KEEPING RESEARCH DATA

SAFE 2

Neil Beagrie, Brian Lavoie and Matthew Woollard

with contributions by the Universities of Cambridge, Oxford, and Southampton, the Archaeology Data Service, OCLC Research, UK Data Archive, and University of London Computer Centre.

Final Report - April 2010

Prepared by:

Charles Beagrie Limited

www.beagrie.com

A study funded by

JISC

With support from OCLC Research and the UK Data Archive

PREFACE

The first Keeping Research Data Safe study funded by JISC made a major contribution to understanding of long-term preservation costs for research data by developing a cost model and indentifying cost variables for preserving research data in UK universities (Beagrie et al, 2008). However it was completed over a very constrained timescale of four months with little opportunity to follow up other major issues or sources of preservation cost information it identified. It noted that digital preservation costs are notoriously difficult to address in part because of the absence of good case studies and longitudinal information for digital preservation costs or cost variables.

In January 2009 JISC issued an ITT for a study on the identification of long-lived digital datasets for the purposes of cost analysis. The aim of this work was to provide a larger body of material and evidence against which existing and future data preservation cost modelling exercises could be tested and validated.

The proposal for the KRDS2 study was submitted in response by a consortium consisting of 4 partners involved in the original Keeping Research Data Safe study (Universities of Cambridge and Southampton, Charles Beagrie Ltd, and OCLC Research) and 4 new partners with significant data collections and interests in preservation costs (Archaeology Data Service, University of London Computer Centre, University of Oxford, and the UK Data Archive).

A range of supplementary materials in support of this main report have been made available on the KRDS2 project website at http://www.beagrie.com/jisc.php. That website will be maintained and continuously updated with future work as a resource for KRDS users.

ACKNOWLEDGEMENTS

The authors would like to thank all contributors to the cost survey and additional contributors to our case studies (Kevin Ashley, Simon Coles, Catherine Hardman, Luis Martinez Uribe), and other colleagues for peer review, feedback on and contributions to the draft report including: Robert Beagrie, Julia Chruszcz, Neil Grindley, Simon Hodson, Hervé L'Hours, Rob Read, Elin Stangeland, Mark Thorley, and Grant Young.

CONTENTS

1. EXECUTIVE SUMMARY

Data has always been fundamental to many areas of research but in recent years it has become central to more disciplines and inter-disciplinary projects and grown substantially in scale and complexity. There is increasing awareness of its strategic importance as a resource in addressing modern global challenges and the possibilities being unlocked by rapid technological advances and their application in research (NAS 2009).

The first Keeping Research Data Safe study funded by JISC made a major contribution to understanding of long-term preservation costs for research data by developing a cost model and indentifying cost variables for preserving research data in UK universities (Beagrie et al, 2008). The Keeping Research Data Safe 2 (KRDS2) project has built on this work and delivered the following:

- A survey of cost information for digital preservation, collating and making available 13 survey responses for different cost datasets;
- The KRDS activity model has been reviewed and its presentation and usability enhanced;
- Cost information for four organisations (the Archaeology Data Service; National Digital Archive of Datasets; UK Data Archive; and University of Oxford) has been analysed in depth and presented in case studies;
- A benefits framework has been produced and illustrated with two benefit case studies from the National Crystallography Service at Southampton University and the UK Data Archive at the University of Essex.

Our main findings are presented in full in the Conclusions (section 9). Some examples of our key findings are:

Long-term Costs of Digital Preservation for Research Data:

- The costs of archiving activities (archival storage and preservation planning and actions) 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 in KRDS2. This confirms and supports a preliminary finding in KRDS1.

Benefits of Preserving Research Data:

- We have recognised that the identification and promotion of the "near term benefits" are particularly important in advocacy to researchers: we can show in our benefit case studies and also our costs work at Oxford that there are significant benefits in the short-term to current researchers as well as long-term benefits to future research.

Our Survey and Sources of Information for Costs:

- 11 responses were received from the UK and two from mainland Europe. Unfortunately a further two offered from the USA could not be available within the deadline for publication of KRDS2. Cost information from respondents is available for most of the KRDS2 main activity phases (pre-archive, archive, access, support services, and estates), although the depth and breadth of information available from different collections varies considerably (see section 6 for individual responses).

Application of the KRDS Activity Model:

- The KRDS activity model has been reviewed by partner institutions and found to be broadly robust and fit for purpose: some small changes have been made to the sub-activities as part of KRDS2 (see section 4) and guidance on its application extended;
- We have recognised that the activity cost models should be applied at different levels of detail for different purposes: as a result KRDS2 now caters for potential dual application of the activity model with two versions presented at different levels of detail (see sections 5.2, 5.3, and 5.4).

Our work has confirmed the strengths of the approaches underlying the original Keeping Research Data Safe report produced in 2008 but also allowed some limitations and areas needing further development to be defined. In section 9 we have discussed these areas and made the following recommendations for future work as follows:

Recommendation 1: 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 (and future prospective evidence accumulated to it). This is more amenable to mapping into a consistent framework for analysis and often more complete than more

historic cost evidence. A range of potential sources of such cost evidence are identified in our survey.

Recommendation 2: The KRDS project team should seek future opportunities to extend the costs survey; raise awareness of KRDS internationally; and develop research partnerships on digital preservation costs.

Recommendation 3: From KRDS2 outcomes, it is likely that the largest potential cost efficiencies will come from future tool development supporting ingest and access activities. Funders may wish to focus on investigating the potential benefits that could arise from further automation of these activities.

Recommendation 4: Examine further development of the pre-archive phase of the KRDS2 activity model and produce versions of the model from a researcher's perspective.

Recommendation 5: Seek to implement KRDS2 in cost spreadsheets and continue research on implementation variables and metrics that could enhance them.

Recommendation 6: 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.

Recommendation 7: Elements from KRDS2 and its findings should be considered by JISC for inclusion in its Research 3.0 campaign to disseminate the results and findings to other end users.

Recommendation 8: JISC and other funders to consider further work on identifying and quantifying the benefits of research data preservation.

In summary, in KRDS2 we have identified and analysed collections of long-lived research data and information on associated preservation costs and benefits and provided a larger body of material and evidence against which existing and future research data preservation cost modelling exercises can be tested and validated. We believe this work will be critical to developing preservation costing tools and cost benefit analyses for justifying and sustaining major investments in repositories and data curation.

2. INTRODUCTION

Data has always been fundamental to many areas of research but in recent years it has become central to more disciplines and inter-disciplinary projects and grown substantially in scale and complexity. There is increasing awareness of its strategic importance as a resource in addressing modern global challenges and the possibilities being unlocked by rapid technological advances and their application in research. However, there are several significant challenges facing the UK academic community relating to the long-term curation, storage, retrieval and discovery of research data. One of these challenges is developing a better understanding of the costs involved in long-term preservation of research data.

The Keeping Research Data Safe2 ("KRDS2") project aims to build on previous work on digital preservation costs for research data contained in the first Keeping Research Data Safe ("KRDS1") report (Beagrie et al 2008).

It has identified and analysed collections of long-lived research data and information on associated preservation costs and benefits and provides a larger body of material and evidence against which existing and future research data preservation cost modelling exercises can be tested and validated. We believe this work will be critical to developing preservation costing tools and cost benefit analyses for justifying and sustaining major investments in repositories and data curation.

3. OBJECTIVES AND METHODOLOGY

3.1. OBJECTIVES

The objectives of this study were to:

- Understand current requirements for the gathering of evidential material that will increase understanding of the long-term costs (and where possible the cost benefits) of research data preservation;
- Review international literature for relevant initiatives;
- Establish suitable criteria for identifying appropriate sources of information on preservation costs for research data;
- Undertake a survey of likely sources of information that may be appropriate for the aims of this study;
- Analyse identified research data collections and associated preservation cost information to determine their validity for the purposes of this study;
- Liaise and negotiate with research data collection owners and cost information providers to establish the terms on which their preservation cost information may be used;
- Analyse the cost components and variables associated with the long-term management of the identified research data collections and to compare and contrast them with the model proposed in the “Keeping Research Data Safe Report”;
- Make recommendations of suitability for the further analysis and exploitation of specific sources of information.

3.2. METHODOLOGY

To achieve these objectives we utilised the Keeping Research Data Safe cost framework as a tool for organising and scoping our work. We undertook a combination of desk research; a data survey; analytical work with national and disciplinary digital archives that have existing historic cost information for preservation of digital research data collections; and interaction with digital archives in research universities who have little or no historic cost information but

a strong interest in identifying criteria and metrics for capturing cost information going forward and in quantifying benefits.

We were already familiar with most of the international literature for relevant initiatives from work on Keeping Research Safe and more recently the cost/benefit work and literature review for the UK Research Data Service (UKRDS) Feasibility Study (Serco 2008 a and b), and participation in a review workshop for the proposed third-stage of the LIFE project (www.life.ac.uk). We updated and reviewed our existing research library from these projects to include recent work on LIFE2 (Davies (ed), 2008) and other relevant initiatives (Fry et al 2008, Blue Ribbon Task Force 2008 and 2010).

In addition to literature review, desk research involved contacting existing relevant projects to obtain and share emerging reports, information and methodologies to feed into our data survey and analytical work. For example, we contacted NASA who agreed to share the latest phases of development for the NASA Cost Estimation Tool (Hunolt et al 2008a, Hunolt et al 2008b, Hunolt et al 2008c, Hunolt et al 2008d).

As our project progressed, our work was also shared with new related projects funded by the JISC and others which started up during the course of our work including LIFE3, and the Cost of Digital Preservation Management project run by the Danish National Library and the Danish National Archives.

4. REVIEW OF THE KRDS1 ACTIVITY MODEL

4.1. INTRODUCTION

All of our project partners undertook a detailed review of the activity model published in KRDS1 against their existing preservation activities and had an opportunity to suggest potential changes or areas of difficulty in the published model. The overall finding from this review was that the KRDS1 Activity Model was robust and broadly a good fit to their activities. Some changes were suggested, mainly to the wordings of definitions and edits to the existing text.

One specific area of concern for some was the use of Open Archival Information System (OAIS) terminology (CCSDS 2002) and its potential for acting as a barrier to understanding for some user groups. After discussion it was decided that the original justification for use of OAIS terminology where appropriate in KRDS still stood. OAIS terms are well-defined, published, and well-established in the preservation community. However, we believe it will be important for the wording of the activity table to be reviewed and adapted as needed locally by users for their intended audience and their specific application.

In addition, three substantive changes or additions to activities were also identified by two or more reviewers and were agreed as changes to the KRDS2 activity model (see section 5):

- **The need to divide the "outreach and depositor support" sub-activity under Acquisition in the Archive phase in KRDS1**. Several national services reported that outreach providing data management advice was a significant activity for those charged with supplying advice and guidance to researchers preparing grant proposals in the pre-archive phase. A high percentage of these proposals would not be funded and would therefore not generate deposits. Other data producers than researchers could also be a significant target community for outreach. Similar concerns were raised by a university partner establishing a central support service for its researchers where outreach working with researchers from the moment they create their datasets to ensure that appropriate preservation actions are taken early in the research life cycle; and audits to understand the research data management requirements of their research groups, will be crucial pre-archive phase activities. It

was therefore agreed to introduce a new “Outreach” activity in the pre-archive phase and change the sub-activity under Acquisition to “depositor support” and amend definitions accordingly.

- **The need to divide the development of the archive’s Selection Policy and its application within the selection sub-activity of Acquisition.** Several reviewers pointed out the development of a selection policy is episodic as a cost and best separated out from the day-to-day application of policy. We have therefore inserted a new sub-activity for “develop policy and standards” under the administration activity and amended the selection sub-activity accordingly.
- **The need to cover staff training and development as a specific activity**. We have therefore inserted a new sub-activity for staff training and development under Common Services.

We also agreed that the presentation of the activity model should be altered to make it more user-friendly. For easier comprehension of the overall structure, we have provided a simple single page overview of the KRDS2 activity model showing the main phases and activities; and also modified how the detailed KRDS2 activity model is presented to the user.

5. THE KRDS2 ACTIVITY MODEL

5.1. INTRODUCTION

The KRDS2 Activity Model is an example of a lifecycle costing method applied to research data. Lifecycle costings model a lifecycle 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.

The first Keeping Research Data Safe report sets out the broader cost framework and guidance within which the KRDS2 activity model can be applied (Beagrie et al 2008).That cost framework consists of three parts:

- A list of key cost variables and units. This section describes key variables which affect the cost of preservation activities. The cost variables are divided into two major groups: economic adjustments and service adjustments.
- An activity model (now updated and replaced in KRDS2) for research data identifying activities with cost implications for preservation. This is sub-divided into Pre-Archive, Archive, and Support Services. Typically Pre-Archive activities relate to research projects in universities, and Archive activities to data archiving repositories run by universities or third-parties. Both of these relate to lifecycle 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. These are included in calculating full economic costs.
- A resources template. This 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.

Typically the activity model will help identify resources required or 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.

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

In addition to "macro" applications within or between institutions, the Framework can also be used to focus on particular activities and tasks within the two main lifecycle stages of Pre-Archive and Archive in the model.

5.2. Additional Observations on applying the Model from KRDS2

- We would stress that the activity model is generic and our expectation is that end-users will tailor it to their specific institution and requirements. In particular we have often re-used terms and definitions from the OAIS Reference Model (see KRDS1 activity model for annotated sources, Beagrie et al 2008). OAIS terms will be capitalised in the scope notes, e.g. Archival Information Package (AIP). For audiences unfamiliar with OAIS terminology, these may need further explanation or "translation" as appropriate for local use.
- The activity model is designed for costing preservation activities where there is a distinct archiving phase based on a designated archive centre or function. Although these exist within our case study sites and many universities, we have also encountered specific research disciplines and sub-disciplines where this is not the norm and the locus of preservation is a research group or even the individual researcher. We recognise the KRDS2 activity model contains many activities and sub-activities which are relevant to preservation in these scenarios but the presentation and structure of the KRDS2 model itself will need significant future adaptation if it is to be tailored specifically for them.
- In addition we have recognised that the KRDS2 activity model could be applied at different levels for different purposes. As noted by Gerlach in discussion of activity

based costings for IT services (Gerlach 2002, p 64-5), 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. Detailed activity modelling is usually needed for operations planning and process improvement, whereas more general high-level activity models are sufficient for cost management.

KRDS2 now caters for potential dual application of the activity model with two "versions" presented at different levels of detail. A single page overview (section 5.3) of the KRDS2 consisting of just the main phases, e.g. archive; and sub-phases e.g. ingest, has been produced 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. The detailed activity model provides options for more detailed operations planning and process improvement as well as the necessary definitions and scope of the phases and activities.

