User Guide

For

Keeping Research Data Safe

Assessing Costs/Benefits of Research Data Management, Preservation and Re-use

Version 2.0 – July 2011

Prepared by:

Charles Beagrie Limited

www.beagrie.com

funded by

JISC

Contents

1. Introduction

1.1. Why Use This Guide?

Keeping Research Data Safe (KRDS) is a cost framework that can be used to develop and apply local cost models for research data management and long-term preservation. The exact application may depend on the purpose of the costing, which might include:

- identifying current costs;
- identifying former or future costs;
- comparing costs across different collections and institutions which have used different variables;
- developing a charging policy or appropriate archiving costs to be charged to projects;
- focussing in more selectively on particular activities and modelling the effect of changes to specific processes.

In addition, it includes a Benefits Analysis Toolkit and discussion of benefits which provides a valuable starting point and framework for assessing the impact and benefits of research data management and preservation activities.

Finally, KRDS has been a significant research project establishing many key “rules of thumb” for digital preservation costs and approaches to sustaining digital research data. Even those who do not wish to or cannot allocate the resources to develop local models based on KRDS are likely to benefit from its key findings and exemplars, covered in later sections of the Guide.

The major outputs from KRDS have been the project final reports (KRDS 2008 and KRDS 2010), the Benefits Analysis Toolkit (http://www.beagrie.com/krds-i2s2.php) and the supplementary materials to the KRDS2 final report available from the KRDS2 project website (http://www.beagrie.com/jisc.php). The KRDS final reports have been extremely well received by the community. However the project outcomes such as case studies and guidance are now split over two long reports, appendices and supplementary material.

The KRDS User Guide has been developed to support easier assimilation of the combined work of the KRDS1 and KRDS2 projects by those wishing to implement the tools or key findings. The User Guide is an edited selection and synthesis of the KRDS reports combined with newly commissioned text and illustrations. It provides a succinct summary of key implementation guidance and tools, links to prepared extracts such as case studies from the reports, and additional guidance on its application.

1.2. Audience

Although based on UK experience and practice, there are many aspects of its work and approach which are relevant to an international audience. Similarly although tailored to research data and aimed primarily at a research data audience, there are broader lessons in terms of digital preservation costs and benefits that can be transferable to other sectors.

2. The KRDS Costs Framework

2.1. Introduction

KRDS is an example of a life-cycle costing method applied to research data. Life-cycle costing models a life-cycle for a specific process(es) and then identify measurable component activities, cost drivers (variables that affect the costs of the activity e.g. volumes, formats etc), and resources (staff time, equipment etc) to provide an understanding of costs for that process.

KRDS sets out the broader cost framework and guidance within which the KRDS Activity Model can be applied. That cost framework consists of three parts:

- **KRDS Activity Model.** A generic activity model for research data identifying activities with cost implications for preservation and ordering them in a nested hierarchy of Phases, Activities, and Sub-activities.
- **Cost Drivers**. Key variables (e.g. salary levels or rates of inflation), which affect the cost of preservation activities. The cost drivers are divided into two major groups: economic adjustments and service adjustments.
- **Resources Template**. This presents categories ("resource pools") of cost (e.g. staff or equipment) and duration (year 1, year 2, etc) in a simplified, generic form closer to that used in the cost methodologies of UK universities based on the Transparent Approach to Costing method (TRAC) – http://www.hefce.ac.uk/finance/fundinghe/trac/.

Typically the KRDS Activity Model will help identify activities on which resources are expended, the economic adjustments help spread and maintain these over time, and the service adjustments help identify and adjust resources to specific requirements. The resources template provides a framework to draw these elements together so that they can be implemented in a TRAC-based cost model. Typically the cost model will implement these as a spreadsheet, populated with data and adjustments agreed by the institution.

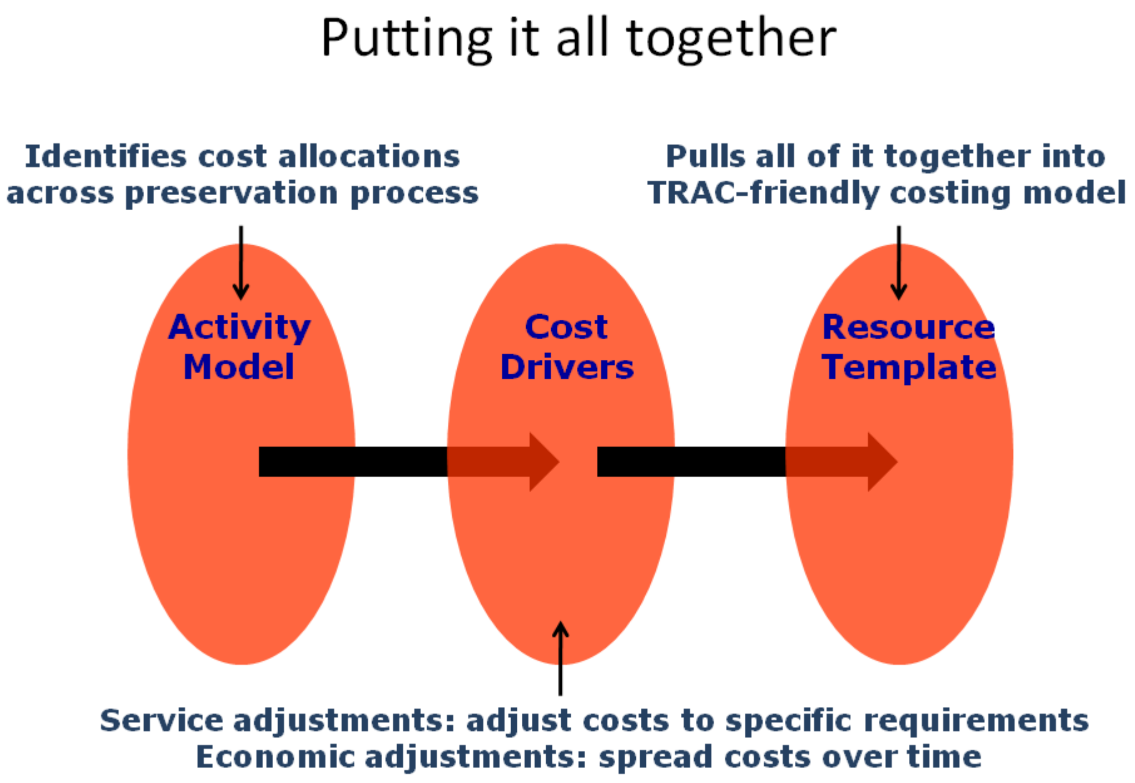

Putting it all together
Identifies cost allocations across preservation process
Pulls all of it together into TRAC-friendly costing model
Activity Model
Cost Drivers
Resource Template
Service adjustments: adjust costs to specific requirements
Economic adjustments: spread costs over time

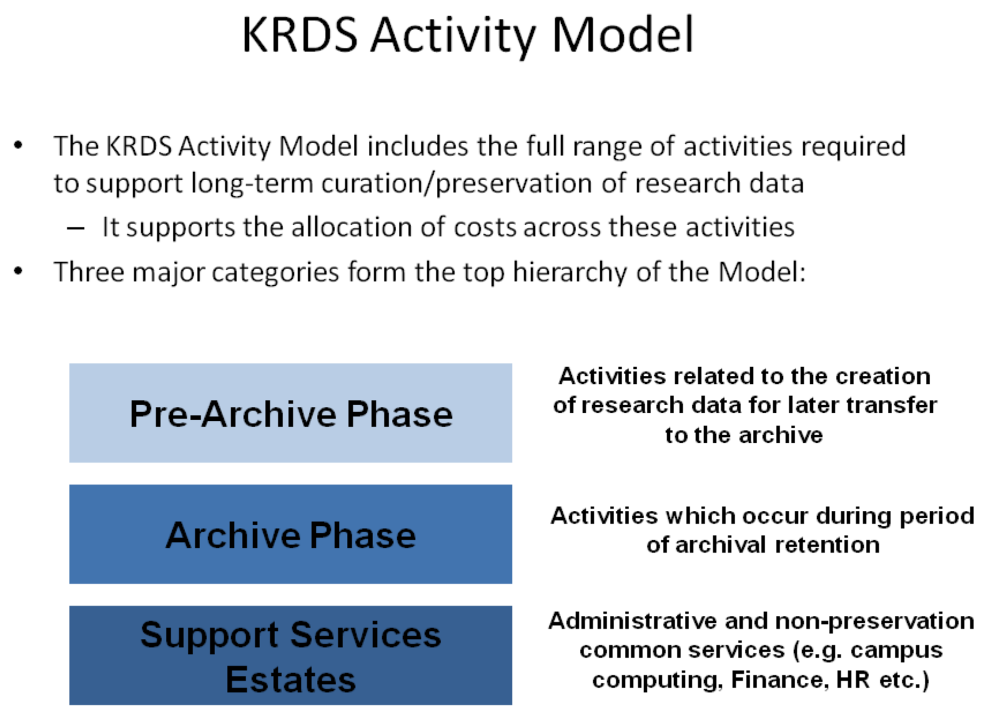
Figure 1: Putting it all Together – an Overview of the KRDS Costs Framework

The three parts of the cost framework can be used in this way to develop and apply local cost models.

Each of the main components of the cost framework is described in more detail below beginning with the next section, which sets out how different activities can be classified and mapped into a consistent model and costs attributed to activities. An approach to assessing the benefits arising from these activities is described in section 5.

