below with examples from our studies.



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 (e.g. 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 enough information on how the data was created. In these circumstances, well-curated data has clear short and medium-term benefits. (KRDS2, p.60)

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 difference to policy thinking than any other project for which he had been responsible. (KRDS2, p.72)

Internal Benefits

A curated and preserved research data set may generate internal benefits if the research data set is made publicly available and is frequently used and re-used by external researchers, this may increase the visibility and impact of the original research, and by extension, enhance the reputation and standing of the researcher and the institution in which it was created. (KRDS2, p. 62)

External Benefits

External benefits may manifest themselves on a variety of scales: across a group of collaborating universities, across the scientific community as a whole, and even on an economy-wide scale, to the extent that long-term preservation of research data enhances the prospects for commercialising scientific discoveries, catalysing new companies, and expanding opportunities for high-skill employment. (KRDS2, p.62)

Additional examples of benefits are available in the KRDS Benefits Toolkit and in the KRDS final reports.

Institutional issues: repository models

Once the generic benefits and those specific to your institutional goals and ambitions articulated within institutional strategies are well understood, it is important to define requirements. You can then apply the KRDS cost model to these requirements to estimate the level of investment needed for the preservation of research data. This will then contribute to building a business case, necessary to release funds or attract investment. When decisions and plans are being made to progress research data preservation initiatives within universities, there are a number of areas that require careful thought. These will cover both the wider institutional considerations, and the detail of applying a cost model to the preservation of research data. One will inevitably influence the other.

Repository models and structures

- There are a number of different service models and structures for research data preservation at international, national, and local level. There are significant differences and needs between disciplines.
- Research data is not as homogenous as research publications and is less likely to be available through a single institutional repository.
- Subject knowledge, preservation and curation skills are needed for long-term management of research data.
- The staffing and storage requirements are more substantial for research data preservation than for e-print repositories. Annual recurrent costs for central data repositories are therefore higher than for e-print repositories.
- In some disciplines national and occasionally international data repositories are/will be available.
- Potentially there is considerable scope for economies of scale across HEIs through either shared services or disciplinary data centres or centralised repositories at national level.
- Individual researchers are likely to feel alienated if archiving only occurs at an institutional level.

(KRDS1, pp.67-75)

Type of repository	Reference	Staff	Equipment (capital depreciated over 3 years)
Institutional Repository (e-publications)	SHERPA project	1 FTE	£1,300 pa
Federated Institutional Repository (data)	KCL case study	2.5 FTE	£27,546 pa
Federated Institutional Repository (data)	Cambridge case study	4 FTE	£58,764 pa

Annual Recurrent Costs: central data repository vs typical institutional repository for e-publications (KRDS1, p.4)

Recommendations to Institutions

- Consider federated structures for local data storage 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.
- 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.

(KRDS1, pp.77-78)

In their own words

It is important to consider the Department level in this landscape, in addition to the overall institutional level. It is an academic's natural affiliation and an environment they understand and can often have an influence on, i.e. it is at this level where money can be raised and decisions surrounding 'what is important' can be made by the most appropriate people.

(KRDS1, pp.123-124)

Institutional issues: cost variables and data collection levels

Key Cost Variables

- Institutions may control some of the unpredictability of future costs by limiting the future effect of some service variables e.g. the timing of actions has important implications for costs.
- Access costs are potentially the most variable area of costs. Considering some of the access functions as value-added services can make it easier to predict long-term costs.
- Evolution of technology and the availability of commercial off the shelf software or mature open source software will have significant effect on costs.
- Data collection levels and preservation aims have a major overall influence on a number of key cost variables.

(KRDS1, pp.24-35)

Recommendation to Funders and Institutions

- Implement KRDS in cost spreadsheets and continue research on implementation variables and metrics that could enhance them.
- Future researchers and their funders should note from our work that longitudinal studies of digital preservation costs are best developed from relatively recent cost evidence.

(KRDS2, pp.82, 84)

Data Collection Levels

Service requirements for different data collections are likely to vary considerably with data having different value and requirements for access over time. Significant costs are associated with moving data collections from one level to another over time. (KRDS1, pp.59-60, 164-165)

Recommendation to Institutions

Consider utilising the US National Science Board (the governing body for the National Science Foundation) long-lived data collection levels (research database collection; resource or community data collection; reference collection) to aid understanding and categorisation of user requirements and costs over time. (KRDS1, p.77)

References

KRDS website and KRDS Benefits Analysis Toolkit: http://www.beagrie.com/krds.php

KRDS1: Beagrie, N., Chruszcz, J., and Lavoie, B. (2008), *Keeping Research Data Safe: a cost model and guidance for UK universities*, Final Report April 2008, available from http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

KRDS2: Beagrie, N., Lavoie, B., and Woollard, M. (2010), *Keeping Research Data Safe 2*, Final Report April 2010, available from http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Acknowledgements

This factsheet has been prepared by Charles Beagrie Ltd and JISC. The Keeping Research Data Safe studies have been funded by JISC and conducted by a partnership of the following institutions: Charles Beagrie Ltd, OCLC Research, the UK Data Archive, the Archaeology Data Service, the University of London Computer Centre, UKOLN/DCC, and the universities of Cambridge, King's College London, Oxford, Southampton and UCL. For more information see the KRDS website or contact: info@beagrie.com.

Charles Beagrie Ltd
2 Helena Terrace, College Street, Salisbury SP1 3AN, United Kingdom
Tel: +44 (0)709 204 8179, Fax: +44 (0)709 204 8179, Email: info@beagrie.com

JISC Executive
King's College London 1st Floor, Brettenham House, 5 Lancaster Place, London WC2E 7EN, United Kingdom