5.3. Overview of the Main Phases and Activities in the KRDS2 Activity Model

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	

5.4. Detailed Version of KRDS2 Activity Model

Pre-Archive Phase

Scope Notes: Primarily relates to research projects in universities creating research data for later transfer to a data archive. However activities can be adapted for first stages in piloting and development of a new data archive if required.

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.
Creation	The project activities involved in creating research data. Included to note any significant implications for preservation costs or archive access/use downstream.	

Activity	Sub-activity	Scope Notes
	Negotiate IPR/licensing/ ethics	These need to be dealt with at the earliest stages by the data creator so that when data is deposited into an archive there are no residual issues around IPR, licensing, or ethics. These can be very difficult to resolve at a later stage. Guidance on IPR, licensing and ethics may be available from the archive or funder to assist in this.
	Generate descriptive metadata	Generating the Descriptive Information for research data. This will form part of the Submission Information Package (SIP) deposited with the archive at a later stage.
	Generate user documentation	The producer of the data needs to take into account whether users outside of the project may access the data and document accordingly.
	Generate customised software	This includes custom interfaces and applications if required. Such software will require specification, testing and implementing and include detailed documentation. Standardising on a set of supported software will be more cost effective and should be encouraged.
	Data management	Services and functions for populating, maintaining, and accessing a wide variety of data by the project.
	Create submission package for archive	Format/contents and the logical constructs used by the producer and how they are represented on each media delivery or in a telecommunication session. Submission Information Package (SIP): an Information Package that is delivered by the producer to the archive for use in the construction of one or more Archival Information Packages (AIP).

Archive Phase

Scope Notes: The activities required for long-term archiving of research data.

Activity	Sub-activity	Scope Notes
Acquisition	The processes involved in acquiring research data for an archive.	
	Selection	The application of the archive's selection policy.
	Negotiate submission agreement	The communication and negotiation of submission agreements with producers/depositors.
	Depositor support	Support and encouragement for researchers and others with data to deposit.
Disposal	The transfer to another archive or controlled destruction of material by the archive.	
	Transfer to another archive	Transfer material to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.
	Destroy	Destroy material which has not been selected for long-term curation and preservation. Documented policies, guidance or legal requirements may require that this be done securely.
Ingest	The Ingest functional area includes receiving, reading, quality checking, cataloguing, of incoming data (including metadata, documentation, etc.) to the point of insertion into the archive. Ingest can be manual or electronic with manual steps involved in quality checking, etc.	

Activity	Sub-activity	Scope Notes
	Receive submission	This provides the appropriate storage capability or devices to receive a submission of data. Submissions may be digital delivered via electronic transfer (e.g., FTP), loaded from media submitted to the archive, or simply mounted (e.g., CD-ROM) on the archive file system for access. Non-digital submissions would likely be delivered by conventional shipping procedures. The Receive Submission function may represent a legal transfer of custody for the Content Information and may require that special access controls be placed on the contents. This function provides a confirmation of receipt to the producer, which may include a request to resubmit in the case of errors resulting from the submission.
	Quality assurance	The Quality Assurance function validates (QA results) the successful transfer of the data submission to the staging area. For digital submissions, these mechanisms might include Cyclic Redundancy Checks (CRCs) or checksums associated with each data file, or the use of system log files to record and identify any file transfer or media read/write errors. In addition to these basic integrity checks, it may also include many more discipline-specific tests on the quality of data and metadata.
	Generate information package for archive	This deals with the transformation of the submitted data (Submission Information Package) into a format suitable for the archive. Archival Information Packages within the system will conform to the archive's data formatting and documentation standards. This may involve file format conversions, redaction, disclosure checking, data representation conversions or other reorganisation of the content information.
	Generate administrative metadata	Administrative metadata about the preservation process: • pointers to earlier versions of the collection item • change history

Activity	Sub-activity	Scope Notes
	Generate/upgrade descriptive metadata and documentation	Includes the development (or upgrading of received) data and product documentation (including user guides, catalogue interfaces, etc.) to meet adopted documentation standards, including catalogue information (metadata), user guides, etc., through consultation with data providers.
	Co-ordinate updates	Provides a mechanism for updating the contents of the archive. It receives change requests, procedures and tools from Manage System Configuration (Operating system services).
	Reference linking	The semantic linking of primary data to textual interpretations of that data.
Archive Storage	Services and functions used for the storage and retrieval of Archival Information Packages (AIPs).	
	Receive data from ingest	The Receive Data function receives a storage request and an AIP from Ingest and moves the AIP to permanent storage within the archive. This function will select the media type, prepare the devices or volumes, and perform the physical transfer to the Archival Storage volumes.
	Manage storage hierarchy	The Manage Storage Hierarchy function positions, via commands, the contents of the AIPs on the appropriate media based on storage management policies, operational statistics, or directions from Ingest via the storage request. It will also conform to any special levels of service required for the AIP, or any special security measures that are required, and ensures the appropriate level of protection for the AIP.
	Replace media	This provides the capability to reproduce the Archival Information Packages (AIPs) over time

Activity	Sub-activity	Scope Notes
	Disaster recovery	Disaster recovery is the process, policies and procedures related to preparing for recovery or continuation of technology infrastructure critical to an organisation after a natural or human-induced disaster. Disaster recovery planning should include planning for resumption of applications, data, hardware, communications (such as networking) and other IT infrastructure. It is a subset of a larger process known as business continuity planning that includes planning for non-IT related aspects such as key personnel, facilities, and crisis communication. It should provide a plan for and testing of mechanisms for duplicating the digital contents of the archive collection and storing the duplicate in a physically separate facility and recovery from them. This function is normally accomplished by copying the archive contents to some form of removable storage media (e.g., digital linear tape, compact disc), but may also be performed via hardware transport or network data transfers. The details of disaster recovery policies are specified by Administration.
	Error checking	Provides statistically acceptable assurance that no components of the AIP are corrupted during any internal Archival Storage data transfer. It requires that all hardware and software within the archive provide notification of potential errors and that these errors are routed to standard error logs that are checked by the Archival Storage staff.
	Provide copies to access	The archive design will reference the preservation strategy and policy, considering off-site copies and any discipline specific requirement for multiple versions or editions. The number of versions and copies affects storage and management costs.
Preservation Planning	The services and functions for monitoring, providing recommendations, and taking action, to ensure that the information stored in the archive remains accessible over the long term, even if the original computing environment becomes obsolete.	

Activity	Sub-activity	Scope Notes
	Monitor designated user community	The Monitor Designated User Community function interacts with archive Consumers and Producers to track changes in their service requirements and available product technologies. Such requirements might include data formats, media choices, and preferences for software packages, new computing platforms, and mechanisms for communicating with the archive.
	Monitor technology	The Monitor Technology function is responsible for tracking emerging digital technologies, information standards and computing platforms (i.e., hardware and software) to identify technologies which could cause obsolescence in the archive's computing environment and prevent access to some of the archive's current holdings.
	Develop preservation strategies and standards	The Develop Preservation Strategies and Standards function is responsible for developing and recommending strategies and standards to enable the archive to better anticipate future changes in the Designated User Community service requirements or technology trends that would require migration of some current archive holdings or new submissions.
	Develop packaging designs and migration plans	The Develop Packaging Designs and Migration Plans function develops new Information Package designs and detailed migration plans and prototypes. This activity also provides advice on the application of these Information Package designs and migration plans to specific archive holdings and submissions.
	Develop and monitor SLAs for outsourced preservation	Where a decision is made to outsource some or all archive functions a contractual relationship will be established and to ensure service requirements are understood and met a Service Level Agreement (SLA) needs to be put in place and monitored.

Activity	Sub-activity	Scope Notes
	Preservation action	Preservation action covers the process of performing actions on digital objects in order to ensure their continued accessibility. It includes evaluation and quality assurance of actions, and the acquisition or implementation of software to facilitate the preservation actions. Preservation has a feedback loop back into/through Ingest functions in the activity model.
	Generate preservation metadata	The information an archive uses to support the digital preservation process. Specifically, the metadata supporting the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context. Preservation metadata thus spans a number of the categories typically used to differentiate types of metadata: administrative (including rights and permissions), technical, and structural. The documentation of digital provenance (the history of an object) and to the documentation of relationships, especially relationships among different objects within the archive.
First Mover Innovation	Where preservation functions and file formats are evolving a high-degree of expenditure might be required in implementation phases and in R&D developing the first tools, standards and best practices. This cost is highly variable for individual institutions and significantly dependent on how much is done solely by the institution or by a wider community. Communities or vendors can make significant up-front investments in first solutions and standards which affect downstream preservation costs. Most data archives participate in these activities to some degree although leadership and significant effort may be restricted to a few large institutions. Added as it has significant implications for cost modelling or potential for use/re-use.	
	Develop community data standards and best practice	Whilst preservation functions are evolving professional involvement in developing community standards and best practises is a cost effective approach to the delivery of efficient solutions.
	Share development of preservation systems and tools	Combining effort with others in the community can deliver significant developments for relatively small cost to individual institutions, and may even attract external funding.

Activity	Sub-activity	Scope Notes
	Engage with vendors	This might include beta-testing, participation in user groups, and development of commercial partnerships.
Data Management	The services and functions for populating, maintaining, and accessing both descriptive information which identifies and documents archive holdings and administrative data used to manage the archive.	
	Administer database	Responsible for maintaining the integrity of the Data Management database, which contains both Descriptive Information and system information. Descriptive Information identifies and describes the archive holdings, and System Information is used to support archive operations.
	Perform queries	Receives a query request from Access and executes the query to generate a result set that is transmitted to the requester.
	Generate report	Receives a report request from Ingest, Access or Administration and executes any queries or other processes necessary to generate the report that it supplies to the requester. Typical reports might include summaries of archive holdings by category, or usage statistics for accesses to archive holdings.
	Receive database updates	Adds, modifies or deletes information in the Data Management persistent storage. The main sources of updates are Ingest, which provides Descriptive Information for the new AIPs, and Administration, which provides system updates and review updates.
Access	Services and functions which make the archival information holdings and related services visible to consumers.	
	Search and ordering	This includes providing access to catalogue information and a search and order capability to users, and receiving user requests for data. “Order” implies a request /permission step, regardless of how implemented (e.g. manual or automated), where a request for a set of data or product instances, perhaps the results of (or a selected subset of the results of) a search, is processed and accepted or denied.

Activity	Sub-activity	Scope Notes
	Generate information package for dissemination to user	This function accepts a dissemination request, retrieves the Archival Information Package from Archival Storage, and moves a copy of the data to a staging area for further processing. The types of operations, which may be carried out, include statistical functions, sub-sampling in temporal or spatial dimensions, conversions between different data types or output formats, and other specialised processing. See also Generate Information Package for Archive in Ingest – as some archives may generate archive and dissemination version simultaneously,
	Deliver response	The Deliver Response function handles both on-line and off-line deliveries of responses (Dissemination Information Packages, result sets, reports and assistance) to consumers.
	User support	The user support functional area includes support provided in direct contact with users by user support staff, including training for users, user demonstrations, responding to queries, taking of orders, staffing a help desk (i.e., staff awaiting user contacts who can assist in ordering, track and status pending requests, resolve problems, etc.), etc. User support staff includes specialist expertise to assist users in selecting and using data and products.
	New product generation	Initial generation and reprocessing with quality checking of new data products produced from data or products previously ingested, or generated. Note that this has as a feedback loop back into/through Ingest functions.

Support Services

Scope Notes: Services and functions needed to control the operation of the other functional entities on a day-to-day basis.

Activity	Sub-activity	Scope Notes
Administration	The functions needed to control the operation of the other functional entities.	
	General management	Management includes management and administration at the data service provider level ("front office") and direct management of functional areas. Management also includes staff with overall responsibility for internal and external disciplinary specialist activities, information technology planning, and data stewardship.
	Customer accounts	To facilitate billing and payment receipts from "customers". Also useful for reporting usage and restricting access as appropriate to closed collections with specific license conditions.
	Administrative support	Administrative support and control provided by office managers, personal assistants and clerical staff.
	Develop policies and standards	This function is responsible for establishing and maintaining the archive's standards and policies. These include initial format standards, documentation standards, model deposit agreements, user agreements and data licensing, the archive's selection policy and the procedures to be followed during the Ingest process. They will normally involve a large initial effort to develop and then regular review and small updates over time and rarer major re-drafting.
Common Services	These are the other shared supporting services supplied by the institution or located within the archive.	
	Operating system services	Provide the core services needed to operate and administer the application platform, and provide an interface between application software and the platform.

Activity	Sub-activity	Scope Notes
	Network services	These provide the capabilities and mechanisms to support distributed applications requiring data access and applications interoperability in heterogeneous, networked environments.
	Network security services	Network security services include access, authentication, confidentiality, integrity, and non-repudiation controls and management of communications between senders and receivers of information in a network
	Software licences and hardware maintenance	Ensure that correct software licenses are in place and that they are renewed in a timely way. Also, determine the most appropriate level of hardware maintenance for the configuration and put in place call procedures and reporting with the supplier. Renew in a timely way.
	Physical security	With reference to facility and infrastructure. The service will have a disaster recovery plan to deal will all eventualities and to mitigate risk.
	Utilities	Supply of uninterrupted power supply, air conditioning, water etc.
	Supplies inventory and logistics	Management of supply chain, movement of goods, and recording of purchases and deliveries.
	Staff training and development	Support for training or developing archive staff to carry out particular roles.

Estates

Scope Notes: Estates management and attendant costs includes leasing of premises, space management and maintenance. Treated as a cost element in TRAC separate from other common services and charged at variable rates according to function, e.g. laboratory/non-laboratory.

6. THE COSTS DATA SURVEY

6.1. INTRODUCTION

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. We used our desk research and input from the project partners to prepare selection criteria for identifying appropriate sources of information to feed into our data survey. Our selection criteria and definition of scope for research data (http://www.beagrie.com/KRDS2_selectioncriteria.pdf) for this may be downloaded from the project website.

We prepared a survey proforma to identify key research data collections with information on preservation costs and issues. Between September and November 2009 we made an open invitation via email lists and the project blog and project webpage for others to contact us and contribute to the data survey if they had research datasets and associated cost information that they believed may be of interest to the study. The project partners in KRDS2 also contributed to the data survey. This section provides a short overview of the results.

6.2. OVERVIEW

13 survey responses were received: 11 of these were from UK-based collections, and 2 were from mainland Europe. Two further potential contributions from the USA were unfortunately not available in time to be included.