2.2. KRDS Activity Model

The KRDS Activity Model identifies research data activities with cost implications for preservation. It is organised in a nested hierarchy of levels beginning with Pre-Archive, Archive, and Support Services and Estates. Typically Pre-Archive activities relate to all activities related to data creation and management for research projects in universities or other research institutes prior to archiving, and Archive activities to data archiving repositories run by universities or third-parties. Both of these relate to life-cycle costs for research data. Activities in Support Services can support either Pre-Archive or Archive activities and typically will be part of the existing infrastructure for finance, IT, and other common services. Estates are the TRAC category for buildings and other infrastructure. Support Services and Estates are included in calculating full economic costs.



KRDS Activity Model

- The KRDS Activity Model includes the full range of activities required to support long-term curation/preservation of research data
  - It supports the allocation of costs across these activities
- Three major categories form the top hierarchy of the Model:

Pre-Archive Phase

Activities related to the creation of research data for later transfer to the archive

Archive Phase

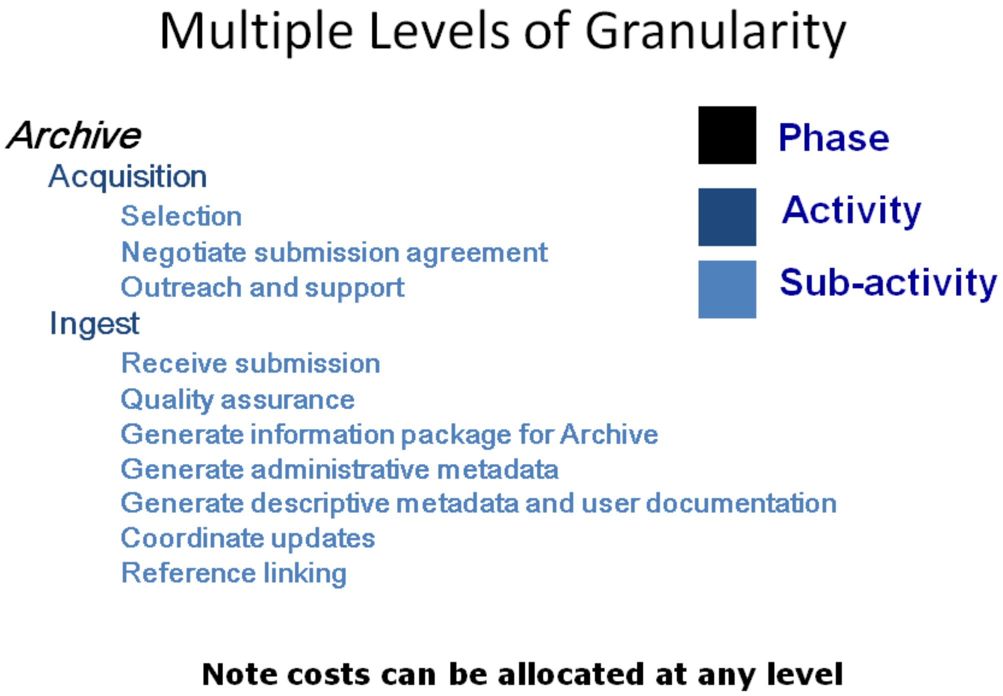
Activities which occur during period of archival retention

Support Services Estates

Administrative and non-preservation common services (e.g. campus computing, Finance, HR etc.)

Figure 2: Top Hierarchies of the KRDS Activity Model

Costs can be allocated at any level in the Activity Model and the Model can be applied at different levels of granularity for different purposes.



Multiple Levels of Granularity

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

■ Phase

■ Activity

■ Sub-activity

Note costs can be allocated at any level

Figure 3: Multiple Levels of Granularity in the KRDS Activity Model

KRDS caters for potential dual application of the Activity Model with two “versions” presented at different levels of detail to allow for different types of costing.

2.2.1. KRDS “Lite” Version of the Activity Model

A single page overview, the KRDS “Lite” version of the Activity Model (see Figure 4) has been produced subsuming the sub-activities and consisting of just:

- the main Phases, e.g. Archive; and
- the activities e.g. Ingest.

This provides a high-level granularity of the Model for allocating costs which could be suitable for a cost management application (sufficient to understand overall allocation of costs). This can be obtained with a much lower overhead in terms of capturing the required cost information and may be helpful to some institutions.

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

Figure 4: “Lite” version of the KRDS Activity Model (First published in KRDS2, p. 14)

2.2.2. KRDS "Detailed" Version of the Activity Model

The "detailed" version of the KRDS Activity Model provides options for more detailed operations planning and process improvement requiring finer levels of granularity. It is also essential for providing the definitions and scope of the phases, activities and sub-activities, necessary for mapping local activities accurately and consistently to KRDS.

A sample extract of the Model is provided below. The full detailed version of the Model can be downloaded from http://www.beagrie.com/KRDS2_Activity_Model_detailed.doc.

Activity	Sub-activity	Scope Notes
Outreach	Guidance on best practice and archiving requirements and other support and training by the archive for researchers submitting funding proposals or creating research data. This may be targeted at potential depositors and/or broader communities and data producers.	
Initiation	The activities involved in initiating research activity that will generate research data. Included to note any significant implications for preservation costs downstream.	
	Project design	Take into account implications of any data creation or acquisition activity including data formats; metadata; volume and number of files, etc.
	Data management plan	Should include plans for future preservation and data sharing.
	Funding application	Include Full Economic Cost (FEC) elements including activity relevant to preparation for preservation where applicable.
	Project implementation	Allows for ramping up and staff investment in project starting-up activity. The project must define an 'implementation period' over which the implementation effort and cost are estimated.

Figure 5: Sample Extract of Detailed Version of the Activity Model (Full version published in KRDS2, pp. 15-26)

2.3. KRDS Cost Drivers

The cost drivers are the different variables which can affect the overall costs of preservation. They are divided into two categories in KRDS: economic adjustments and service adjustments. A number of general considerations are also discussed including the idea of varying collection levels. These help frame discussion of the cost drivers in subsequent sections.

2.3.1. General Considerations

Collection Levels and Preservation Aims

Collection levels and preservation aims have a major overall influence on a number of key cost variables. It is very important to recognise that data collections vary substantially in terms of their anticipated user community and levels of use and therefore the associated preservation aims and costs. We suggest that HEIs consider using the collection levels of research, resource or community, and reference data collections proposed for long-lived data collections by the National Science Board (NSB) in the USA (NSB 2005). In brief these are:

- research data collections, which serve a limited group often the Principal Investigator and immediate participants in the research project;
- resource or community data collections, which serve a specific science or research community;
- reference data collections, which serve large segments of the general scientific and education community.

These collection levels provide indicators of likely number of users/user communities and levels of user support, periods of retention and preservation, and application of standards and quality control and validation of data and its accompanying metadata and documentation. These are significant cost factors so the collection levels and indicators for them may assist in identifying similar collections and cost estimation from “peer” collections with known cost data. Differences between collection levels can also be significant in terms of differences in service adjustments that may apply.

Note data collections may move up or down between these collection levels over time making it possible for a collection and its preservation aims and intended user community to change. Such changes may be infrequent but are likely to incur significant preservation upgrade costs. The reasons for this are:

(a) migrating from research data, where much of the knowledge required to interpret the data is in the form of tacit knowledge within the research group, to community or reference data requires that this knowledge is made explicit in user documentation and metadata describing the collection so that it is independently understandable to other researchers;

(b) A high degree of adherence is needed for resource/community and reference collections to: community standards for file formats; standards for metadata structure and content such as terminology from controlled vocabularies and ontologies; use of standards for encoding such as XML or RDF to make this metadata machine processable; thorough clearance of IPR and ethical consent for re-use; and validation and audit of these by the Archive to make them accessible and usable by others. This is not normally required or not required to the same degree of rigour for research collections.

The majority of data collections in HEIs are likely to be at the research collection level intended only for use of the project team and sometimes a very small number of external users. Retention periods and preservation requirements may be set by the funder's grant terms and conditions or by legal requirements (e.g. for clinical trials). Note preservation costs may be highest in the early years and become less significant over time. Preservation requirements are likely to be at a basic "secure storage" level for a set number of years with sufficient description to allow retrieval over that period.

However HEIs may also hold a number of data collections at resource/community or reference collection levels particularly if they host national or subject data centres. These collections will require significantly more investment in acquisition, ingest, and user support and these costs will be reflected in the service adjustments.

Controlling Future Costs

It is possible for institutions to control some of the complexity and unpredictability of future costs by limiting the future effect of some of the service adjustments. For example by taking action to regulate variables such as file formats during acquisition and ingest. This can be seen in the practice of a number of research data archives in the case studies (for an example see the preferred data formats in the ADS Guidelines for Depositors http://archaeologydataservice.ac.uk/advice/guidelinesForDepositors).

Timing

The timing of actions within the life-cycle has important implications for costs and is a significant dependency within the model. This is particularly true in relation to generating descriptive or preservation metadata and user documentation in the Pre-Archive phase rather than generating new/upgrading deficient metadata and documentation during ingest in the Archive phase. We provide examples such as that from Digitale Bewaring Project which estimated costs c. 333 euros for the creation of a batch of 1000 records in the pre-archive phase. In contrast once 10 years have passed and material has been transferred to an archive it may cost 10,000 euros to 'repair' a batch of 1000 records with badly created metadata (KRDS1, p.6). Similarly preservation action to address technology obsolescence may change from easily solvable and inexpensive while the technology is familiar and relevant staff and equipment are available, too expensive or even impossible once access to relevant staff and equipment are lost.