The responses cover a broad area of research including the arts and humanities, social sciences, and physical and biological sciences and research data archives or cultural heritage collections. Each survey response is approximately 6-8 pages in length. The British Atmospheric Data Centre (BADC) response is 22 pages as we have included supplied supplementary material and mappings. The Dutch Data Archiving and Networked Services (DANS) indicated that a study of their costs is nearing completion which will provide detailed information of all their operational costs. More detailed cost studies and analyses have also been undertaken at a number of our KRDS2 project partners. Further analysis and discussion of preservation costs or benefits for collections at the Archaeology Data Service (ADS), the National Crystallography Service/eCrystals (Southampton University), National

Digital Archive of Datasets (NDAD) at the University of London Computer Centre, University of Oxford, and UK Data Archive (UKDA) are available in sections 7 and 8.

Cost information is available for most of the KRDS2 main activity phases (pre-archive, archive, access, support services, and estates) although the depth and breadth of information available from different collections varies considerably (see individual responses). Most cost information is relatively recent at least in terms of information which would be amenable to comparative analysis. Most of the data is potentially available for research subject to confidentiality or other terms and conditions.

Summary of KRDS2 Data Survey Responses

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 1: 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).

The survey questionnaire sought to identify cost information available for the main KRDS2 activities in the Pre-Archive and Archive phases. The number of institutions with information on each main activity in the Pre-Archive and Archive phases is also shown in Figure 2 below. Information for some activities is very high (archival storage cost information is available in 100% of the responses). Other more infrequent activities such as disposal (and perhaps also preservation planning) are less well represented. Knowledge of acquisition costs is also relatively low (46%).

Despite the fact that all responses were received from archives, some information on pre-archive costs was forthcoming: either because institutions were also involved in data creation (e.g. digitisation, research experiments) themselves, or because they had access to costs of research projects via funding bodies.

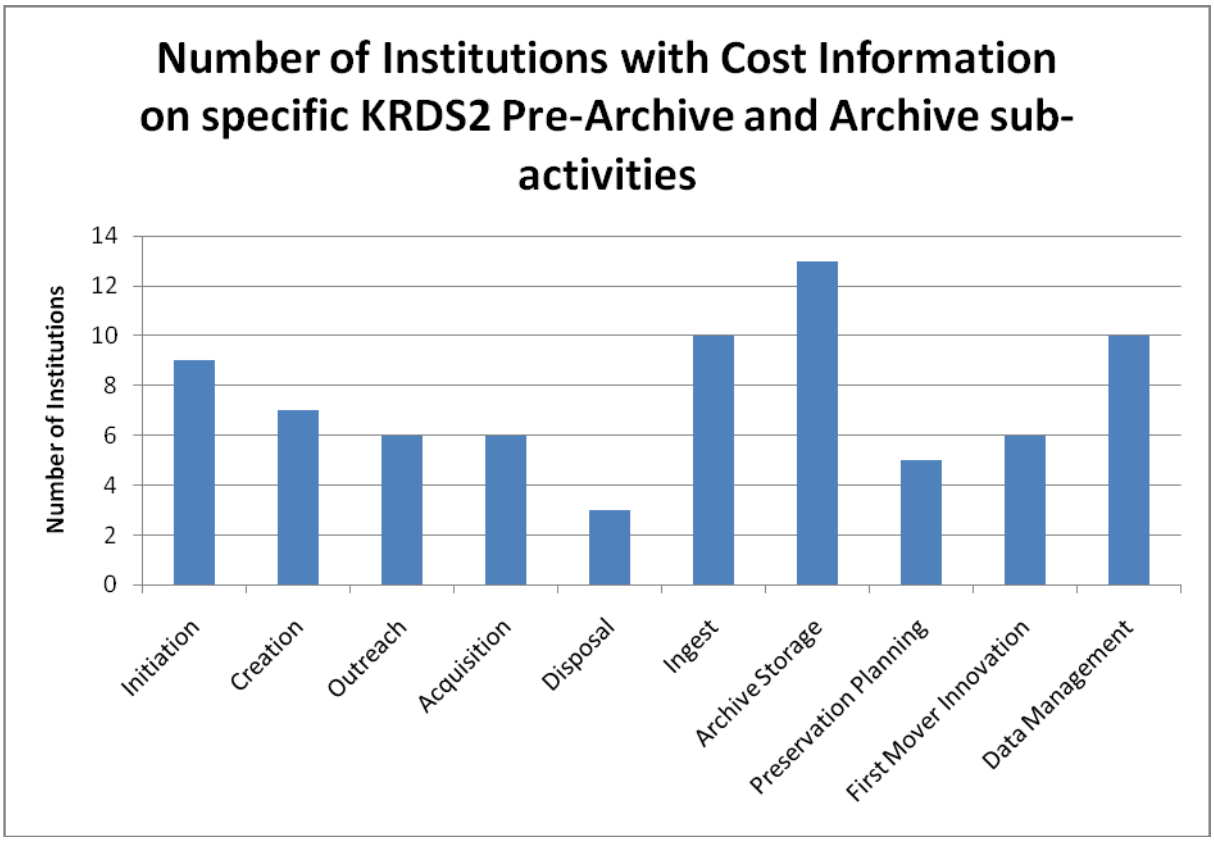


Figure 2: Number of Institutions with Cost Information on specific KRDS2 Pre-Archive and Archive sub-activities

6.3. Data Survey Responses

Individual completed responses to the data survey provide more detail and are available on the project website from the links below (urls for the links are also provided for those working from a print copy).

UK Responses

ADS (Archaeology Data Service) - http://www.beagrie.com/survey/ADS.doc

BADC (British Atmospheric Data Centre) - http://www.beagrie.com/survey/BADC-NERC.doc

eCrystals (National Crystallography Service, University of Southampton) - http://www.beagrie.com/survey/ecrystals.doc

EDINA (UK Borders Service, EDINA, University of Edinburgh) - http://www.beagrie.com/survey/Edinburgh.doc

Linnean Society (Linnean Society Collection, University of London Computer Centre) - http://www.beagrie.com/survey/ULCC-Linnean.doc

NDAD (National Digital Archive of Datasets, University of London Computer Centre) - http://www.beagrie.com/survey/ULCC-NDAD.doc

NLW (Welsh Journals Online, National Library of Wales) - http://www.beagrie.com/survey/NLW.doc

Oxford (University of Oxford) - http://www.beagrie.com/survey/Oxford.doc

Rutherford (Rutherford Appleton Laboratory, Science and Technology Facilities Council), UKDA (UK Data Archive) - http://www.beagrie.com/survey/RAL-STFC.doc

VADS (Visual Arts Data Service) - http://www.beagrie.com/survey/VADS.doc

International Responses

BABS (Bibliothekarisches Archivierungs- und Bereitstellungssystem -The Library Archiving and Access System- Bavarian State Library, Germany) - http://www.beagrie.com/survey/BSL.doc

DANS (Data Archiving and Networked Services, The Netherlands) - http://www.beagrie.com/survey/DANS.doc

7. ANALYTICAL WORK ON PRESERVATION COSTS

7.1. INTRODUCTION

We selected three organisations with collections identified during KRDS1 and the EIDCSR project at Oxford which we had felt had promising preservation costs information for further analysis during KRDS2. We used the Keeping Research Data Safe cost framework as a tool for organising and scoping our work. All our partners in this work analysed their activity costs associated with the long-term management of the identified research data collections and compared and contrasted them with the model proposed in the Keeping Research Data Safe1 Report. This work has fed into our review of the KRDS2 activity model (sections 4 and 5 above) and our Costs Data Survey (section 6). For the four organisations and their collections selected for more detailed work, cost datasets were then collated or in some cases generated for KRDS2 by the partners and analysed. The results of these analyses are presented below.

7.2. ARCHAEOLOGY DATA SERVICE COSTS ANALYSIS

Introduction

The Archaeology Data Service (ADS) supports research, learning and teaching with high quality and dependable digital resources (see http://ads.ahds.ac.uk/). It does this by preserving digital data in the long term, and by promoting and disseminating a broad range of data in archaeology. The Collection is a broad church, from pdfs of journal back runs to downloads of excavation data including PDF, TIFF, databases, spreadsheets, CAD (dxf, dwg), geophysics data (xyz), and GIS (shp) video. The total size of the collection is 1.5 Terabytes. Access to the collection is via the internet.

The ADS featured in the KRDS1 report, and a case study was devoted to its charging policy (Beagrie et al 2008, p87-94). It is a partner in KRDS2 and completed a response to the KRDS2 survey of digital preservation costs data, which is available from the KRDS2 website (see http://www.beagrie.com/jisc.php). In addition, it made a confidential detailed costs spreadsheet for the archiving over the last 5 years of 24 of its collections available for further analysis in the project. These costs cover current expenditures and do not factor in the

amortization of the initial set-up costs of the archive. All costs are expressed in GB Sterling (£). A summary analysis of the costs data from ADS is provided below.

Key Summary Data

24 collections

Total size: 164.3 GB

Total preservation cost: £251,437.88

Average cost per MB: £1.53

Data Analysis

Although there are many factors that can impact per-unit costs, the ADS cost data suggests that scale may be significant. Examination of the cost data for the 24 ADS collections suggests a correlation between archive size and total costs. For the 12 smallest ADS collections, median cost-per-MB was £88.06. For the 12 largest ADS collections, median cost-per-MB was £1.54. Economies of scale usually emerge when fixed costs represent a substantial component of total costs. In the context of the ADS data, staff costs seem to represent the "fixed costs" of data curation: these costs appear to be substantial and not strongly correlated (if at all) with collection size. Larger collections therefore reduce per unit cost by spreading staff costs over higher volumes of data curation activity.

Examination of the data indicates the prominence of staff costs in the overall costs of data curation:

- Total staff costs (exclusive of FEC) as percent of all costs: 50%;
- Total storage costs as percent of all costs: 20%;
- Total staff costs are about 2.5 times larger than total storage costs.

Even though the archive collections vary considerably in size (ranging from 8 MB to 39.9 GB), staff costs as a percentage of total costs varied within a much narrower band across the collections:

- 12 of the collections exhibited a staff cost/total cost ratio of 60-62%;
- 6 of the collections exhibited a staff cost/total cost ratio of 50-59%;
- 2 of the collections exhibited a staff cost/total cost ratio of 40-49%;

- 1 of the collections exhibited a staff cost/total cost ratio of 30-39%;
- 2 of the collections exhibited a staff cost/total cost ratio of 20-29%;
- 1 of the collections exhibited a staff cost/total cost ratio of 10-19%;

Only 4 of the 24 archive collections exhibited a staff-to-total cost ratio less than 45%.

Given that staff costs appear to be at best only weakly correlated with collection size, this seems to suggest that expansion of archival capacity, in addition to lowering the average per unit cost of curation, is relatively inexpensive in absolute terms.

Looking at the distribution of staff costs over five major cost categories derived from the KRDS2 activity model (pre-archive, acquisition, ingest, archive, and access), the largest proportion is accounted for by the access category (31%). However, the activities leading up to and including ingest of the materials into the archive collectively account for 55% of total staff costs. Somewhat surprisingly (compared to some public perceptions), the process of actually preserving the materials (archive category) accounts for only 15% of total staff costs.

Looking at the combined archiving costs of pre-archive, acquisition, and ingest, it is interesting to see whether this cost varies with the size of the collection. Pre-archive is assigned the same figure for all collections, and is therefore uncorrelated with collection size. Acquisition costs and ingest costs do not seem to have a discernable correlation with collection size. It would be interesting to know more about the nature of these costs, and why they do not correlate with the size of the collections and this could be an area for future research. Given the data at hand, however, it would seem that the costs of expanding the size of the archive are primarily fixed.

Key Observations

Examination of the ADS cost data yields the following general observations:

- Economies of scale seem to be a salient feature of the ADS cost profile, with substantial savings in per unit cost achieved as the size of collections increase;
- The costs of long-term data curation/preservation are dominated by fixed costs, or more generally, costs that do not vary with the size of the collection. Once archival

capacity has been set up, the marginal cost of adding another MB of content seems to be quite low;

- The origin of the fixed costs component of overall costs seems to lie with the high proportion of staff costs associated with data curation. For the most part, staff costs only weakly correlate, or do not correlate at all, with the size of the archive;
- The cost of setting up and maintaining an apparatus for getting material into the archive seems to be much greater than the cost of setting up and maintaining an apparatus for preserving these materials over the long term.

7.3. UK Data Archive Activity Costs Analysis

Introduction

The UK Data Archive (UKDA) has a staff of approx. 50 FTEs and holds over 5,000 datasets with accompanying documentation, of value predominantly to the social science and history communities. Data is generally quantitative (microdata/macrodata) or qualitative. The microdata are usually coded responses to survey questions; microdata are aggregate numerical data (often erroneously known as “statistics”). Qualitative material includes in-depth interviews, diaries, anthropological field notes and complete answers to survey questions. Data comes, in general, from two main sources, academic researchers and government departments/national statistical agencies.

The UKDA was founded in 1967 and is one of the oldest digital archives in the UK. The UKDA contributed some cost information to KRDS1 (Beagrie et al 2008) and it was originally hoped in KRDS2 that the historic UKDA cost information could be analysed in detail as a cost series. In practice however, the existing historic cost data from 2002-3 and 2005-6 proved more difficult to work with and has more limitations than anticipated. The decision was taken therefore by UKDA to capture an entirely new costs dataset in June 2009 which

could be analysed in greater detail for KRDS2. Only limited comparison and analysis over time could be attempted given the limitations of earlier cost datasets.

This section describes in different levels of detail, the three activity based costing exercises held at the UK Data Archive in 2002-3, 2005-6, and 2009. However it only deals with the most recent study in depth for the reasons noted above.

Methodology for the 2009 study

All members of staff were asked to complete the (approximate) number of hours which they had carried out on KRDS2 activities during the month of June 2009. There is no reason to suppose that June is any more atypical than any other month, except that the hours spent answering user queries may be slightly lower than average because of the examination season. On the other hand, since most UK Data Archive staff are employed within the same broad functional areas as covered by the KRDS Activity model, any temporal effects may be minimized by the organisational model of the UK Data Archive, and by the aggregation of sub-activities in the analysis.

Staff had been made aware in mid-June that this questionnaire was to be circulated, and were in a position to keep personal timesheets. The final questionnaire was not circulated until the first week of July, and it is it possible that some staff may have been more diligent than others in keeping a record of their activities.

The questionnaire included the three levels of activity heading provided in KRDS down to the most granular sub-activity level but with some omissions for activities which are not carried out in the UK Data Archive. The definitions of all the activities were revised to use terminology more appropriate to the UK Data Archive's internal practices. A number of additional headings were added to account for organisational activities which did not fall within the KRDS headings. Some of these were subsequently added to the revised activity headings finalised in KRDS2 as they are appropriate to the model; others were not as they were almost purely related to the organisational structure of the UKDA and unrelated to digital preservation activities.

It is important to note that this activity based costing exercise included all UK Data Archive staff. The UK Data Archive is an umbrella organisation which co-ordinates and runs national

services, as well as carrying out a number of related research projects. Hence the costs of the UK Data Archive as a whole do not reflect the costs associated with digital preservation. Despite this these costs can be seen as a reasonable proxy for such costs. It is also critical to take into account the particularly service-based nature of the Economic and Social Data Service, one of the services partially hosted at the UK Data Archive. The particular emphasis on user support in this service means that the costs of user support are higher than they might otherwise be in an institutional repository. Similarly, the UK Data Archive uses its "Acquisitions Review Committee" to appraise and select datasets for ingest into the collection; consequently, the proportion of time and thus cost associated with this activity will be much greater than for organisations which have a less strict selection policy. However, we believe that using all staff in the UK Data Archive gives a realistic indication of the costs of the activities which are carried out there. Even costs which may at first seem to some to be unrelated to digital preservation, e.g. providing data management guidelines to researchers, do in fact help to reduce the ingest costs of the digital preservation cycle and are key to doing digital curation properly in the long-term.

Results of the 2009 Cost Analysis

During the data capture process, staff asked a number of questions relating to the activities and how they should be included in their responses, despite revising the terminology of the descriptions for a UK Data Archive audience. Not only were some basic activities misunderstood, e.g. the basic ingest process 'Generate Information Package for Archive', but there were troubles interpreting the differences between line and general management. Any organisation attempting to track costs using this activity model should scrupulously check that the definitions are relevant to local circumstances. We recommend that the internal practices of the organisation are taken into account whenever similar activity costings take place.

Despite these definitional problems, the experiment has shown that the activity model is reasonably robust in itself. Organisational practice may mean that individuals find it difficult to differentiate between different activities, but they can be reasonably certain about the top level headings. However, even within the "revised" model used for this survey, there were some minor potential overlaps between these high level categories which have made data

analysis problematic. These overlaps occur most notably within the management function. Internal line management, where members of staff are given tasks for a period were considered by some managers to be under the heading of “general management” and by others as under the specific activity which was being carried out. In analysis, all time reported on internal meetings and line management has been reallocated. Consequently, the reporting of this survey concentrates on these highest level activities, and only refers to the most granular activities where it seems appropriate.

It is worth mentioning that further complications in responses arose due to unfamiliar terminology in the model (and questionnaire) and that some responses were based on individuals’ reaction to the activity heading and not the description. One staff member allocated two hours to the “provide copies to access” activity which is a fully automated process. This was simply a misunderstanding of the activity title. This misunderstanding may not have occurred if the activity had simply been “ingest”. A further consequence of providing three levels of sub-activity in this exercise was that some of the activities are too granular for some of those who carry them out to recognise the differences between them. One member of staff simply ascribed their entire activity to ingest, given that the tasks carried out encompassed all of the different ingest activities.

It is also worth noting that the very wide range of activities carried out by the UK Data Archive also means that some members of staff are not employed directly related on the key services of the UK Data Archive in “keeping research data safe”, and thus some of their activities, e.g. project management, are not related to any of the activities in the model. Furthermore, some self-reported activities, including “university business” are likely to be specific to an organisation. Hence, it is worth keeping in mind when examining these results that they are organisation-specific, and that the organisation’s costs relate to the overall remit of the organisation and not just their “keeping research data safe” remit.

To make the questionnaire more straightforward, activities were to be measured by hour over the month. The result of this may have been that some members of staff exaggerated their hours of work. Consequently all hourly activities have been converted to a proportion of “paid-hours”, on the basis that regardless of the number of hours an individual works in a month their cost to the organisation is the same. This should be borne in mind. Of just the

full time members of staff, the total hours recorded (including sick, holiday, etc.) ranged from 108 to 233, a considerable range.

An initial analysis was made using the exact headings provided by staff. The outcome included approximately 40% of cost (not time!) of activities which were not explicitly included with the activity model. A very large proportion of these costs related to internal meetings, informal communications and leave of absence – either through leave or sickness. This is not a fault with the Activity Model per se, rather a demonstration that people find it very hard to respond to surveys of this nature, and it was considered at the outset to ask people to explicitly record leave of absence rather than allocate it themselves to an activity.

Furthermore all of the Information Development and Programming Section of the UK Data Archive were unable to satisfactorily sub-divide their working hours into the specific sub-activities in the model. After discussion with the project team these activities were included with the Data Management activity.

Finally for the purposes of this exercise we renamed the activity First Mover Innovation to 'Research and Development' to make it explicit that the responses included in this activity were generally reported as R&D. This is not precisely First Mover Innovation as defined within the activity model but there are likely to be similarities. Once activities had been reallocated and the total costs (salary and on-costs) for each individual included, the overall activity costs in the UK Data Archive were as follows:

Activity	% cost	% time
Archive: Acquisition	5.8	4.8
Archive: Ingest	21.5	22.2
Archive: Archive Storage / Preservation Planning	3.1	2.8
Archive: Research and development	6.9	6.9
Archive: Data Management / Information Development	15	14.7
Archive: Access	16.9	16.3
Support Services: Administration	21	23.8
Support Services: Common Services	5.1	4.9
Other	4.8	3.7
Total	100	100.1

Figure 3: Proportion of costs and time spent for different UKDA Archive and Support Service Activities in 2009.

The other activities, making up approximately 5% of the spend covered: disclosure checking, reformatting services, project management of external projects as well as some non-work related activities.

It is instructive to note that for the UK Data Archive that the percentage cost expended on specific activities is not hugely different to the proportion of time (hours) expended. For the UK Data Archive, it would be possible to estimate the cost of activities with reasonable accuracy from the number of hours expended by activity without having to recalculate on the basis of each individual's salary.

Key Observations from the 2009 Cost Analysis

- The applicability of these costs to other organisations must be seen in the light of the particular mission of the UK Data Archive which may differ from other organisations involved in digital preservation, and consequently provide a different spread of costs;
- The applicability must also be tempered by the fact that the majority of data ingested into the UK Data Archive is social or economic survey based data (though the overall collection is quite diverse) which means that both the subject-matter of the data and their file formats are relatively discreet, allowing steady throughput and for subject-based staff to be employed;
- The activity costs of an established organisation can often be a priori allocated. With a total FTE staff of 50.5,[1] the UK Data Archive is sometimes able to transfer staff members from one activity to another, but this is not always possible, and organisations with staff with specialised skill sets may suffer disproportionately;
- The whole organisational structure in which a digital repository sits may heavily affect the spread of costs which can be reported. The UK Data Archive, hosted by the University of Essex, carries out almost all its own financial and human resources activities with limited assistance of the host institution. Removing these activities from the costs of the UKDA would reduce the cost of support services by around 12%, but might increase the indirect overheads charged by the University to the UKDA;

[1] Figure correct at June 2009; excluding one FTE on secondment and including one FTE long-term sick leave.

[2] Plattering is a term used in the UKDA, since at least the early 1980s to denote the process by which

- The UKDA 2009 cost analysis showed particular difficulty with the practical allocation of tasks such as internal meetings between activities in the archive or administration phases and the need for overall adjustments to reflect activities such as annual leave and other absences. Capturing activity costs for an organisation at the most granular level of KRDS2 (i.e. down to sub-activities) across ALL of its activities is extremely onerous. Capture of costs data at higher levels from KRDS2 (i.e. activity or phase) would be easier to implement and may be more appropriate. This mirrors similar experience elsewhere (Gerlach et al 2002). These lessons have been reflected in our advice in section 5.2 on implementing the KRDS2 activity model and in providing a high-level overview version of the model (section 5.3) to guide most applications.

Comparison with the UKDA 2002-3 Cost Analysis

Activity Based Costing exercises had been carried out twice within the UK Data Archive before the KRDS2 case study. The first of these studies was carried out in 2002-3 and was designed to inform internal planning. It was a more formidable task than the 2009 study owing to the lack of predecessors. A few of the headings used in this data capture process are reproduced here to provide an insight into some of the difficulties encountered in mapping historical data to more recent KRDS2 headings and definitions:

- Acquisitions - booking in
- Acquisitions – general
- Acquisitions - negotiation for data
- Acquisitions - queries to depositors
- Data/documentation initial checks
- Library work
- Preservation (plattering/migration)[2]
- Research
- Translation

The headings used in the 2002-3 exercise had been slimmed down for analytical purposes and the raw data had been destroyed. While there is some level of comparability between

[2] Plattering is a term used in the UKDA, since at least the early 1980s to denote the process by which an ingested data collection is transferred to permanent storage. A number of checks are made to ensure the consistency of the file structures and the integrity of the data within the data collection. It was named after the optical platters used to store data.

surviving aggregate headings and data from this earlier exercise and the 2009 study, a precise mapping is impossible and only a heavily qualified comparison can be attempted.

The slimmed down headings and data from 2002-3 have been mapped (approximately) to those used in the 2009 study, and in a similar manner to the 2009 study, sick leave and annual leave have been re-factored into the main activities on a proportional basis. The results are presented in the table below.

Owing to the impossibility of recasting the 2003 "other" costs and separating out Research and Development from the main headings, the two sets of figures are not fully comparable. The major differences between the 2003 and 2009 costs is a considerable reduction on administrative costs, some of which may in 2009 have been included within the other activities; an increase in access costs which is partially due to increased usage and concomitant user support. What is noticeable however, it that the percentage of time spent on any activity was roughly the same as the percentage of overall cost spent on any activity, though there have been some interesting small changes. The method of collection and interpretation of the 2009 figures are discussed above.

Activity	**2003**	**2003**	**2009**	**2009**
	% cost	**% hours**	**% cost**	**% hours**
Archive: Acquisition	3.9	3.9	5.8	4.8
Archive: Ingest	16.2	20.1	21.5	22.2
Archive: Archive Storage / Preservation Planning	2	1.9	3.1	2.8
Archive: Research and development	N/A	N/A	6.9	6.9
Archive: Data Management / Information Development	21.1	20	15	14.7
Archive: Access	9.4	10.5	16.9	16.3
Support Services: Administration	35.4	32.6	21	23.8
Support Services: Common Services	4.6	3.9	5.1	4.9
Other	7.4	7.1	4.8	3.7
Total	100	100	100	100.1

Figure 4: Comparing the proportion of costs and time spent for different UKDA Archive and Support Service Activities in 2003 and 2009.

The 2005-6 UKDA Cost Analysis

The UKDA also carried out an additional activity-based costing exercise in 2005-6 for the East of England Digital Preservation Regional Pilot Project (DARP). The published report for DARP only dealt with some top-levels activities and generalised costs based on those activities which were considered to be relevant to the purposes of the specific needs of regional archives setting up digital preservation units (EERAC 2006). The focus of attention was on the unit of ingest rather than the overall costs of the organisation. Understanding the costs per unit of ingest may provide an additional method for organisations setting up digital preservation systems. The table below shows the indicative timings per activity per study in 2005/6. These timings have altered since this analysis and will continue to alter as automation of tasks increases, but they demonstrate the particular challenges for an archive working with a particular service element and dealing predominantly with the same forms of data.

Activity	Average Time (or range)	Variability	Notes
Acquisition			
Pre-deposit evaluation	3 hours	Low	
Licence & copyright agreement	1 hour - 2 days	High	
Completion of deposit forms	3 hours	Low	
Check basic elements in place	2 hours	Low	
Reception			
Secure transfer of records	30 minutes	Low	
Integrity check - data & metadata	1 hour	Medium	
Risk analysis - data vulnerability/specialist user support	45 minutes	Medium	
Conversion of data & metadata to preservation format	1 day	Low	
Conversion of data & metadata to dissemination format	5 minutes	Low	Done automatically via pre-programmed scripts. Time/cost is in developing and maintain automated routines.
Processing			
Disclosure control checks	1 hour	Medium	
Production of catalogue record	2 to 8 hours	High	

Activity	Average Time (or range)	Variability	Notes
Preservation			
Secure transfer of records to repository	30 minutes	Low	
Record & metadata storage (multiple media)			Most of the following tasks are automated and are carried out at different levels on a daily, weekly and monthly basis. The real time/cost issue here is establishing and maintaining the system.
Record storage (multiple secure environment)			
Preservation watch			
Refreshment			
Fixity checks			
Migration of file formats			
Access			
Direct from catalogue to preservation front end	1 to 5 days	High	
Delivery of multiple file formats			Done automatically once conversion to dissemination format complete. Time/cost is in developing and maintaining automated routines.
Delivery on multiple media	1 - 2 hours	Medium	
Front end authentication	30 mins per user	Low	
Access via intermediary	30 minutes	Low	
Provision of views of records			Achieved automatically once data mounted in on-line browsing software. Time/cost is in purchase or development of on-line system and subsequent maintenance
User support			
Technical support	15 - 30 mins per query	Medium	
Content support	30 minutes to 4 hours	Medium/Low	

Figure 5: The 2005-6 UKDA Activity Cost Analysis

Key Observations from UKDA 2002-3 and 2005-6 Cost Analysis

- The UKDA 2002-3 and 2005-6 Cost Datasets illustrate the inherent difficulties of retrospectively constructing a time series for digital preservation costs from historic data when survival of data is partial or it had been compiled for different purposes;
- Opportunities for developing a longitudinal series of cost information to analyse digital preservation costs may be best developed prospectively rather than retrospectively. Consistent data collection and terminology could then be applied.

7.4. ULCC National Digital Archive of Datasets (NDAD) cost analysis

Introduction

The National Digital Archive of Datasets (NDAD) is operated under contract by the University of London Computer Centre (ULCC) on behalf of The National Archives. NDAD contains UK government databases which have been designated for permanent preservation as public records. As well as the data itself, NDAD also contains supporting documentation (some born-digital, some digitised) and extensive contextual descriptive information.

As part of KRDS2, ULCC has contributed the Excel Cost Spreadsheet for the NDAD service (http://www.beagrie.com/KRDS2_NDAD_Costs_Spreadsheet.xls) together with a Guide to Interpreting and Using the NDAD Cost Spreadsheet (http://www.beagrie.com/KRDS2_NDAD_Spreadsheet_Guide.doc) authored by Kevin Ashley. Both are included in the supplementary materials for the KRDS2 project report on the project web page. The NDAD Cost Spreadsheet has previously been used as an exercise in digital preservation training events and may be particularly useful in training covering digital preservation costs. The accompanying Guide provides guidance to those wishing to understand and experiment with the spreadsheet.

The opening section (Context) of the Guide provides the background to NDAD and the Cost Spreadsheet. The next section (Service Model) describes the type of service that this Cost Spreadsheet was used for. The following section (Variables) explains the parts of the Cost Spreadsheet you might find it useful to adjust and why. The final section (Limitations) explains some of the limitations of the financial model.