Cost Dependencies, Linkages and "Ripple Effects"

The above illustration of the effects of timing is one example of cost dependencies which exist and need to be captured within any model for preservation costs of research data. An implementation of a cost model should aim to capture ripple effects for costs from one function to another as variables change and allow "what if" scenarios to be constructed. Typical ripple effects are changes in volumes ingested on other archive functions, or changes in other archive activity on costs for support services such as software development and maintenance.

Sensitivities to Workload and Process Time Scheduling

Staff resources are not easily or quickly adjusted to changes in overall volume of deposits, or short-term fluctuations in workload particularly if the archive has little control over when the deposits will arrive or has fixed requirements for the speed with which they must be processed. Sensitivity will be greatest for inherently labour intensive, un-automated functions.

Evolution of Preservation Technology and Availability of Commercial off the Shelf (COTS) or mature Open Source Software/ and Community Standards and Best Practices ("First Mover Innovation")

Evolution of technology and the availability of COTS or mature open source software for use in different preservation functions and parts of the life-cycle will have significant effects on costs. Where preservation functions are evolving, a high-degree of R&D expenditure might be required in implementation phases. Similarly the pre-existence or development of community standards and best practices may have a major effect on preservation costs. These developments normally represent relatively small costs for most institutions individually but in aggregate can be considerable cumulative investments spread over many years and different institutions. Often they may be suitable for external funding and/or collaborative development. They are included as part of the "first mover innovation" function in the Activity Model.

Access

Access costs are potentially the most variable area of costs. It can simplify things to take a view of the archive where one can treat many of the access functions as being 'outside' the archive, since some of them are value-added services which could be removed and still leave a fully-functioning archive. This makes it easier to predict long-term costs. For an example of this see the ADS charging policy case study (see KRDS Costs Case Study 1).

2.3.2. Economic Adjustments

Economic adjustments consist of:

- inflation/deflation;
- depreciation;
- and the cost of return for financing and investment.

Procedures for applying inflation/deflation, depreciation, and cost of return for financing and investment and other adjustments will be available from Finance departments in institutions and the guidance in TRAC.

Inflation/Deflation

Inflation rates are typically agreed between the institution and funders and applied to cost categories such as staff. Deflation rates are typically applied to some equipment categories such as computer storage media with known long-term trends in price reduction.

Depreciation

There are several methods for calculating depreciation generally based on either the passage of time or the level of activity (or use) of the asset, which attribute the historical or purchase cost of an asset, across its useful life.

Cost of Return

The cost of return for financing and investment covers the cost of financing and generating a minimum level of retained surplus to permit rationalisation, updating and development. The TRAC methodology includes two cost adjustments for infrastructure costs and the return for financing and investment. The infrastructure cost adjustment is applied to the institution's approach to depreciation of major assets such as buildings to ensure they better reflect the full long-term costs of replacing that infrastructure.

2.3.3. Service Adjustments

We have selected below some of the key service cost drivers from KRDS that you are likely to need in implementing the cost framework. KRDS also identifies a number of additional service adjustments that may apply in specific circumstances and these are listed in Appendix A of the User Guide.

Staff Costs and Labour Rates

Staff costs should be recorded inclusive of salary, national insurance, and superannuation (pension) costs. Institutional rates and expectations will be available for pay progression and inflation costs. A mixture of different skill sets will be required for management, technical support, domain expertise, and administrative support: appropriate salary scales will be available from the institution.

Staff costs are likely to be the major cost in any preservation activity within an HEI. 70% or more of the costs of preservation services in the case studies relate to staff costs and historically these have always been seen as the major component of preservation costs.

Activity Duration

The duration of activities (year 1, year 2, etc) will need to be recorded so that costs and adjustments for inflation/deflation can be captured and modelled. It is significant to note that KRDS found annual preservation costs per unit preserved decline over time. We believe this reflects a number of factors: (a) the year-on-year decline in digital storage costs; (b) the effects of collections continuing to grow and adding economies of scale. Projections from the Archaeology Data Service (ADS) for preservation costs over 20 years based on experience and costs in its first 10 years of operation were as follows:

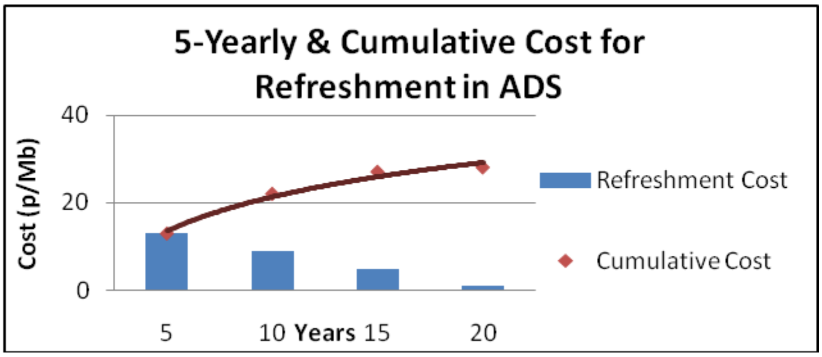


Figure 6: Costs for archival storage and preservation may decline to a minimal level over 20 years

Start-up and Operational Phasing of Activity

In addition to activity duration it is helpful to consider the phasing of the activity. The key difference between the cost profiles of these phases is that the former will emphasise the fixed costs of setting up the infrastructure/capacity of the repository system, while the latter will emphasise the variable costs of operating that capacity over time. Most of the up-front investment will necessarily occur in the start-up phase. Typically both research projects and data archives will have a start-up phase and operational phases in which the cost profiles will change over a period of months or years. The start-up phase is likely to reflect both the ramping-up of activities e.g. recruitment of staff and specific start-up activities e.g. developing new policies and procedures for the archive. The start-up costs particularly in terms of staff time can be substantial. The operational phase is likely to reflect increasing productivity and efficiency as procedures become established, tested and refined and the volume of users and deposits increases. In other sectors it has been suggested that operational services can show around 20% reduction in costs for each doubling of capacity due to this Experience Curve effect (Henderson 1974, Grant 2004).

Economies of Scale and Ingest Volumes

We identified the importance of economies of scale and the impact this has on unit costs for digital preservation. As an example, the University of London Computer Centre (ULCC) which runs the National Digital Archive of Datasets, provided us with costs for accession rates of 10 or 60 data collections: a 600% increase in accessions only increases costs by 325% as a result of economy of scale effects (KRDS1, p.6).

Data volumes need to be recorded. Typically these will be measured in Mb, Gb, Tb, or Pb volumes and the overall number of files. In general, higher volumes will lead to higher costs but the ratio of cost to volume is not a linear relationship as economies of scale and efficiency gains lower per unit costs. Some disciplines require petabyte stores. Institutions need to establish a policy that deals with both local demands of researchers together with a balancing of opportunity to effectively use shared national and subject repository services.

Levels of Automation

Given the overall impact and significance of staff costs, levels of automation (or conversely the levels of manual intervention required per dataset) are a significant variable for overall costs. The level of impact will be dependent on the economies of scale that can be achieved. In areas such as archive storage a high-level of automation e.g. robotic tape storage is widespread. In other areas such as ingest it will be most beneficial for high-volume accessions with relatively homogenous content.

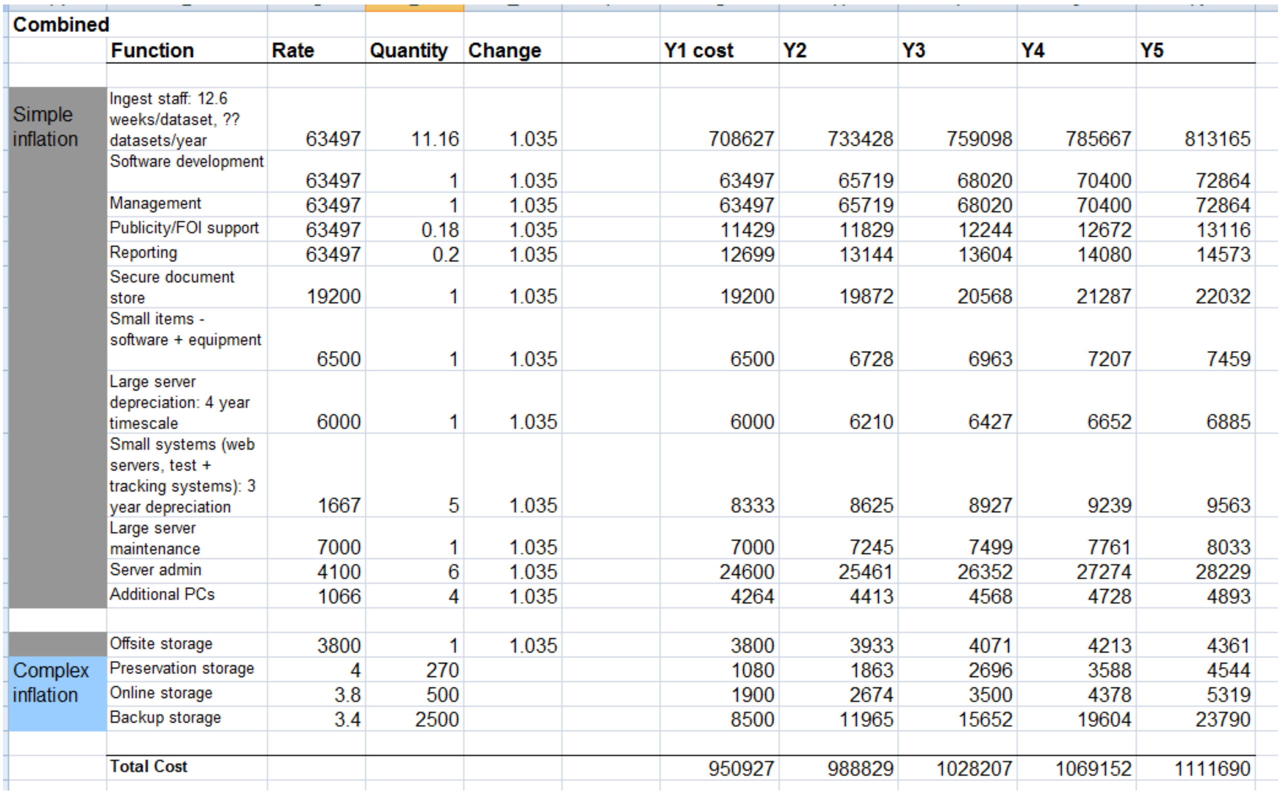
2.4. Resources Template

The resources ("resource pools") are derived from our activity model divisions of Pre-Archive, Archive, and Support Services, and TRAC cost categories (Joint Costing & Pricing Group 2005) with specific additions for archive charges and outsourcing for the requirements of the KRDS methodology. Cost categories taken from TRAC are Staff, Equipment, Travel, Consumables, Estate Costs, and indirect costs. In a full TRAC presentation staff costs would also be divided into direct or directly allocated costs, and economic adjustments (inflation/deflation, depreciation/infrastructure cost adjustment, cost of return for financing and investment) would be subsumed in calculations and applied as approved by the institution and funder to staff and other costs.

The resources template (Figure 7) provides a framework to draw together other elements of activity model and cost variables. The template presents categories of cost (e.g. staff) and duration (year 1, year 2, etc) in a simplified, generic form closer to that used in the cost methodologies of UK HEIs based on TRAC. It is a summary model as in practice the cost categories would be expanded to cover specific items e.g. individual members of staff and items of equipment, etc. Typically the cost model will implement these as a spreadsheet, populated with data and adjustments agreed by the institution. See the extract from the NDAD Cost Spreadsheet in Figure 8 that provides an example of this (The NDAD Cost Spreadsheet and guidance were made available as part of KRDS2 2010 supplementary materials).

Repeat Pre – Archive/Archive/Support Services or sub-components as required for purpose of costing	**Repeat duration (year 1, year 2, etc. as required for purpose of costing)**
Staff costs	
Equipment costs	
Travel	
Consumables	
Estate costs	
Indirect costs (where applicable)	
Outsourcing/ Archive Charges	

Figure 7: KRDS Resource Template to modify, expand, and implement to your requirements (first published in KRDS1, p. 47-8)



Combined

	Function	Rate	Quantity	Change		Y1 cost	Y2	Y3	Y4	Y5
Simple inflation	Ingest staff: 12.6 weeks/dataset, ?? datasets/year	63497	11.16	1.035		708627	733428	759098	785667	813165
	Software development	63497	1	1.035		63497	65719	68020	70400	72864
	Management	63497	1	1.035		63497	65719	68020	70400	72864
	Publicity/FOI support	63497	0.18	1.035		11429	11829	12244	12672	13116
	Reporting	63497	0.2	1.035		12699	13144	13604	14080	14573
	Secure document store	19200	1	1.035		19200	19872	20568	21287	22032
	Small items - software + equipment	6500	1	1.035		6500	6728	6963	7207	7459
	Large server depreciation: 4 year timescale	6000	1	1.035		6000	6210	6427	6652	6885
	Small systems (web servers, test + tracking systems): 3 year depreciation	1667	5	1.035		8333	8625	8927	9239	9563
	Large server maintenance	7000	1	1.035		7000	7245	7499	7761	8033
	Server admin	4100	6	1.035		24600	25461	26352	27274	28229
	Additional PCs	1066	4	1.035		4264	4413	4568	4728	4893
	Offsite storage	3800	1	1.035		3800	3933	4071	4213	4361
Complex inflation	Preservation storage	4	270			1080	1863	2696	3588	4544
	Online storage	3.8	500			1900	2674	3500	4378	5319
	Backup storage	3.4	2500			8500	11965	15652	19604	23790
	Total Cost					950927	988829	1028207	1069152	1111690

Figure 8: An Example of an Institutional Cost Spreadsheet showing functions, staff and other resource allocations over years 1-5 (from NDAD Costs Spreadsheet, published as KRDS2 2010 supplementary materials).

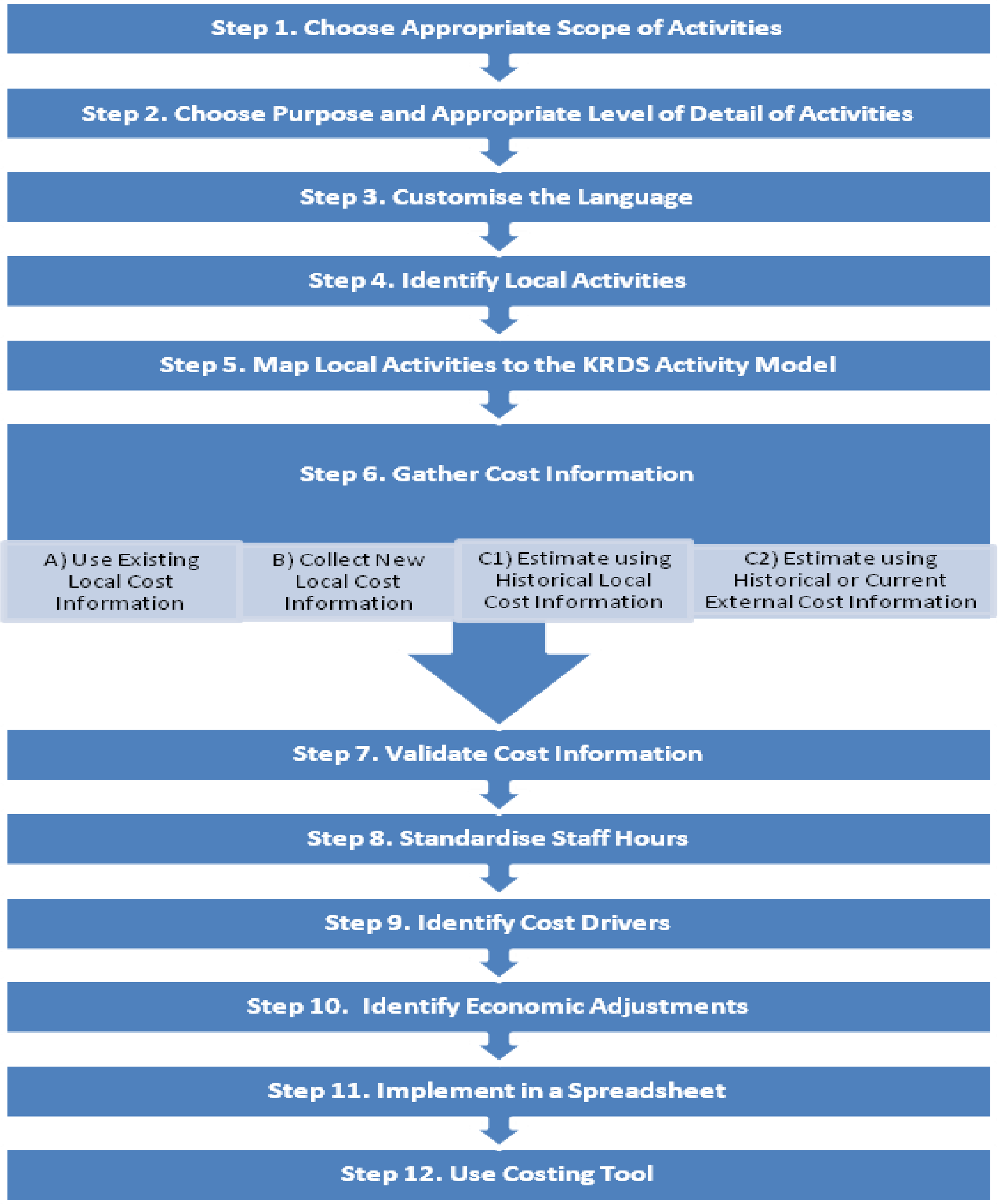
3. A Brief “How To” Guide For Life-Cycle Cost Analysis

A particularly powerful tool for assessing costs over long time spans is life-cycle costing. Life- cycle costing models a life-cycle for a specific process(es) and then identifies measurable component activities, cost drivers (variables that affect the costs of the activity e.g. volumes, formats etc), and resources (staff time, equipment etc) to provide an understanding of costs for that process. This considers the ownership as well as the acquisition costs of collections or physical assets over their -cycles from “cradle to grave”. It allows not just the assessment of the cost of creating or purchasing an asset but assessing what is needed for future maintenance. This makes it particularly appropriate for costing activities such as research data management and data curation. To undertake a life-cycle costing, there is a sequence of general considerations and steps that you need to follow as detailed below.

General Considerations: Before applying the KRDS cost framework to your institution take into consideration that:

- Dedicating a person to be responsible for collecting the cost information will save you effort and deliver results of better quality. The person should be responsible for checking the progress of the survey. Use someone who will be seen as independent and trusted by all staff – be aware of likely staff concerns over information related to individual performance, for the potential exaggeration in hours of work, or misunderstanding of activities and their definition;
- Overwhelming staff with information during the cost information collection procedure can have the opposite of desired results. Keep it simple and do not expect everyone to be on the same page from the beginning;
- Running an initial trial for a day or a week will give you an insight to the issues occurring when applying the model to your organisation and help the staff gain an understanding the Model and to query any points that are unclear;
- Asking staff to report separately on activities which are outside those in the Activity Model. Leave, sickness or absence should be specified separately. Allow for a general comments field to cover any other points of relevance.