Readers should be aware that although the NDAD Cost Spreadsheet is based on real costings and reflects the actual process of calculation ULCC used for the service in 2007-2010, the figures are not those which ULCC tendered for the contract. Some critical variables, particularly those relating to volume of work, were different. In addition, this reflects a costing exercise undertaken in mid-2007 using underlying data which itself mainly dated from 2006 and in which ULCC were trying to estimate costs for 3-5 years in the future in the context of bidding for a commercial contract.

A brief discussion and analysis of the costs data in the Spreadsheet itself for NDAD is provided below.

Data Analysis

The ULCC/NDAD cost data reflect a repository ingesting 36 data sets a year, with each data set 5 GB in size, for a total of 180 GB per year, and 900 GB over 5 years. Cost per GB is £5,282.93 (or £5.28 per MB) in the first year, and increases roughly at the designated rate of inflation over the succeeding four years. Because the cost projections for Years 2 through 5 are essentially the first year's costs adjusted for inflation, the analysis below focuses on the first year cost's as representative of costs incurred over the entire 5-year cycle.

One of the most salient features of ULCC's cost profile is the predominance of ingest staff costs as a fraction of overall annual costs. Ingest staff costs account for three-quarters of overall costs, or £3,936.82 per GB (£3.94 per MB). Staff costs in general (i.e., staff ingest costs plus development, management, publicity, and reporting) constitute the vast majority – 90% – of overall annual costs. These findings underline the conclusion (corroborated by ADS and Oxford) that curation of research data exhibits a labour intensity that is quite high. It also suggests that an area ripe for innovation may be automated solutions for certain aspects of the data curation process with very high staff costs such as pre-ingest and ingest.

According to the ULCC figures, fixed costs (i.e., those costs which are invariant to growth in the size of the archive in terms of newly ingested databases) account for about a quarter of overall costs. About 65% of these costs are associated with administrative or overhead staff costs (development, management, publicity, reporting); the remaining fixed costs pertain to capital equipment such as servers and PCs. The fact that about a quarter of overall costs is invariant to the rate of ingest suggests opportunity could exist (if NDAD was not operating on

a fixed budget) for lowering per-unit costs by expanding the scale of the repository and spreading variable costs over higher volumes of preservation activity.

Fixed costs are eventually not fixed but you have to scale up quite a way before that becomes an issue for ULCC, hence they have not factored this into the spreadsheet. As growth in ingested data continues, computing resources will eventually reach capacity and investment in additional equipment will be necessary.

Nearly all of the costs enumerated in the ULCC cost profile are subject to inflationary pressures. For the purposes of the data reported, a constant inflation rate of 3.5% was assumed. In practice, however, the rate of inflation can vary considerably: for example, in October 2009, the UK Retail Price Index (RPI) was estimated to be -0.8% (i.e., prices were actually falling). But as recently as October 2008, the RPI stood at about 5%. Long-term activities like data curation are especially subject to the vagaries of external economic forces, which increase the likelihood that actual costs will diverge substantially from projected costs.

ULCC maintains an environmentally-controlled “paper store” as part of the NDAD preservation activity for any original paper documentation that accompanies the datasets. Annual costs for operating the paper store are £19,200, and are included in the overall annual costs reported by ULCC.

The ULCC data indicates that the costs of simple bit preservation are relatively low compared to overall data curation costs. ULCC estimates that the per-MB cost of bit storage on tape (including multiple copies, multiple sites, periodic re-reading and checking, and periodic migration to new media) is £0.004. Maintaining accessible copies of preserved objects on disk adds another £0.0038 per MB. The annual administration and depreciation costs of one server is about £0.0561 per MB. Therefore, the total per-MB cost of simple bit storage is about £0.0639, or only 1 percent of the overall annual cost per MB (£5.28) calculated above. This suggests that the cost of simply ensuring that digital data persist and nothing more is in fact a very small proportion of overall curation costs.

Finally it is quite possible to contemplate a model in which ULCC could put far less effort into the ingest process, and value-added metadata and user documentation and hence transfer effort to the consumer. This would greatly reduce ingest costs, but would effectively change

the Designated User Community to a smaller set of people who could make use of any original supporting documentation to interpret and access the data themselves.

Key Observations

- As with other data analyses, the ULCC data exhibits a heavy predominance of staff costs in relation to overall curation costs. This in turn suggests that as currently practiced, data curation is a highly labour-intensive activity;
- Also corroborating other analyses, the cost of simple bit storage for ULCC appears to constitute a very small proportion of overall curation costs;
- The costs of ingest – receiving data, preparing it for long-term storage, and incorporating it into the digital archive – receives the largest allocation of resources. In comparison, the resource allocation devoted to storage management and related activities is quite small;
- The ULCC data illustrates the importance of inflation and other elements of the external economic environment, which might serve to drive a wedge between projected costs and actual costs;
- In a fixed-cost environment, the Designated User Community is also partly shaped by the access and ingest services which one can afford to provide.

7.5. University of Oxford Cost Analysis

Introduction

The Embedding Institutional Data Curation Services in Research (EIDCSR) project (http://eidcsr.oucs.ox.ac.uk/) is addressing the research data management and curation challenges of two research groups in the University of Oxford. In recent months the EIDCSR Project has been taking part in a Keeping Research Data Safe 2 (KRDS2) case study on cost benefits.

The KRDS2 case study in Oxford aimed at gathering cost information related to the creation, management and curation of the research data produced by the research groups participating in EIDCSR. This Oxford perspective complements other KRDS2 participants as it provides access to data from multiple disciplinary domains and is not a national data

centre. Furthermore the remit of the EIDCSR project means that a much greater emphasis has been placed on being involved with researchers at the creation stage (while national data centres are increasingly involved in this, it has not been their primary focus which has been on receiving data from researchers and then to manage, provide access and preserve them). In that sense, national data centres and other centralised archives are more organised along the lines of the open archival information systems (OAIS) reference model (CCSDS 2002) and thus able to capture costs in the Activity Model relatively easily. In contrast, capturing costs information at Oxford presented a real challenge, as data management and curation are undertaken as an institutional federation of services provided by a variety of departments such as Libraries, Computing Services, Research Services and the research departments themselves.

This distributed environment means that the Oxford case study needed to be approached from a slightly different perspective. In Oxford, the activities around data management and curation did not follow an ideal OAIS model but proved to be a combination of local actions (at the research group level) dealing with the creation and some form of data management and some central (service provider level) curatorial activities such as metadata management and back-up.

Data Analysis

Some of the most interesting costs from the Oxford survey were those related to the creation of the data. Researchers were able to provide accurate estimates on the costs of generating their data in terms of staff time as well as costs of lab equipment.

One of the established central services included a back-up and long-term file store service provided by Oxford University Computing Services. This file store is used to keep copies of the data safe and relies heavily on researchers to decide what to keep or destroy, for how long to keep it and when to undertake any preservation actions needed.

Other related curatorial services and activities where costs were identified included the audit of data requirements and practices, creation of metadata, development of policy and implementation of workflow tools that allow researchers to easily make use of central services such as metadata management or archive and back-up. All these are undertaken as part of the EIDCSR Project and are not yet established as services for researchers in the

University. The diagram below (Figure 6) uses a bubble chart to present the aggregated costs of creating data and managing them locally by the research groups, the cost of curation and the back-up and long-term file store. 73% of the total expenditure across eight years is related to the creation of the data. Curatorial activities undertaken as part of the EIDCSR project cover 24% of the costs. It is important to note that EIDCSR is a research and development project to establish policies and methods and therefore costs of an established curation service would be expected to be lower. Finally the back-up and long-term filestore that ensures the data to be securely stored for five years and the local data management are only 2% and 1% respectively of the total costs.

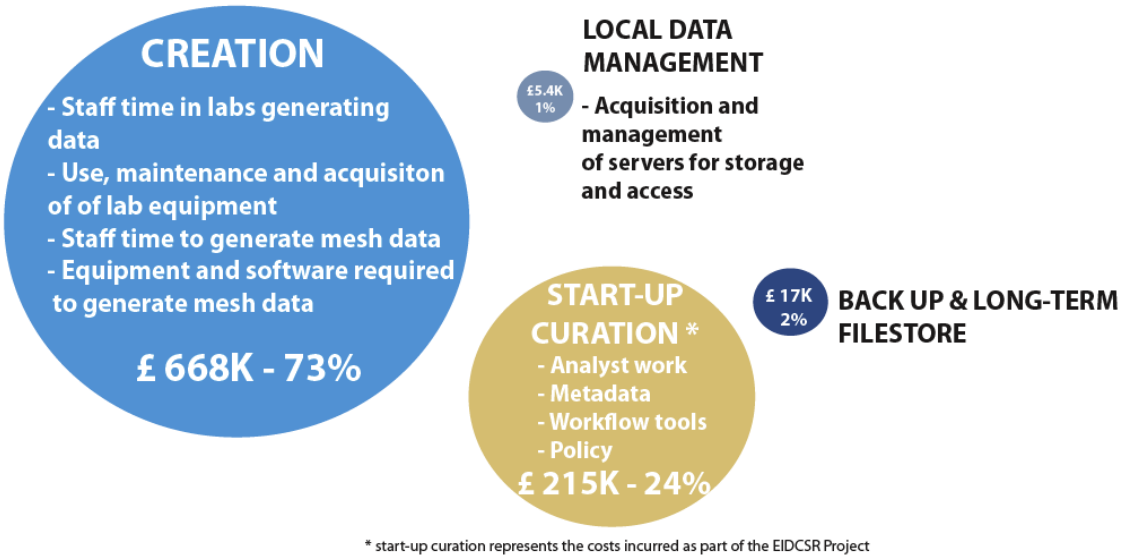


Figure 6: Data Management and Curation Costs from the Oxford Survey

The following diagram (Figure 7), shows how the activities with costs associated take place in time with creation and local management occurring in the first three years, curation starting before the end of those first three years and over and finally back-up and long-term filestore taking place for the following five years. The biggest proportion of the costs is concentrated at the beginning of this lifecycle and then they go down with time.

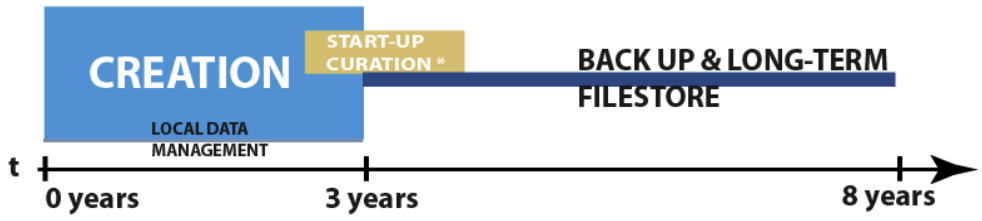


Figure 7: Data Management Activities Placed in Time.

It is extremely hard to estimate the future steady-state curation cost for a University like Oxford given our current knowledge. Resolving this will require further observation and analysis over a period of years as local curation services develop. Although the curation in this case is undertaken through the EIDCSR project, it is of a research and development nature. Established data curation services will always need to have an element of research and development to ensure their continuous service improvement. Therefore it is foreseen that the cost of institutional curatorial services in Oxford will not represent such a high percentage but it is currently unclear how much lower they will be and how long it will take to develop them.

All the costs identified through the survey used Full Economic Cost (FEC) models to take account of the direct, indirect and estates costs. FEC is widely used across the University, and it is well understood and accepted. Therefore it makes sense to build on this model to develop data management and preservation costing tools.

The cost information gathered was organized using the activity model developed by KRDS2 without any further normalisation of measuring units like size of research team or size of data. Further work is required to find measures to normalise the data so that it can be applied in different cases.

After collecting and organizing the cost information, it was useful to think in terms of benefits using the KRDS2 benefits taxonomy. One of the dimensions present in this taxonomy was of particular relevance in Oxford, near-term benefits. Attempting to curate researcher's data requires a strong engagement with researchers. This may be fostered by understanding researchers' challenges with data and highlighting the near-term benefits of curating data relevant to them. Examples from Oxford have been included in the discussion of the KRDS2 Benefits Taxonomy (see section 8, and Macdonald and Martinez-Uribe 2009).

Key Observations

Examination of the Oxford cost data yields the following general impressions:

- The costs of data curation are small in comparison with the costs of data creation. The majority of lifecycle costs are incurred before the data ever enters the preservation repository;
- The cost data suggest an extra cost of roughly 20% over and above the cost of data creation to maintain the data for five years;
- As with other data curation efforts, staff costs represent a major portion of overall curation costs;
- The costs of simple bit storage seem to represent a small proportion of curation costs, and an even smaller proportion of overall costs.

According to the Oxford data, only 12% of overall costs are assigned to Archiving activities. As might be expected given the research and development nature of the EIDCSR project, the majority of these costs are represented by First Mover Innovation activities, indicating that only £23,599.20, or less than a third of the Archiving costs and only 3% of total costs, are directly attributable to archival functions (metadata management and HFS archiving resources). Drilling down still further, the Oxford numbers suggest that metadata management within the project accounts for nearly 60% of these direct archiving costs, and therefore greatly exceeds the core costs of bit storage.

Oxford reports:

Pre-archive costs:	£268,619.00
Archive costs:	£95,226.20
Support costs:	£451,507.40
Total cost:	£815,352.60

The cost data for the Archiving category is calculated for 4 terabytes. Assuming that the total cost of £815,352.60 represents the “lifecycle” (3 years of creation and 5-years of long-term storage) costs for 4 terabytes, this indicates a per-MB cost of £0.20.[3] As with the ADS data,

[3] Note that this figure is not the total preservation cost as all other costs associated with the other preservation activities would need to be included.

this finding once again highlights the importance of scale in reducing the per-unit cost of long-term data curation.

The data curation involves three different kinds of data: histology data, MRI data, and mesh data. The salaries, equipment, indirect costs, and estate costs associated with the creation of the three forms of data total to £666,290.00, or 82% of the overall costs of data preservation. This finding suggests that the vast majority of costs are incurred before the data is even ingested into the repository.

Turning to the cost figures for the HFS Archive, Oxford reports cost data for the following components of the service: staff, non-staff, estate, and indirect. Non-staff costs (primarily the cost of media and media maintenance) account for the largest share of this cost at 40%; staff cost account for the next largest share at 29%. Indirect costs comprise 27%; estate costs are negligible. These numbers may require more nuanced interpretation, however, since one would surmise that much of the indirect costs and estate costs are attributable to staff. In this case, the share of total cost directly or indirectly linked to staff would rise considerably, possibly making this category the largest component of overall archiving cost. This would corroborate the analysis of the ADS data, which suggested that a significant proportion of overall costs were allocated to staff; indeed, only 4 of the 24 archived ADS data collections exhibited a staff-to-total cost ratio less than 45%.