Then implement the following steps:



Step 1. Choose Appropriate Scope of Activities

Step 2. Choose Purpose and Appropriate Level of Detail of Activities

Step 3. Customise the Language

Step 4. Identify Local Activities

Step 5. Map Local Activities to the KRDS Activity Model

Step 6. Gather Cost Information

A) Use Existing Local Cost Information

B) Collect New Local Cost Information

C1) Estimate using Historical Local Cost Information

C2) Estimate using Historical or Current External Cost Information

Step 7. Validate Cost Information

Step 8. Standardise Staff Hours

Step 9. Identify Cost Drivers

Step 10. Identify Economic Adjustments

Step 11. Implement in a Spreadsheet

Step 12. Use Costing Tool

Figure 9: Step by Step Guide to Making an Activity Based Cost Analysis

Step 1. Choose Appropriate Scope of Activities. Choose the appropriate scope of activities you wish to use from the KRDS activity model. The model allows for pre-archive, archive, and support services phase activities. Select phases and main activities from the model to reflect your requirements for cost or benefit analysis and arrangements for support or calculating indirect costs. The availability of pre-archive cost information will vary between organisations but it is particularly useful for some benefits case work such as assessing impact/value for depositors/creators/funders from archiving and re-use, or impact/value of changing timing of some activities such as metadata creation, etc. Support activities vary between organisations: in some it may be provided entirely in-house or via a parent organisation and generic budgets. Similarly some sectors and countries have established practice for calculating indirect support costs via agreed formulae.

Step 2. Choose Purpose and Appropriate Level of Detail of Activities. Choose the appropriate level of costing detail you need by choosing either the “Lite” or the “Detailed” version of the KRDS Activity model. A critical decision in a cost model’s design is the defining of activities at an appropriate level of detail. This is because the choice of activity level greatly affects the accuracy and cost of developing and maintaining the model. Just the high-level activities in the lite version of the model are usually sufficient for cost management and to understand the overall allocation of costs. This can be obtained with a much lower overhead in terms of capturing the required cost information. The detailed activity model provides options for more granular operations planning and process improvement as well as the necessary definitions and scope of the phases and activities you will need for reference.

Step 3. Customise the Language. Customise the language of the Activity Model according to the local needs: we have often re-used terms and definitions from the OAIS Reference Model (see KRDS1 activity model for annotated sources as published in KRDS 2008 p.36-46). OAIS terms will be capitalised in the scope notes in the detailed version of the KRDS Activity Model, e.g. Archival Information Package (AIP). For audiences unfamiliar with OAIS terminology, these may need further explanation or “translation” as appropriate for local use (for examples and further discussion see the cost case studies 2-8 in section 4).

Step 4. Identify Local Activities. Identify main relevant activities performed in your organisation. This can be carried out by interviewing employees on their daily activities and grouping those or by using existing sources of information (look at step 6a).

Step 5. Map Local Activities to the KRDS Activity Model. Map the activities performed in your organisation against the version of KRDS Activity Model selected in steps 3/4 in order to ensure full understanding of the Activity Model and avoid overlapping problems at a later stage. Thoroughly familiarise yourself with the detailed version of the Activity Model and use the definitions provided to assist with accurate mapping from local activities. Most mappings will be straight-forward but some may be more problematic and need careful translation and explanation. Not all KRDS activities may be undertaken in your organisation and some may need to be omitted. Do it with as much staff participation as possible and produce an agreed version of the Activity Model for your local implementation.

Step 6. Gather Cost Information. Use one of the methods mentioned below to gather information on activity costs. Accuracy of results depends on the chosen method:

(a) Use **Existing Local Cost Information**. Obtain available budget information and staff timesheets from existing sources.

(b) Collect **New Local Cost Information**. The most accurate and most costly procedure is the customised collection of new local cost information. In most cases, a collection procedure must be developed and support mechanisms (spreadsheets, timesheets, etc) for collection may need to be agreed. Collection of the information will need to be timely and skilled collectors may be required. One option may be **Local Sampling**. Ask staff to record the number of hours spent on defined activities over a certain period of time [a generic example for high-level activities in an Archive based on the “lite” version of the KRDS Activity Model is provided here]. When selecting the sample time period consider possible distortions arising from the phasing of the activities – particularly differences between start-up activities or projects, and established operational activities. The start-up phase emphasises the fixed costs of setting up the infrastructure/capacity of the repository system and “initial learning by doing”, while the operational one emphasises the costs of operating that capacity over time once routines and infrastructure are well established.

(c) **Estimation**. In the case where real local information cannot be obtained or information collection efforts cannot be financially justified, estimation may be possible from (i) **Historical Local Cost Information** if such information exists e.g. from previous costing attempts. We would not advise using this method though if the information is not detailed enough or are known to be significantly different from current experience; (ii) **Historical or Current External Cost Information**. Cost information from other comparable organisations or external sources may also be used for estimation. Again the information must be fit for purpose and its value for estimation will vary according to the close alignment of relevant activities and costs between your organisation and external sources.

Step 7. Validate Cost Information. Check for discrepancies in information (e.g. assignment of 5 hours per day to an automated activity). Apply modifications if needed.

Step 8. Standardise Staff Hours. Convert hourly activities to a proportion of “paid-hours”. For many professional staff regardless of the number of hours an individual works in a month their cost to the organisation is the same (unless it is not the case e.g. overtime is agreed). Note that staff costs should be recorded inclusive of salary, national insurance, and superannuation (pension) costs.

Step 9. Identify Cost Drivers. Investigate and acquire information related to cost drivers – the relevant KRDS variables termed service adjustments (see section 2.3.3 and Appendix A). This step can be performed in parallel with steps 1-8.

Step 10. Identify Economic Adjustments. Identify economic adjustments to all resources such as inflation/deflation, depreciation, and cost of return for financing and investment (see section 2.3.2).

Step 11. Implement in a Spreadsheet. Typically the activity model will help identify allocation of staffing required, the economic adjustments help spread and maintain these over time, and the service adjustments help identify and adjust resources to specific requirements. The resources template in KRDS1 provides a framework to draw these elements together (see section 2.4). Typically the cost model will implement these as a spreadsheet, populated with data and adjustments agreed by the institution. For implementations in the UK, there is also guidance (KRDS 2008, p. 21-22, 27) on conforming to the Transparent Approach to Costing (TRAC). See the NDAD costs spreadsheet as a UK example of an institutional cost spreadsheet. A number of other generic cost spreadsheets are also available on the Web and can be examined to see general layout and principles for constructing more complex spreadsheets for example the NASA Cost Estimation Tool, or the Cost Model for a Data Management Center.

Step 12. Use Costing Tool. In working through steps 1-12 you have developed a detailed local costing tool based on KRDS. Use the results for the initial purpose of costing. The exact application may depend on the purpose of the costing, which might include:

- identifying current costs;
- identifying former or future costs;
- or comparing costs across different collections and institutions which have used different variables.

These are progressively more difficult. The model may also be used to develop a charging policy or appropriate archiving costs to be charged to projects. It can also be used to contribute to cost/benefit or value chain analysis.

4. KRDS Benefits Analysis

4.1. Introduction

Analysis of the costs of preserving research data sets is not enough to assess the economic feasibility of a particular digital preservation activity. Cost analysis should be accompanied by a framing of the benefits from preservation – in other words, the value that is anticipated to emerge from the investment in maintaining the long-run existence and accessibility of research data. The benefits conferred from investment in digital preservation often are either assumed to be common knowledge, or are expressed in terms far too generic to be of practical use for decision-making purposes (e.g., "preserving society's digital record for future generations", etc.).

Unfortunately, measuring benefits is often quite challenging, especially when these benefits do not easily lend themselves to expression in quantitative terms. Part of the reason why characterising the benefits from digital preservation activities has been neglected is no doubt a consequence of the difficulty of the task.

Despite the challenges, it is still useful to think carefully about the nature of the benefits an investment in digital preservation is expected to bring. As a first step in this process, KRDS has created a Benefits Analysis Toolkit framing a few important dimensions that illuminate the broad contours of the benefits digital preservation investments potentially generate. These dimensions serve as a high-level framework within which thinking about preservation benefits can be organised and then sharpened into more focused value propositions.

The development of the Toolkit has been funded by JISC as part of the "KRDS/I2S2 Digital Preservation Benefit Analysis Tools" Project, which has tested, reviewed and developed further the Keeping Research Data Safe (KRDS) Benefits Framework and the KRDS/I2S2 Value Chain and Benefit Impact Analysis tools. The current Toolkit therefore consists of two tools: the KRDS Benefits Framework (version 3.0 July 2011); and the Value-chain and Benefits Impact tool (version 2.0 July 2011). Each tool contains a more detailed guide and worksheet(s). Both provide a series of common examples of generic benefits which have been revealed by their application to frequently arise from the curation/preservation of research data. Users are encouraged to modify or add to these as required.

4.2. The KRDS Benefits Framework Tool

The KRDS Benefits Framework (Tool 1) requires less experience and effort to implement and can be used as a stand-alone tool in many tasks. It can also be the starting point and provide input to the use of the Value-chain and Impact analysis (Tool 2). It is a tool for identifying, assessing, and communicating the benefits from investing resources in the curation/long-term preservation of research data and is especially useful in supporting and organizing early-stage brainstorming on the benefits associated with a particular activity. Once potential benefits have been identified it can also assist in articulating them to a broad audience of stakeholders and in customising their expression to address different stakeholder audiences.