The HFS archiving cost seems to represent the core cost of simple bit storage (i.e., exclusive of other data curation costs such as metadata management). Oxford notes that for Research Council-funded projects, only 80% of this cost is recouped. The data suggests that the vast majority of lifecycle costs are for activities other than actual storage of the bits. This would seem to suggest that the core cost of simply ensuring that the bytes persist over time is an extremely small proportion of overall data management expenditure.

8. BENEFITS TAXONOMY AND BENEFIT CASE STUDIES

8.1. 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. Much of the literature addressing economic issues related to digital preservation focuses on the cost side of the cost/benefit equation. Comparatively little attention is paid to articulating the benefits to stakeholders arising from the preservation activity. Instead, 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.).

Serious analysis of the economic feasibility of prospective digital preservation projects requires projected costs to be weighed against expected benefits. 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.

Several recent studies – e.g., Beagrie, et al. (2008) and Fry, et al. (2008) have addressed the question of benefits arising from the long-term preservation of research data. Both studies articulate a diverse set of benefits that can potentially accrue from ongoing accessibility to research data sets. Also Currall and McKinney (2007) aims to identify the intangible benefits of digital preservation. Despite the challenges associated with actual measurement of the benefits from digital preservation, 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, it is useful to frame out 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 to organise thinking about

preservation benefits, and may provide some insight into how generic expressions of preservation benefits can be sharpened into more focused value propositions.

A taxonomy for categorising the benefits from long-term preservation of research data is presented below. The taxonomy is illustrated with examples from case studies drawn from the experiences of the UK Data Archive, the National Crystallography Service, and the University of Oxford. Lead authorship of a case study is by the contributing partner institution with comments and additional analysis by the team of three lead authors.

8.2. BENEFITS TAXONOMY- SUMMARY

Dimension 1	
Direct Benefits	**Indirect Benefits (Costs Avoided)**
New research opportunities	No re-creation of data
Scholarly communication/access to data	No loss of future research opportunities
Re-purposing and re-use of data	Lower future preservation costs
Increasing research productivity	Re-purposing data for new audiences
Stimulating new networks/collaborations	Re-purposing methodologies
Knowledge transfer to industry	Use by new audiences
Skills base	Protecting returns on earlier investments
Increasing productivity/economic growth	
Verification of research/research integrity	
Fulfilling mandate(s)	

Dimension 2	
Near Term Benefits	**Long-Term Benefits**
Value to current researcher & students	Secures value to future researchers & students.
No data lost from Post Doc turnover	Adds value over time as collection grows and develops critical mass
Short-term re-use of well curated data	
Secure storage for data intensive research	
Availability of data underpinning journal articles	

Dimension 3	
Private Benefits	**Public Benefits**
Benefits to sponsor /funder of research/archive	Input for future research
Benefits to researcher	Motivating new research
Fulfil grant obligations	Catalysing new companies and high skills employment
Increased visibility/citation	
Commercialising research	

Figure 8: Summary Overview of the KRDS2 Benefits Taxonomy.

8.3. BENEFITS TAXONOMY -DETAILED DESCRIPTION

Dimension 1: Direct Benefits and Indirect Benefits

Direct benefits are what most people think of when they think of preservation benefits – that is, positive statements about the value created by maintaining persistent access to digital materials. For example, we might say that preservation of a certain set of research data permits future scholars to undertake particular forms of scholarship (conversely, of course, we can say that failure to preserve the data would mean certain forms of future scholarship would not be possible). Other examples of direct benefits include transfer of knowledge from current researchers to future researchers, increases in research productivity from using well-curated, easily accessible data; and the coalescing of new disciplinary and inter-disciplinary networks of collaboration around key research data sets. One can imagine circumstances where the long-term preservation and accessibility of research data could diminish obstacles to the commercialisation of scientific discoveries, leading to the formation of new companies, increased demand for highly-skilled workers, and higher levels of productivity and economic growth. Direct benefits from digital preservation might even include fulfilment of mandated data preservation obligations attached to a funding award. In general, direct benefits take the form of a value proposition along the lines of "if preservation occurs, an outcome will occur which is of value to some group of stakeholders."

The use of the word "outcome" is important in the statement above. It is important to keep in mind that when stating the direct benefits of digital preservation, the focus should be on the outcome from preservation, not the process. Preservation is not by itself a desired outcome; it is a process by which preserved digital objects (i.e., preserved data sets) are produced. It is the value-generating activities associated with use of preserved research data that is the true outcome of preservation, and the source of the "return on investment" to preservation. Consequently, a compelling value proposition for digital preservation is more than just a commitment that certain digital objects will persist over time; rather, it should articulate as plainly as possible the sorts of value-generating outcomes that can be realised through the ongoing availability of the preserved digital objects.

Both the UKDA and NCS observe direct benefits extending primarily from the opportunity for ongoing access to, and use of, preserved research data. For example, NCS notes several

particular benefits following directly from preservation of research data, including the transfer of knowledge from current to future researchers; and increased probability of commercialisation of scientific knowledge. UKDA emphasises that availability of preserved research data creates direct benefits in terms of verification of past research and motivation for (and input to) new research, but notes several important nuances in characterising these benefits. First, it is difficult to assess benefits based on accessibility alone: more specifically, the fact that a data set was accessed or downloaded does not necessarily mean the data was actually used. Given this, a more concrete measure of benefits associated with ongoing availability would be evidence that demonstrated use of a preserved research data set: for example, citations in scientific papers or even popular media such as newspapers. Second, it is important to note that even if a preserved data set has not been accessed or used, it does not necessarily follow that it has no value; an implicit value still arises from inclusion of the data in the permanent scholarly record, and there is always a possibility that future use will occur. However, a value proposition for digital preservation is more compelling when based on demonstrable use of the preserved content, rather than the possibility of future use. Finally, it is important to understand patterns of use for preserved research data when assessing benefits. Intensity of use at a particular point in time may not always be an accurate indicator of long-term future value. For example, some data sets may enjoy heavy usage, but only for a relatively short time period, while others may exhibit a comparatively low rate of use, but one that persists steadily over long periods of time. All of these factors must be taken into account when assessing the direct benefits realised from long-term preservation of research data.

Indirect benefits are another form of benefit that can potentially emerge from digital preservation. They are best understood as “costs avoided.” For example, investing now in the preservation of a particular research data set might be justified on the basis that if the data were allowed to disappear, re-creating it at a later time would be extremely – and possibly prohibitively – expensive (e.g. see costed examples for data creation and data loss in the NCS case study). In these circumstances, investment in preservation now avoids a larger cost sometime in the future. Of course, the validity of this argument rests on the likelihood that future demand for the data will in fact materialise. Even if the data is in no imminent danger of disappearing, engaging in curation activities early in the digital life cycle

may be a less expensive strategy than postponing them until the future. The costs of preserving uncurated data long after it was originally created can be quite costly; retrospective metadata creation, for example, is often extremely expensive.

Indirect benefits can also be framed from the perspective of protecting earlier investments in research and digital collection development. Universities and other institutions invest significant sums in acquiring and/or developing digital assets, or more generally, funding the research activities that produce these assets. Failure to provide resources for the ongoing maintenance of important research outputs – i.e., failure to ensure that the outputs persist in a state such that they continue to release value to their users – reduces the return on the original investment in creating and/or acquiring them. Research data sets and other digital assets are durable goods; that is, they can continue to generate value over extended periods of time. Just as resources are allocated toward the ongoing maintenance of other durable goods like houses or automobiles, it is important to provide for the ongoing maintenance of expensive investments in research and research outputs.

In general, indirect benefits represent situations where incurring a preservation cost now diminishes the likelihood of incurring an even larger cost sometime in the future. The indirect benefits of digital preservation can be as compelling as the direct benefits, and should not be overlooked. The experiences of UKDA and NCS provide useful illustrations of this point.

Much of the social science data managed by UKDA is, for all intents and purposes, unique; very little of this data can be re-created should it be lost. For example, a data set like the General Household Survey for 2001 could be replaced in the sense that a new project could be launched that repeated the exercise of collecting the data contained in the original survey (at a cost of roughly £500,000). However, such an exercise would not re-create the 2001 data; it would replace it with new data of a similar nature. The 2001 data can never be precisely replicated. However, UKDA also notes that there are circumstances where it may be less expensive to re-create data than to preserve it; for example, in the context of a recent project involving scanned page images, UKDA determined that the cost of carrying out complete preservation of these images was more expensive than re-scanning the materials in the future if needed.

NCS provides some interesting data on the indirect benefits of archived crystal structures:

- Depending on the original storage medium, long-term preservation (essentially byte storage) of the raw data of a crystal structure can range from £21.95 to as little as £1.48;
- NCS also preserves "results data" based on analysis of the raw data, at a cost ranging from £30 to £2.15;
- In general (unlike the UKDA social science data) results data can be regenerated; however, the costs of doing so will vary enormously depending on whether the raw data has been preserved or not (£50 to £400 if it has; as much as £20,000 if it has not);
- Preservation of the results data avoids a significant future cost. More generally, preservation of both the raw data and the results data provides a dual hedge against incurring substantial future costs: by preserving the results data, the costs of reproducing it are avoided; by preserving the raw data, the costs of reproducing the results data are minimised in the event that re-creation is unavoidable.

Dimension 2: Near-term Benefits and Long-term Benefits

Another dimension along which the benefits from digital preservation can be characterised is the time horizon over which they are projected to be distributed. Typically, the benefits from preservation are assumed to be long-term in nature; in some cases, there is an implicit assumption that the benefits are conferred exclusively on future generations of stakeholders. The implication then becomes that current decision-makers incur the costs of preservation, while future generations reap the benefits of the investment. While this may be true in some circumstances, it is likely that in others the distribution of preservation benefits over time will be more nuanced.

Digital materials are fragile in comparison to other media. If neglected, it is possible that important research data sets, along with ancillary materials such as data documentation, digital lab notebooks, and so on, may become corrupted or simply disappear within a very short time span. These materials may be of immense value to current researchers and students, if steps were taken to preserve them. In this sense, preservation confers benefits

on today's stakeholders, as well as future stakeholders. In framing the benefits of digital preservation, it is useful to consider how these benefits impact the current array of stakeholders.

Our costs case study at the University of Oxford (see section 7.5) also highlighted a number of ways digital curation/preservation services can offer tangible, practical benefits to current researchers. For example:

- Some research centre directors noted that the constant turnover of post-doctoral researchers often resulted in lost data, since there are currently no established mechanisms to routinely collect and organise the data these post-doctoral researchers generate;
- In some cases, researchers generated data several years ago and now could not make sense of them as they had not kept enough information on how the data was created in the first place. In these circumstances, well-curated data would have clear short-term benefits;
- In scientific disciplines, research groups require secure storage for their large volume of data generated by instruments such as electronic microscopes or by computing simulations run in GRID systems;
- Some clinical research centres compile data for decades and spend months migrating data formats in order to avoid format obsolescence;
- In many cases, researchers want to make their articles' accompanying data available online in a sustainable way and they do not have the institutional infrastructure to do this, so they just publish the data on their departmental website.

In the UKDA's experience, the cost of providing near-term access overlaps considerably with the costs of preparing research data for long-term preservation and access. Certain costs are incurred independent of the length of time the data is to be kept; the additional costs to preserve the data over the long term represent an increment over these costs. Given this, much of the cost that on the surface appears to be an allocation for preserving materials over the long term is in fact expended to provide short-term access as well. Hence, relatively

little additional work has to be carried out to gain both short and long term benefits related to ongoing access.

NCS notes that near-term benefits for the individual researcher are an important element of their data preservation activities, in the form of:

- ongoing access to raw data during an experiment;
- the ability to fulfil funders' mandated data deposit requirements;
- and the establishment of a chain of provenance for their research data and subsequent findings.

NCS also notes a long-term benefit (i.e., distinct from short-term benefits) in regard to preservation of embargoed data. In this case, there are no short-term benefits from preservation. However, it is still important to make the necessary current preservation investments in order to realise the future benefits; well-curated data is much easier to make accessible and use in the future than data that has been neglected.

In general, near-term benefits and long-term benefits are not mutually distinct, but instead intrinsically connected. Ensuring that important research outputs persist over the near term and are available to today's researchers, is a necessary condition for securing the opportunity to generate long-term benefits and making them available for tomorrow's researchers.

Dimension 3: Private Benefits and Public Benefits

The first two dimensions deal with the questions "what are the benefits?" and "when are the benefits realised?". The last dimension deals with the question "on whom are the benefits conferred?". As a general matter, benefits from preservation can be classed into two broad categories: those that accrue to the direct constituents of the entity that sponsors and/or pays for the preservation (private benefits), and those that extend to the community at large (public benefits). For example, a curated and preserved research data set may generate private benefits on several levels: first, it may fulfil the individual researcher's grant obligations to deposit the data in secure storage; second, 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.

But preservation of the research data set may also confer benefits on the wider academic community, in particular by motivating and serving as input for future work by scholars at other institutions. In this sense, the institution preserves the research data set for use not just by its direct constituents, but for the benefit of scholars and learners everywhere. Public benefits should not be overlooked when characterising the value returned to an investment in digital preservation. These public 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.

The appropriate mix of public and private benefits can be an important element of a compelling value proposition for digital preservation. In some cases, decision-makers may be primarily concerned with the private benefits of preservation; since the investment is being made by a particular institution, there is an expectation that the benefits should accrue primarily if not exclusively to that institution. On the other hand, many mission-driven institutions consider themselves tightly embedded in broader networks of collaboration and collective interest. In this case, contributions to the "greater good" may be valued, in addition to the private benefits that accrue directly to the individual institution. In any event, decision-makers should be aware of and consider carefully the nature of the benefits a digital preservation investment confers along the private/public dimension.

The UKDA generally considers itself to have three categories of stakeholders, or direct constituents:

- Users of data which UKDA preserves on behalf of others;
- Creators of data who wish to ensure that their research outputs are accessible and verifiable in the future OR (in the case of government departments) those whose data has public re-use value and can be shared at no cost to the creator;

- Investors in research who wish to provide access to data which they have funded (e.g., ESRC research data) and which is likely to be valuable to their research community for re-use.

The stakeholder categories are not mutually exclusive. Beyond these categories, however, UKDA has discerned evidence that the benefits from data preservation sometimes extend more widely to the general public. For example, there is evidence that many of the UKDA data sets have been re-purposed as pedagogical tools to assist in certain forms of learning. This is a specific instance where a digital curation activity set up to serve one set of constituents nevertheless has "unintended consequences" in terms of providing benefits to other communities as well, such as researchers in other disciplines or even the general public. NCS also perceives wider public benefit emerging from its digital preservation activities, in the form of a transfer of knowledge across time and space that can be used to validate past research and motivate new research. These benefits, which extend beyond UKDA's and NCS's perceived direct constituencies, should be noted when articulating the value proposition for digital preservation.

The three dimensions of direct/indirect, near-term/long-term, and private/public are intended to help organise thinking about the nature of the benefits associated with investments in the long-term preservation of research data, in order to better assess their relative weight in comparison to the cost of the preservation investment itself. Clearly, there are other dimensions that might be added to this list, and much more work needs to be done to characterise specific examples of benefits within each category. In addition, quantification of many of these benefits is difficult, and in some cases, impossible; however, articulation of benefits even in just a qualitative way can help raise awareness on the part of funders and other decision-makers. At the least, this taxonomy will hopefully encourage a deeper understanding of the nature of the benefits long-term preservation of research data can offer, and in doing so, help clarify the benefit side of the cost/benefit equation.

8.4. Benefits Case Study: National Crystallography Service, Southampton University

This benefits case study on research data preservation was developed from longitudinal cost information held at the Department of Chemistry in Southampton and their experience of data creation costs, preservation and data loss profiled in Keeping Research Data Safe (Beagrie et al 2008). A fuller description of the background, methodology and a breakdown and analysis of costs presented is available as NCS benefits study supplementary material (http://www.beagrie.com/KRDS2_NCS_benefits_supplementary.doc) on the KRDS2 project website.

This case study covers several primary activities in the Pre-Archive and Archive phases of the KRDS2 model, namely:

- Initiation: Project design & Data management plan;
- Creation: Generate descriptive metadata, Data management & Create submission package for archive;
- Acquisition: Selection & Depositor support;
- Disposal: Transfer to another archive & Destroy;
- Ingest: Receive submission.

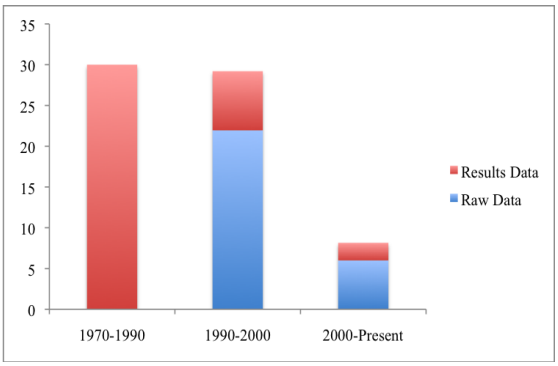
A comparative study of the costs to a) preserve (in original storage format) and b) migrate (to new storage format), data collected on the National Crystallography Service (NCS) from the longitudinal time period of 1970-2009 is presented in this benefits case study. During this timeframe experimental instrumentation, computational capability, and data storage media (e.g. paper, digital video disc (DVD), robotic tape store) have radically changed. When considering these elements of change one can roughly group transitions between technologies e.g. the introduction of personal computers, a new generation of instrumentation, or the advent of online storage, to fall into three roughly similar periods (1970-1990, 1990-2000 & 2000-present).

The outcome of an NCS experiment is a crystal structure, which is the product of collecting raw experimental data and processing it into results data – cost data presented throughout are those relating to the generation of a single crystal structure. It is important to note that

these data are taken as a 'snapshot at a point in time' i.e. at the time of writing, as the migrations (and therefore the costs ascribed to them) are all priced at that point.

The most pertinent costs from the study are depicted below.

(a)



(b)

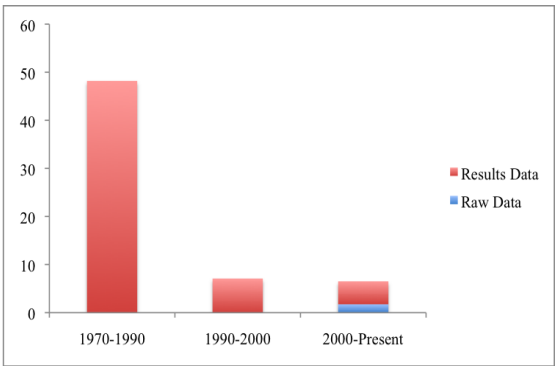


Figure 9 (a): Relative and total costs (£s) per dataset of preserving raw and results data (Note: 1970-1990 it was not possible to store & preserve raw data); (b): Relative and total costs (£s) per dataset of migrating raw and results data (Note: It was not possible to migrate raw data acquired during 1970-1990 & 1990-2000).

It is important to note that the cost to generate a structure with current equipment is £328, however the cost to recreate a structure from the 1970 and 1990 periods is around sixty times this amount, c. £20K (see NCS costs in KRDS1, Beagrie et al 2008). The reason for this is the differentiation between raw and results data: as with most experimental science, these are treated differently in terms of data management and preservation. The cost of recreating historical data is defined by the need to re-synthesise the sample from which that data were generated, which includes all the expertise and laboratory infrastructure from an entire research project – it is not simply a matter of "doing the experiment (or analysis) again". The reason the sample needs to be re-synthesised is that it has not been possible in these eras to store and preserve raw data. In more recent times raw data can be preserved, in which case the cost of recreating the data is that involved with the (re)interpretation of the raw data. The most obvious points from these data are that:

- The cost of preserving data has dramatically reduced;
- The cost of migrating data from recent eras when computing has been more prevalent is significantly less;

- The cost of preserving raw data is around 70% of the total (raw + results) data preservation cost.

It is therefore a noteworthy conclusion that the preservation of raw data, as opposed to results data, is the significant factor in crystal structure data preservation.

This study was concerned with capturing accurate costs for the migration of historic data across media as a preservation exercise. Again, the differentiation between raw and results data is made (however it is not possible to capture costs for raw data in the first two eras, as historically it was not possible to store it) and results are summarised in Figure 10.

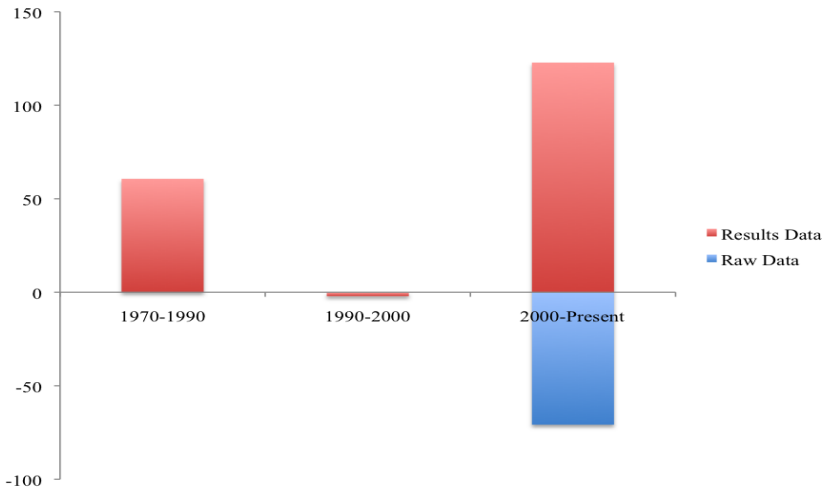


Figure 10: Migration Costs as a Percentage of Preservation Costs (see NCS benefits study supplementary material, Table 1, for detailed figures).

Migration costs for different original media from particular eras are represented as a percentage of the cost of preservation of the same data from that era. That is, negative values indicate that the cost of preservation outweighs that of migration and vice versa. This therefore tells us that:

- It is more expensive to migrate raw data than to preserve it. This is due to the fact that these data are large in volume and the formats cannot be migrated through transformation due to the proprietary binary format – therefore the only possible actions are copy or destroy. This factor must be balanced against the fact that without raw data it is not possible to regenerate the results and therefore the effective

cost of this is £20K. Therefore it is recommended that raw data is preserved, appraised regularly and either migrated or destroyed;

- It is considerably more costly to migrate results data than preserve them. This is due to the variety of formats and the storage media used over the years. With modern approaches the preservation of results data is becoming well understood and addressed and it is recommended that these routes be taken;
- There is considerable fluctuation in the relative cost of migration against preservation with different eras (that is media, data types, instruments etc) and it does not necessarily follow that modern (or indeed any era) approaches make it cheaper to migrate as opposed to preserve with respect to other eras;

Migrating raw and results data highlighted some important points regarding data loss:

- During migration of raw data from CDs/DVDs to online storage there was a 7% loss of data: in principle this corresponded to a financial loss of £2.8 million, due to the fact that to recreate the data the samples would have to be entirely re-synthesised;
- Migration of results from floppy disks resulted in a 5% data loss, with a perceived financial loss of around £2 million for the same reason as above;
- Less than 1% of results were lost in the migration from paper, however the cost of that migration was extremely variable (depending on archive quality).

The stakeholder benefits that have been highlighted by this study are mainly counter factual and can be aligned to the KRDS framework taxonomy as figure 11 below:

NCS Stakeholder Benefits	
Dimension 1	
Direct Benefits	**Indirect Benefits (Costs Avoided)**
increased research productivity and successful outcomes arising from implementing correct and useful metadata for preservation; transfer of knowledge about the process from current to future generations; larger contribution to the body of knowledge; knowledge transfer resulting in increased commercialisation of discoveries; fulfilling funders mandates;	understanding of the counter factual aspects of the 'what if' scenarios that this study presents; protection of earlier investments.
Dimension 2	
Near Term Benefits	**Long-Term Benefits**
For researchers: an ability to return to raw data during the analysis; an ability to provide a provenance chain to the raw data for validation in the early stages of dissemination; increased visibility of their research outputs;	Preservation of embargoed or unpublished data (currently estimated to be around 80% of research outputs); It is considerably cheaper to return to well curated data long after collection and make it public; the ability to reinterpret data with next generation software.
Dimension 3	
Private Benefits	**Public Benefits**
the ability to manage personal research data for the future, so that it may be exploited at a later time;	increased value for money; increased knowledge transfer; growth of the body of data available for mining and new science.

Figure 11: NCS Stakeholder Benefits - KRDS2 Tabulated Summary.

8.5. Benefits Case Study: UK Data Archive

Measuring the impact (outcomes) of investment (costs) in preservation projects and services is an area of increasing interest across all sectors. It is also a major area of interest to UKDA and this benefits case study focusing on social science and historical datasets was developed as part of their contribution to the KRDS2 project. This section is structured around the KRDS2 benefits taxonomy and is intended to provide some concrete examples about the benefits accruing from digital preservation at the UK Data Archive. Each section is preceded by a short summary from the taxonomy.

The benefits taxonomy commences with a rehearsal of an argument about digital preservation analysis which suggests that the costs of preservation have been worked endlessly, but that any good economic analysis of digital preservation activities should be complemented by a discussion of the benefits of those activities. Benefits of these activities, it suggests, are either considered to be common knowledge or framed in such a generic way as to be impossible to use for real decision making purposes. A possible reason for these generic approaches is that it is particularly hard to measure these benefits in any quantitative way and so they have been ignored. In order to present some of the benefits of digital preservation KRDS2 has produced a taxonomy of benefits with the view that by being expressed in more formal terms they may "provide some insight into how generic expressions of preservation benefits can be sharpened into more focused value propositions." The aim of this benefits case study is to present some of the benefits accrued by preservation at the UKDA within this framework in order to assist in this process.

Dimension 1: Direct Benefits and Indirect Benefits

1. Direct Benefits

The related value proposition for Direct Benefits is: "if preservation occurs, an outcome will occur which is of value to some group of stakeholders." In one sense, most benefits relating to digital preservation at the UKDA are ones which are of direct value to one group of stakeholders or another.

Access

The major direct benefit of digital preservation for the UKDA is that material which was created well into the past remains accessible. Access/Accessibility should not be mistaken for re-use, which is discussed below, the term is used to denote the possibility of access rather than its occurrence. Access can be viewed as “potential value through re-use”. The value of this benefit is not necessarily able to be calculated; consequently a real cost-benefit analysis cannot occur either. The fact that someone downloads a dataset which had been deposited at the UKDA ten years previously does not necessarily mean that it has value to someone, but it allows us to assume that there is some perceived value as otherwise they wouldn’t download it in the first place. The fact that a study is downloaded after preservation has taken place is de facto evidence that there is value in preserving the data in the first place. (However, we should not assume that because something hasn’t been downloaded yet that it will not have value to a user in the future.)

Re-use

A more common direct benefit from continued access to data at the UKDA is the ability of researchers to use data which they did not create themselves and may not otherwise have had access to. Re-use may be typified as “actual value realised”. While the provision of access per se is a benefit, re-use is a much more concrete benefit, since re-use can often provide demonstrable evidence that data was in fact used productively. Evidence for re-use is not just found in citation of academic papers, but within the newspapers. A recent example is at http://news.bbc.co.uk/1/hi/health/8278742.stm . The Millennium Cohort Study used by the researchers was accessed via the UK Data Archive.

The ability to re-use data can be a benefit both in the short term and in the long term. 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 the provision of 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 and some of this research may in itself have contributed to public policy.

Audience

Future use, actual or possible, is a simple benefit of any series of preservation activities, but these are not necessarily quantifiable, and the relative importance of any "use" does not necessarily have to be related to the relative use. It could potentially be related to the cost of production (see indirect benefits below). An example relates to two studies which were "published" by the UKDA on the same day (23 April 2002); the Genevan Sex Crimes Database, c.1440-c.1790 [SN 4364] and the Health Survey for England, 2000 [SN 4487] (commissioned by the Department of Health). From publication to June 2009, the former had been downloaded nine times, the latter 1,513 times. On the basis of use alone, the 2000 Health Survey has had 150 times more impact than the Genevan Sex Crimes database. If use was related to audience, the latter's usage would seem quite reasonable. Amongst historians of sexual deviance the Health Survey of 2008 is unlikely to have registered much of a blip on their radar! Furthermore, the Genevan Sex Crimes database is likely to provide low but consistent usage over the long term, whereas the Health Survey of England is likely to already have peaked in annual downloads: (400 in 2007 and 141 in 2008). Direct benefits associated with the quantification of usage must be tempered by an understanding of the audience. For example, in this case because the population/target audience of historians interested in sexual deviance is considerable smaller than that of health researchers, the level of direct benefits to all health researchers may be lower than those for all historians interested in sexual deviance.

Academic activity

Direct benefits can also be measured in terms of academic activity. How much research activity is engendered as a result of a study being available? Problems associated with quantifying this form of benefit are a) lack of proper citation; b) uncertainty in measuring "value" of the research itself; c) the time-lag between use and publication and d) the inability of an archive to measure this activity in the first place.

Public policy example

The following paragraph is copied from an internal document on the impact of the Economic and Social Data Service (ESDS). ESDS is one of the services run by the UKDA. ESDS is jointly run at University of Essex and University of Manchester.