The Framework organises benefits along three broad dimensions: the outcome achieved; when the outcome is achieved; and who benefits from the outcome. Each of these dimensions can be subdivided into two categories: direct and indirect benefits, near-term and long-term benefits and internal and external benefits respectively. This is summarised graphically in Figure 10 below. These dimensions are applicable to nearly all research data curation/preservation activities and they can be used without modification in most contexts. Any benefit associated with a data curation/preservation activity can be characterised according to these three dimensions. Each Dimension provides a different but complementary view of potential benefits so individual benefits may be repeated in different dimensions or generate alternate “mirror images” of the benefits when considered from a different perspective.

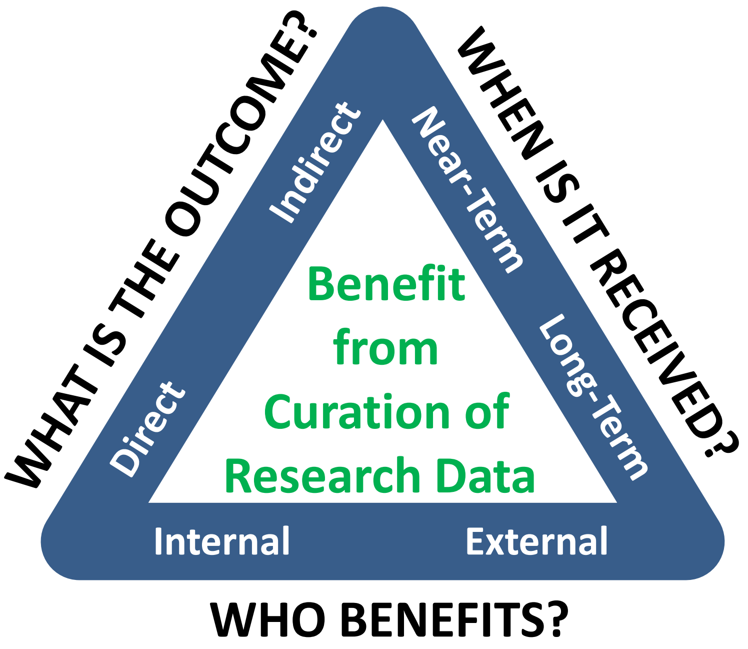


Figure 10: The anatomy of a benefit

Application of the Benefits Framework to a range of projects over the course of its development has revealed a number of common benefits that frequently arise from preservation of research data. Often, these can be simply expressed in a generic form independent of project specifics. Figure 11 below shows the list of examples of these "generic" benefits which is provided as the starting point for applying the framework. *Please note this is not a comprehensive list of potential benefits from preserving research data.*

Examples of Common Benefits

New research opportunities

Input for future research

Motivating new research

New research funding

Increasing research productivity

Stimulating new networks/collaborations

Knowledge transfer to other sectors

Knowledge transfer to industry

No re-creation of data

No loss of future research opportunities

Secures value to future researchers & students

Protecting returns on earlier investments

Lower future preservation costs

Planned management from an early stage in the research life-cycle is ultimately more cost-effective than late intervention (providing proper selection of what to keep is done)

Commercialising research

Increasing skills base of researchers/students/staff

Increasing economic growth

Catalysing new companies and high skills employment

Verification of research/research integrity

Fulfilling organisational mandate(s)

Fulfil research grant obligations

Value to current researcher & students

No data lost from Post Doc turnover

Secure storage for data intensive research

Availability of data underpinning published findings

Re-purposing data for new audiences

Use by new audiences

Re-purposing methodologies

Enhancement of research tools and software by testing on a range of well-curated datasets

Scholarly communication/access to data

Long-term re-use of well curated data

Short-term re-use of well curated data

Adds value over time as collection grows and develops critical mass

Increased visibility/citation

Figure 11: A list of examples of generic benefits

Figure 12 below displays some of the benefits selected and described in detail from real-world situations with the aid of the Framework. It demonstrates how, as a next step in characterizing benefits, the simply expressed common benefits given in the list of examples can be selected and extended into more detailed descriptions specific to particular projects if required.

Direct Benefits

[*New research opportunities*]. A direct benefit from continued access to data at the UKDA is the ability of researchers to use data which they did not create themselves. Survey data collected by government agencies in the UK may never have been accessible to the research (and/or wider) communities had it not been for their preservation at the UKDA. The re-use of government data, especially of the major surveys (e.g., British Social Attitudes Survey), has propelled research across a wide range of disciplines. (KRDS2, pp. 70)

Indirect Benefits (Costs Avoided)

[*Lower future preservation costs*]. The Digitale Bewaring Project in the Netherlands, which focused on government electronic records, estimated that the creation of a batch of 1,000 appropriately-documented records during the Pre-Archive phase would cost approximately 333 euros. Conversely, once 10 years have elapsed 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

[*No data lost from Post-doc turnover*].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 near- and medium-term benefits. (KRDS2, p.60)

Long-Term Benefits

[*Adds value over time as collection grows and develops critical mass*]. 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

[*Increased visibility/citation*]. 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

[*Catalysing new companies and high skills employment*]. 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)

Figure 12: Examples for each of the main dimension and sub-divisions of the KRDS Benefits Framework

The structure of the Framework can be customised and extended as needed, given the circumstances of specific projects or institutions.

For example, Dimension 3 (Internal/External) in the Framework could be further sub-divided by more specific groups of stakeholders if desired. An illustration of this, populated with some examples of common benefits, is provided in Figure 13 below.

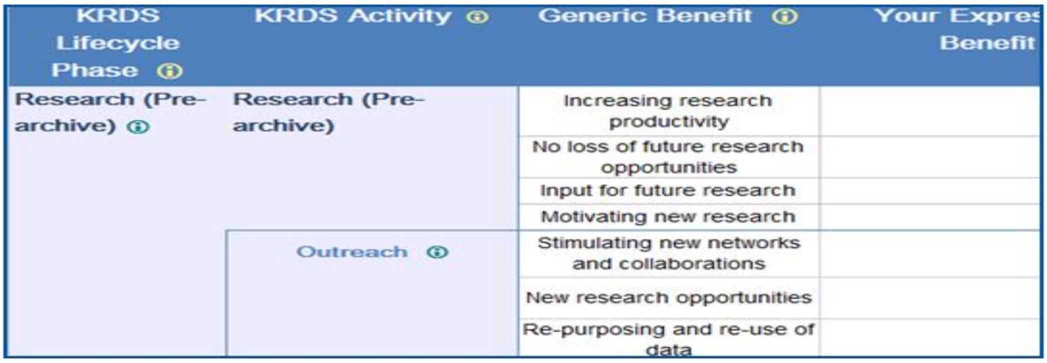
Dimension 3 (Who Benefits) Sub-divided by a University's Stakeholders					
Internal Benefits			**External Benefits**		
Researcher	**Research Group**	**Institution**	**Research Funder**	**Discipline**	**Others** (e.g. NHS, etc)
Increased visibility/ citation	No data lost from Post Doc turnover	Fulfilling organisational mandate(s)	Increasing research productivity	Scholarly communication /access to data	Knowledge transfer to other sectors

Figure 13: Example of Dimension 3 of the KRDS Benefits Framework expanded for a selection of a University's Stakeholders

4.3. The Value Chain And Benefits Impact Tool

The Value Chain and Benefits Impact Tool is the more advanced tool in the Toolkit and requires more experience and effort to implement. Once benefits have been identified and organised within the Benefits Framework, further work can proceed aimed at identifying potential measures or illustrations of the value and impact of those benefits. This next stage is supported by the Value-Chain and Benefits Impact Tool, which can be used in assessing where value is added to outputs in a chain of activities and for use in evaluation, strategic and organisational planning, and reporting. It is likely to be most useful in a smaller sub-set of longer-term and intensive activities.

The Tool consists of a detailed user guide and two worksheets; the Benefits Impact worksheet and the Value-chain and Benefits Impact worksheet. An extract from the later is shown in Figure 14 below. To use this Tool, the worksheet should be selected that most closely matches your needs. Both worksheets have been pre-populated with the selection of common generic benefits also used in the Benefits Framework Tool but you may review, delete or add more to the selection. The tool has been designed to be generic but easily configurable by the user for their specific needs or application.



KRDS Lifecycle Phase ⓘ	KRDS Activity ⓘ	Generic Benefit ⓘ	Your Expres Benefit
Research (Pre-archive) ⓘ	Research (Pre-archive)	Increasing research productivity	
		No loss of future research opportunities	
		Input for future research	
		Motivating new research	
	Outreach ⓘ	Stimulating new networks and collaborations	
		New research opportunities	
		Re-purposing and re-use of data	

Figure 14: An extract from the Value-chain and Benefits Impact worksheet

The Tool is intended help you to identify quantitative metrics and qualitative indicators for the impact of benefits and optionally to support a value-chain analysis. It uses the KRDS Activity Model as a starting point for the value-chain analysis, so it is better suited to the specific needs of research data and its curation/preservation.

It is recommended that both worksheets in the Tool are used by a team with a senior member of staff or independent support (e.g. consultancy). For maximum effectiveness in applying the Tool, ideally at least one person in the team should be very familiar with the KRDS Benefits Framework (tool 1), other KRDS Outputs such as the KRDS Activity Model, and similar assessments of value and impact.

Demonstrating the impact of benefits for research data curation/preservation, either directly via metrics (quantitative impacts) or qualitatively via illustration in case studies (qualitative impacts), is still a relatively novel area. The guide provides discussion and further references to JISC and Research Councils' work on demonstrating impact. This can provide examples to assist working though how to demonstrate the impact of benefits and implement capturing the relevant measures/illustrations identified in completing the worksheets.

Guides for each tool and case studies of completed examples of the worksheets (see the Benefits Analysis Tools Project web site) provide documentation and support for your implementation.

The Toolkit is available to download from the project web site (http://beagrie.com/krds-i2s2.php) and the KRDS web site (http://www.beagrie.com/krds.php).

5. KRDS Case Studies

5.1. KRDS1 and KRDS2 Case Studies

Below is a table listing all case studies carried out in the KRDS project and their main features. They have helped to develop and validate in real world circumstances the approaches to costing and benefits for the preservation of research data proposed in our costs and benefits frameworks. The KRDS cost case studies also provide detailed illustrations and descriptions of issues and costs relevant to KRDS. It is intended that the generic expressions of preservation benefits in the KRDS Benefits Framework should be sharpened into more focused value propositions in local implementations. Two benefit case studies were developed in KRDS2 to illustrate local implementations and expansions of the approach. The first of these was a Benefits Case Study for the UK Crystallography Service at Southampton University; and the second was a Benefits Case Study for the UKDA.

Case Study	**Type of Institution/Data**
Costs Case Study 1: Archaeology Data Service Charging Policy Download Cost Case Study 1:	National Data Centre Archaeology Developing a Charging Policy
Costs Case Study 2: Archaeology Data Service Download Cost Case Study 2	National Data Centre Archaeology Collection Preservation Costs
Costs Case Study 3: University of Cambridge Download Cost Case Study 3	University Data Repositories Chemistry, Social Anthropology, University Library, Digital Images, Digital Preservation Costs
Costs Case Study 4: King's College London Download Cost Case Study 4	University Data Repository National Data Centre Arts and Humanities Digital Preservation Costs
Costs Case Study 5: University of Southampton Download Cost Case Study 5	National Data Centre University Data Repositories Chemistry, Oceanography

Case Study	Type of Institution/Data
	Digital Preservation Costs
Costs Case Study 6: University of Oxford Download Cost Case Study 6	University Data Repositories Range of disciplines Curation, preservation or storage costs
Costs Case Study 7: National Digital Archive of Datasets, ULCC Download Cost Case Study 7	National Data Centre Government Datasets Costing Preservation Services Third-party Outsourcing
Costs Case Study 8: UK Data Archive Download Cost Case Study 8	National Data Centre Social Sciences and History Digital Preservation Costs
Benefits Case Study 1: National Crystallography Data Service, Southampton University Download Benefits Study 1	National Data Centre Chemistry KRDS Benefits Framework
Benefits Case Study 2: UK Data Archive Download Benefits Study 2	National Data Centre Social Sciences and History KRDS Benefits Framework

Figure 15: Brief Summary overview of the KRDS Case studies

5.2. Subsequent Implementations

In addition to the case studies in KRDS1 and KRDS2 there have been a number of implementations and case studies building upon KRDS that may be of interest to you.

Case Study	Type of Institution/Data
JISC Research Data Management Infrastructure (RDMI) projects Benefits case studies Download Case Studies	Subject Data Centres University Data Repositories Research Data Benefits
I2S2 Benefits Case Studies Download Case Studies	Integrated National/Local Services Research Data Benefits

Case Study	Type of Institution/Data
Dryad Repository Download Case Study	Subject Repository Evolutionary Biology and Ecology Research Data Benefits Data Curation Costs

Figure 16: Subsequent Implementations of the KRDS Case studies

6. KRDS Costs Survey

One of the core aims of KRDS2 was to identify potential sources of cost information for preservation of digital research data and to conduct a survey of them. 13 survey responses were received: 11 of these were from UK-based collections, and 2 were from mainland Europe. Data in this survey may be useful for anyone investigating cost drivers and cost data for digital preservation. This section provides a short overview of the results. Individual copies of completed cost data survey forms can be downloaded from http://www.beagrie.com/jisc.php.

Collection	Repository Type		Cost Information						
	Research	Cultural Heritage	Pre-archive	Archive	Access	Support Services	Estates	Dates	Accessible?
UK Collections									
ADS	●		●	●	●	●		2004 - Present	Possibly
BADC	●		●	●	●	●	●	2001 - 2008	Possibly
eCrystals	●		●	●				2002 - 2009	Possibly
EDINA	●	●	●	●		●	●	2006 - Present	Possibly
Linnean Soc	●		●	●	●		●	2007 - Present	Possibly
NDAD		●	●	●	●		●	1997 - Present	Possibly
NLW		●	●	●		●		2007 - Present	Yes
Oxford	●		●	●		●	●	2007 - 2009	Possibly
Rutherford	●	●	●	●	●		●		Possibly
UKDA	●			●	●	●		2009	Possibly
VADS	●			●		●	●	2008	Possibly
International Collections									
BABS	●	●	●	●					No
DANS	●		●	●	●	●	●	2008	Possibly

Figure 17: Summary of KRDS2 Data Survey Responses

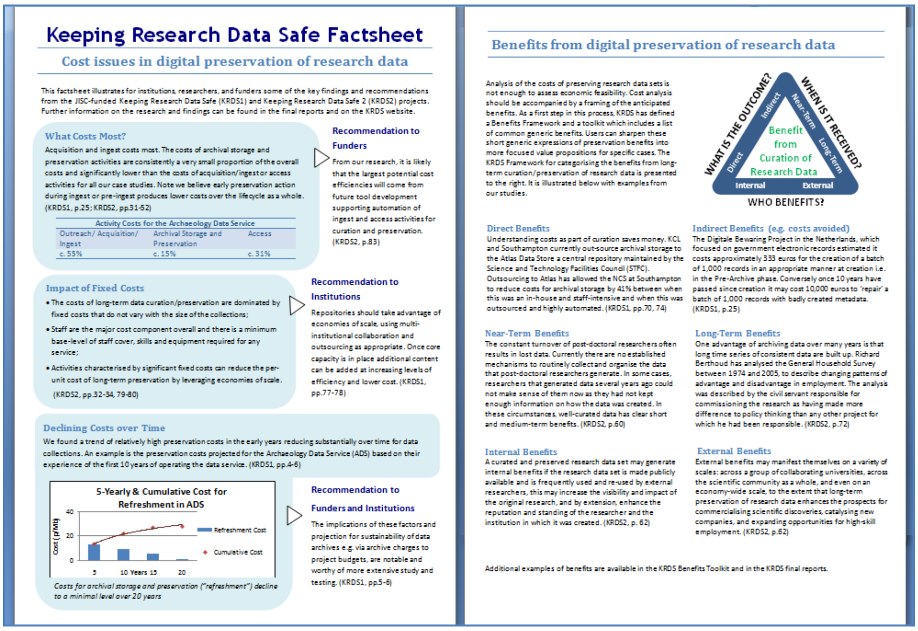
Abbreviations: ADS (Archaeology Data Service, University of York), BADC (British Atmospheric Data Centre), eCrystals (National Crystallography Service, University of Southampton), EDINA (UK Borders Service, EDINA, University of Edinburgh), Linnean Soc (Linnean Society Collection, University of London Computer Centre), NDAD (National Digital Archive of Datasets, University of London Computer Centre), NLW (Welsh Journals Online, National Library of Wales), Oxford (University of Oxford), Rutherford (Rutherford Appleton Laboratory, Science and Technology Facilities Council), UKDA (UK Data Archive, University of Essex), VADS (Visual Arts Data Service, University College for the Creative Arts), BABS (Bibliothekarisches Archivierungs- und Bereitstellungssystem -The Library Archiving and Access System- Bavarian State Library, Germany), DANS (Data Archiving and Networked Services, The Netherlands).

7. KRDS Factsheet

The KRDS Factsheet illustrates in summary form 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. It may be particularly valuable in advocacy and outreach activities. Version two of the Factsheet is available for download as a PDF from http://www.beagrie.com/KRDS_Factsheet_0910.pdf

The A4 four-page factsheet is intended to be suitable for senior managers and others interested in a concise summary of our key findings. It will be relevant to all repositories and institutions holding digital material but of particular interest to anyone responsible for or involved in the long-term management of research data. The factsheet covers the following major areas:

- Cost issues in digital preservation (what costs most, impact of fixed costs, declining costs over time);
- Benefits from digital preservation (benefits taxonomy, direct benefits, indirect benefits, near-term benefits, long-term benefits);
- Institutional issues (repository models and structures, key cost variables, data collection levels).



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.

Figure 18: Illustration of pages 1-2 of the KRDS Factsheet

8. Future Development of KRDS

Defining costs and benefits for research data across a wide-range of disciplines and institutions is a demanding and complex task. KRDS has developed from relatively small-scale incremental projects and we recognise that there are still significant areas for future work. The KRDS2 final report outlined a number of key recommendations for future development, some of which have been progressed further subsequently as noted below:

- Consider further work on identifying and quantifying the benefits of research data preservation [In 2011 JISC funded the development of the I2S2/KRDS Benefits Analysis Toolkit. Elements of the KRDS Benefits Framework were also tested in the JISC Managing Research Data Programme];

- Examine further development of the pre-archive phase of the KRDS2 activity model and produce versions of the model from a researcher's perspective [In 2010 the I2S2 project funded by JISC produced the I2S2 Research Data Lifecycle to do this];

- Seek to implement KRDS2 in cost spreadsheets and continue research on implementation variables and metrics that could enhance them [A number of local implementations including the Dryad Repository- still more potentially to do in this area];

- Raise awareness of KRDS internationally; extend the costs survey; and develop research partnerships on digital preservation costs [The partners continue to promote KRDS internationally and seek research and consultancy partnerships];

- Develop presentation of KRDS as a tool with elements such as guidance notes updated and packaged alongside components such as the activity models and future potential elements such as cost spreadsheets [Ongoing with a website dedicated to KRDS].

These provide a roadmap for future development, which we continue to pursue when opportunity and funding allows. News on future developments will be posted to the blog (you can view/subscribe to the blog at www.beagrie.com).

Section 8: The Future Development of KRDS

8.1. User Feedback and Comments

As KRDS develops further, it is intended to update the User Guide and Factsheet and issue future editions. As such we would welcome feedback and comments from users of this version of the Guide. Feedback can be sent to info@beagrie.com.

9. Further Information and References

Consultative Committee for Space Data Systems (CCSDS), 2002, *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, Blue Book, 2002, (ISO14721:2003). http://public.ccsds.org/publications/archive/650x0b1.pdf

Grant, R. M., 2004, *Contemporary strategy analysis*, (Blackwell Publishing).

Henderson, B. The Experience Curve Reviewed, *Perpectives* No.124 Reprint (The Boston Consulting Group). Retrieved 20/1/08 from http://www.bcg.com/publications/files/experience_curve_I_the_concept_1973.pdf

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.doc

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

KRDS2 2010 supplementary materials: http://www.beagrie.com/jisc.php

KRDS website: http://www.beagrie.com/krds.php

LIFE Project website: http://www.life.ac.uk

National Science Board (NSB), 2005, *Long-lived Digital Data Collections: Enabling Research and Education in the 21st century* September 2005 (National Science Foundation). Retrieved 10/12/07 from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

NASA Cost Estimation Tool website: http://opensource.gsfc.nasa.gov/projects/CET/index.php

Transparent Approach to Costing (TRAC) website: http://www.hefce.ac.uk/finance/fundinghe/trac/

Appendix A – Other Potential Service Adjustments

We have selected and discussed in section 2.3.3 above some of the key service cost drivers from KRDS. KRDS also identifies a number of additional service adjustments that may apply in specific circumstances and these are listed below.

Acquisition, Disposal and Ingest

Number of Depositors

The number of different individual and institutional depositors the archive needs to liaise with will affect acquisition and other archive costs. This is particularly true if different working practices require individual negotiation on deposit terms and bespoke transfer mechanisms to be created.

Number, Mode and Frequency of Deposits

The overall number of deposits needs to be recorded. The frequency of individual deposits (one-off deposit, incremental small deposits over time, etc), and the mode of deposit (automated transfer over the network, via couriered storage media, etc), also affect requirements and therefore costs.

Number, Complexity and Type of File Formats

The number, complexity and type of file formats needs to be considered. Dealing with a small number of widely understood file formats allows for simpler procedures at the time of acquisition and future migration. Each additional format imposes a one-off cost to develop procedures to deal with it. The complexity and type of file formats have similar issues.

Metadata, Documentation, Ethics and IPR

The quality of descriptive or preservation metadata and documentation, and the thoroughness of ethics and IPR clearance have a substantial impact on the potential re-use and value of research data to other researchers. As noted above, timing of these actions in the Pre-Archive phase substantially lowers costs. If any of these issues need to be rectified by the Archive, costs will be substantially higher, and in some cases information may not be recoverable and the value of the data for research significantly degraded.

Levels of Processing, Validation and Calibration

Levels of processing, validation and calibration that need to be undertaken will affect costs. As noted above under collection levels and preservation aims, this may partly be related to data collection levels and the degree and rigour of conformance to standards and overall quality of data required.

De-accessioning Costs

De-accessioning will involve the time of specialist staff for review. Although cost savings may be achieved on archive storage this will need to be assessed and balanced against staff costs for the review. It is worth noting a number of our interviewees and sources suggest the majority of cost for preservation of research data lies in acquisition and ingest rather than in longer-term archive storage and preservation and that given the greatest costs are in acquisition it will often only be worth considering de-accessioning in very few cases on cost grounds.

ARCHIVE STORAGE, PRESERVATION PLANNING, DATA MANAGEMENT

Retention Period

The retention period will impact upon costs. The longer data is retained and therefore require more preservation actions over time to ensure integrity and accessibility, the higher will be the total cost over time. Retention period can be linked to collection levels and preservation aims and legal or grant term conditions as noted above. Consideration should be given by projects at the earliest possible stage as to what data needs to be retained during and beyond the life of the project and how this will be achieved. Costs will be higher for data that needs expert review at the end of the retention period to determine whether it should be disposed compared to data whose deletion/de-accessioning is straight-forward (see de-accessioning above).

Management and Refreshment

The management of data within the archive needs to take account of storage management policies, operational statistics, or directions from the Ingest stages. Cost will be affected by any special levels of service, or any special security / protection measures that are required. These include on-line, off-line or near-line storage, required throughput rate, maximum allowed bit error rate, or special handling or backup procedures. Monitoring is needed to ensure that no corruption of data occurs during transfers.

The size and complexity of the archive will impact both the necessity and the cost of providing operational statistics summarizing the inventory of media on-hand, available storage capacity in the various tiers of the storage hierarchy, and usage statistics.

Data refresh is tied into the archives migration strategy to new systems and storage media. The decisions impacting on costs include policy on frequency of hardware replacement, and the nature of the material in the archive taking into account dependencies.

Number of Versions and Copies

The preservation strategy is likely to include multiple copies of the data including an off-site copy. In some disciplines it will also be common to have multiple versions or editions. The number of versions and copies affects archive storage and management costs.

Storage Media (capacity, costs)

Storage media will be selected on the basis of service requirements e.g. data volumes, required speed of access, or archival properties, and cost. The selection of storage media will influence the frequency of future storage media migration and staff and equipment needed for this task. It is important to remember that the total cost of ownership of archive storage media and systems is substantially higher than the purchase cost alone. Research suggests that the initial capital costs are less than a third of the total costs of ownership.

Archive media monitoring

All storage media need to be monitored for signs of data loss. The sample and frequency with which this is done will influence costs. This will be a more significant cost for storage media requiring manual intervention and inspection compared to automated systems.

Access

Number of Users and User Communities

The size, knowledge base, and number of individual users and user communities will have particular influence on costs and are a significant additional factor in costs incurred by community and reference level data collections. The broader the range of researchers supported the higher the investment will be in user support. Typically large community and reference data collections will involve staff with subject knowledge of the discipline(s) to support designated user communities.

Standard or Custom Interfaces

Systems and/or application interfaces are expensive to develop and then maintain. There are substantial economies from maintaining a small number of standard interfaces and a proportionately high cost to each custom interface the archive needs to develop.

Level of User Support

The demands on user support increase with the volume of users, number of user communities, proliferation of data types, data sources, and user tools. It will be important to define the levels of support at the onset as this has a direct bearing on costs and therefore can impact on the archives policies regarding supported formats etc. The capacity will increase as more automated user support aids become available (beginning with on-line documentation, FAQ, etc.). User support may also include variable potential levels of outreach, education, and training workshops for users.

Access Control

Requirement for access control will add costs on a sliding scale depending on the level of control and methods required. Simple closure of a data collection for a specified time period before access to users is relatively trivial to automate in existing systems. In contrast anything more staff intensive such as manually checking and removing personal information in an access copy can involve a significant cost.

Number and Volume of Accesses

Resources to support access in terms of equipment and staff will be affected by the number and volume of accesses and how these accesses are spread over time and different items/collections in the archive.

Access/Distribution Method

The profile of costs will be affected by the access and distribution method. If access is over a network and largely client lead the cost profile will be very different to ad hoc requests handled directly by staff and supplied offline.

Service Response Times

Users increasingly expect high-speed access to be an inherent part of online systems. Maintaining and configuring access services to consistently meet these expectations will incur higher costs particularly for large volumes of users and accesses.

Processed Products

In some disciplines processing of raw data and the production of value-added editions with standardisation and validation is an essential component of an archive's work. Similarly data may need to be packaged and interpreted for specific user groups e.g. in education. This is labour-intensive and requires appropriately trained staff.

ACKNOWLEDGEMENTS

Keeping Research Data Safe (KRDS) has been funded by JISC. It has been developed over three research projects (KRDS1, KRDS2 and the I2S2/KRDS Benefits Analysis Toolkit project) conducted by a partnership led by Charles Beagrie Limited, involving OCLC Research, UKOLN/DCC, the UK Data Archive, the Archaeology Data Service, the University of London Computer Centre, and the universities of Cambridge, King's College London, Oxford, Southampton and UCL. We would like to thank all partners and others who contributed in various ways to KRDS1, KRDS2 and the Benefits Analysis Toolkit.

In addition, for this User Guide specific thanks are due to Neil Beagrie, Anna Palaiologk, Daphne Charles, Rachel Beagrie, and Rob Beagrie for text, illustrations and edits (Charles Beagrie Ltd); to Brian Lavoie for figures 1-3 and 10; and to our reviewers of various drafts Neil Grindley and Simon Hodson (JISC), Matthew Woollard (UKDA), Brian Lavoie (OCLC), Julian Richards (ADS), Elin Stangeland and Grant Young (University of Cambridge), and Julia Chruszcz (Charles Beagrie Ltd).