"Societal Impact: One of the advantages of archiving data over many years is that long time series of consistent data are built up. Richard Berthoud, of ISER at the University of Essex, has analysed the GHS [General Household Survey] between 1974 and 2005, to describe changing patterns of advantage and disadvantage in employment. A headline finding is that patterns of disadvantage are not fixed – the employment rate of mothers steadily improved over the period, while disabled men's chances of work steadily deteriorated. The initial analysis, undertaken for the Equalities Review, 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." (ESDS 2010).

The major direct benefits ensuing from digital preservation activities at UKDA are: availability of data and its potential for re-use. Social science data can be re-used to inform research and thus public policy. Re-use should also be considered as having two dimensions: verification (i.e., re-use for the same purpose as creation) and new research (re-use for a different purpose to that for which it was originally created).

2. Indirect Benefits

A related value proposition for Indirect Benefits is: "if preservation occurs, what costs can be avoided in order to ensure an outcome which will be of value to some group of stakeholders."

The most straightforward example of an indirect benefit is: "what will it cost to preserve something now in order to avoid the costs of recreation". This proposition is based on the argument that there will be demand for whatever is preserved into the future.

Uniqueness of social science data

Most social science datasets are unique; the UKDA holds only a very small quantity of experimental data which could be repeated, rather "snap-shots" in time. For most of the UKDA's holdings it is not possible to recreate datasets as they are based on unique one off surveys. Consider the annual General Household Survey. The 2001 wave of this survey told

us, amongst other things, that household size was declining slowly, that the prevalence of home ownership and cigarette smoking was flattening out, male employees were less likely to have an employer's scheme pension, but female participation in the same schemes were increasing (Walker et al 2002).The cost of the creation of this dataset is subsumed within a total cost of the GHS (in 2001) which was reported by the National Statistician as being £1.43 million. This figure covered "analysis and reporting for 2000-01, fieldwork for 2001-02 and planning and preparation for 2002-03." (UK Parliament 2001). We can reasonably estimate that the replacement cost for this dataset would be over £500K, but since the results of any replacement would be relating to a different period in time, it would only be a replacement rather than a recreation.

The value of the survey, that is the information which it provided, was worth at least its cost to the Office for National Statistics, and as it was the thirtieth wave of this survey the value was probably higher because it provided another time point in a series of surveys. It is unlikely that the 1,154 researchers who downloaded this dataset from the UKDA (2003-2008 only) would have either been able to afford to recreate the survey or would have wanted to download it if they were supposed to share in its cost of construction.

However, knowing even roughly how much a government survey costs is unusual, and attempting to estimate the proportion of a research grant which is devoted to the collection and management of a dataset is fraught with problems.

In the case of the UKDA indirect benefits are perhaps more clearly identifiable, and potentially possible to be costed (at least in a counter-factual manner). For example, in the case of a single project which the UKDA carried out, the complete preservation of a series of approximately 200,000 scanned images was considered to be more expensive than the cost of re-scanning at a time in the future. The images are backed up; their related metadata is preserved. If the all four separate backups were all to be destroyed accidentally or the TIF format of image files was to become obsolete, or no tools were available to transform this file format into another format, then re-scanning would have to take place. These risks are low, and the "data" is not unique, being derived as it is from printed texts. However, this is not strictly speaking a benefit of preservation, rather a benefit of thinking about the costs of preservation and the costs of re-creating some material.

Re-purposing (data)

However there are less commented upon benefits accruing from preservation. The first is the possibility of repurposing the data for different use at some time after its creation for a different audience. The UKDA produced a cut-down version of the British Crime Survey of 2000 entitled British Crime Survey, 2000: X4L SDiT Teaching Dataset (SN 4918). Between 2004 and 2008 inclusive this study has been downloaded almost 7,500 times. Year on year use has increased annually. It was repurposed as part of a project for Survey Data in Teaching and was designed to be used by A level school children, but is used for both undergraduate and postgraduate research. The availability of data was partially dependent on the UKDA already holding earlier waves of the British Crime Survey and preserving them. So, a direct outcome of preservation of the 2000 British Crime Survey at the UKDA was the provision of a well used teaching dataset based on it. (This could also be an indirect benefit.)

Re-purposing (methodology)

A further example of an indirect benefit relates to a dataset which was first delivered to the UKDA in 1979. This is the National Sample from the 1851 Census of Great Britain. The documentation which related to this dataset, and was preserved alongside the transcription of person information from the 1851 census, was hugely influential in the manner in which a much later accession (the 1881 census returns) was processed and prepared for public access. The earlier investment did not make the costs of creation or preparing for dissemination the later dataset any less, but it improved and informed the research process surrounding those activities. Had the preservation of the 1851 data not been done, the re-purposing and ingesting of the 1881 data may have been done in a very different way.

Re-purposing (data and methodology)

Another benefits example which can't be costed is as follows. Research is currently being carried out at the Centre for Socio-Cultural Change in the University of Manchester which explores people's experiences of family and parenting practices to give insights into the nature of social change and continuity over four decades. The researchers are "explicitly investigating the methodological use of qualitative secondary analysis and are basing their research on a number of 1960s archived classic community and family studies" preserved at the UKDA, including Dennis Marsden's Mother's Alone [SN 5072] and Peter Townsend's

Poverty in the United Kingdom [SN 1671]. This research could not be carried out had the original research data and information relating to its creation and methodological underpinning not been deposited at the UKDA.

Use by non-target audience

Another indirect benefit of preservation is that the UKDA can make some documentation freely available to all comers. Thus the study Workplace Employee Relations Survey: Cross-Section, 1998 [SN 3955] has been downloaded almost 1,000 times since 1999, however the supporting documentation was downloaded 10,500 times in the first half of 2007 alone. Thus supporting documentation which may seem to be only valid to the users of the data has a much larger potential audience from a wider group. Multiple modes of access to material and the ability to keep digital copies available easily (consequence of the digital preservation process) has meant that many more people have been able to use the documentation than they might otherwise have, and users from outside of the core stakeholder community. Thus, these benefits are not only indirect but are public benefits too (see below).

While many of these direct and indirect benefits cannot be quantified easily, if at all, the qualitative evidence provided in examples like those above provide, at the very least, an awareness raising function.

Dimension 2: Near-term Benefits and Long-term Benefits

KRDS2's taxonomy makes a fair distinction between near and long term benefits. It argues that while typically benefits of digital preservation are conferred exclusively on future stakeholders there are benefits to current stakeholders as well.

At the UK Data Archive, and across the social science data archives, most of the data collections would not be accessible at all unless they had been ingested and prepared for preservation by those archives. Access could be provided to certain datasets without preparation for preservation (like www.data.gov/catalog/raw) but the benefits of access would only be short term. At present the time and cost taken to prepare a Dissemination Information Package from a data submission is only slightly lower than the cost to prepare an Archive Information Package as well. The additional "marginal" costs of storage,

migration and refreshment is the actual cost of preservation while the ingest costs can be construed simply as access costs. Thus the majority of users of data lodged at the UK Data Archive are reaping the near-term benefits of preservation activities.

Near and long term use

Evidence from the UKDA for long-term benefits is harder to demonstrate. However, one potentially interesting example is the British Election Study, February 1974; Cross-Section Survey [SN 359] which was deposited with the UKDA in late 1975. Usage of this study shows an almost bimodal distribution with peaks in 1978 a couple of years after it was first made available, and in 2007.

The overall usage of this study is relatively low, but there is a clear resurgence in interest in this study since around 2002 (coincidentally (or not) the same time as UKDA started doing on line downloads). If long-term preservation techniques had not been in practice in those early years and the major costs involved in ingest had not been undertaken in the 1970s, it's possible that in 2000/2001 a decision may have been made to de-archive the study. As this wasn't the case, the study was able to benefit from a new lease of life from 2002. So, the benefit of an integrated and long-term preservation strategy in the 1970s has meant that users who were potentially not even born when the material was ingested are able to access and use these data. It is also worth considering that the main user domain base of any dataset can change over time. Some of the more recent users of this dataset work or study in departments of history.

It is important to show that the benefits of digital preservation at the UKDA occur both in the near term and in the longer term. Preservation activities can be understood to be a superset of data delivery activities. Hence (relatively little additional) work has to be carried out to gain both short and long term benefits.

Dimension 3: Private Benefits and Public Benefits

The third dimension invoked by KRDS2 relates to whom the benefits are conferred. The benefits taxonomy distinguishes between public benefits (the community at large) and

private benefits ("to the direct constituents of the entity that sponsors and/or pays for the preservation").

In general the UKDA can say that it has three major stakeholders:

- Users of data which we hold and preserve on behalf of others;
- Creators of data who wish to ensure that their research outputs are either accessible or verifiable in the future OR (in the case of government departments) those whose data has public re-use value and can be shared at no cost to the creator;
- Investors (who are sometimes creators) in research who wish to provide access to data BOTH which they have funded themselves (e.g., ESRC research data) AND which is likely to be value to their research community for re-use.

A fourth group of "stakeholders" is often ignored: the wider public. As shown above documentation relating to datasets is heavily downloaded: in no month in 2008 did monthly documentation downloads fall below 200,000 items.

Users' benefits are generally understood to be direct benefits, but the fact that the UKDA has a large collection of datasets means that it can act as a single point of entry for a large number of datasets. Over 75% of users who downloaded data in 2008 downloaded more than one study as can be seen in the following:

Number of downloads	Number of users	%
Only 1	1,507	24.6
2-5	2,473	40.5
6-10	895	14.7
>10	1,232	20.2

Figure 12: UKDA Downloads in 2008.

This benefit is not a benefit of preservation per se, rather a benefit occurring from the organisational form of preservation chosen. Providing access to a collection of related datasets, which are made accessible in the same way and with the same protocols, is clearly of value to the end-user.

Creators gain a “free” publishing outlet for their data from a trusted and respected repository. Promotion (and enhancement) of datasets confers an additional level of visibility for organisations and researchers. Re-use of quality data also confers additional benefits to the user through citation.

From the investors’ point of view, ensuring the provision of research data, within a controlled environment, can be costed. It must be considered to be (at least) equivalent to our direct funding. Furthermore, a direct benefit of the investor taking this enlightened approach over a period of 40 years has meant that the UKDA has been able to contribute not only to the creation and analysis of data within the social science community over this period, but it has been able to make a contribution to modes of teaching social science, best practice in researcher data management, standards in digital preservation practice, digital thesauri, etc. These are spin-off benefits which relate to the process of digital preservation, and are the outcomes of the experience and practice of digital preservation.

There are other “spin-off” benefits, which are not directly related to digital preservation but to data curation more widely. The UKDA provides advice and expertise to a variety of national and international advisory bodies. The UKDA, in conjunction with sister archives have clearly been influential in the development of data curation and digital preservation. Immersion in data curation has allowed the UKDA to provide best practice guidance on data sharing and management across the UK Higher Education/Further Education sectors.

9. CONCLUSIONS AND RECOMMENDATIONS

In January 2009 JISC issued an ITT for a study on the identification of long-lived digital datasets for the purposes of cost analysis. The aim of this work was to provide a larger body of material and evidence against which existing and future data preservation cost modelling exercises could be tested and validated. The proposal for the Keeping Research Data Safe 2 (KRDS2) study was submitted in response by a consortium of partners who provided significant in-kind contributions to allow a wider exploration of costs and benefits in the study.

With a relatively modest budget significant achievements have been delivered. Our main findings have been:

Long-term Costs of Digital Preservation for Research Data:

- Although there are disparities between our cost case studies reflecting very different disciplines and missions some consistent broad trends and findings exist (see section 7);
- The costs of archiving activities (archival storage and preservation planning and actions) 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 in KRDS2. As an example the respective activity staff costs for the Archaeology Data Service are Access (c.31%), Outreach/Acquisition/Ingest (c.55%), Archiving (c.15%).This confirms and supports a preliminary finding in KRDS1;
- Some potential opportunities for cost savings and further automation of archive tasks were noted which could be investigated further. Our work suggests the greatest potential cost benefits could arise from future tool development in ingest and access activities;
- “Fixed costs” have a significant impact in our case studies. This largely relates to staff (in particular the minimum viable staffing and skill sets needed to maintain reliable services);

- Economies of scale can be demonstrated in several of our case studies and relate to our observation of fixed costs: once core capacity is in place additional content can be added at increasing levels of efficiency and lower cost.

Benefits of Preserving Research Data:

- A benefits taxonomy has been produced (see section 8) and illustrated with two detailed benefit case studies (see sections 8.4 and 8.5);
- We have recognised that the identification and promotion of the "near term benefits" are particularly important in advocacy to researchers: we can show in our benefit case studies and also our costs work at Oxford (section 7.5) that there are significant benefits in the short-term to current researchers as well as long-term benefits to future research;
- Our benefits case study with the National Crystallography Service and Department of Chemistry at the University of Southampton (section 8.4) has demonstrated the calculation of indirect benefits (costs avoided or counter-factual arguments) for data loss. It highlights the importance of being able to identify costs for pre-archive as well as archive phases of the data lifecycle, to achieve this;
- Our benefits case study with UKDA (section 8.5) illustrates a range of benefits to its stakeholders. Some of these may not have been widely recognised before. For example, the fact that the re-purposing of the methodology as well as data and the use of documentation of a dataset as well as the data itself can be significant: an example is given for the *Workplace Employee Relations Survey: Cross-section, 1998*, where the data has been downloaded 991 times since 1999 but the documentation for the study downloaded 10,500 times in the first half of 2007 alone.

Our Survey and Sources of Information for Costs:

- A survey of cost information for digital preservation has been completed and 13 survey responses collated and made available (see section 6). 11 responses were received from the UK and two from mainland Europe. Unfortunately a further two offered from the USA could not be available within the deadline for publication of KRDS2;

- Cost information from respondents is available for most of the KRDS2 main activity phases (pre-archive, archive, access, support services, and estates) although the depth and breadth of information available from different collections varies considerably (see individual responses);
- Information for some activities is very high (archival storage cost information is available in 100% of the responses). Other more infrequent activities such as disposal (and perhaps also preservation planning) are less well represented. Knowledge of acquisition costs is also relatively low (46%);
- Most cost information is relatively recent at least in terms of information which would be amenable to comparative analysis;
- Most of the data is potentially available for research subject to confidentiality or other terms and conditions.

Application of the KRDS Activity Model:

- The KRDS activity model has been reviewed by partner institutions and found to be broadly robust and fit for purpose: some small changes have been made to the sub-activities as part of KRDS2 (see section 4) and guidance on its application extended;
- We have re-emphasised our guidance in KRDS1 (and strengthened it in KRDS2) on tailoring the model and particularly the language/terminology for local use (see section 5.2);
- We have recognised that the activity cost models should be applied at different levels of detail for different purposes: as a result KRDS2 now caters for potential dual application of the activity model with two versions presented at different levels of detail (see sections 5.2, 5.3, and 5.4);
- Presentation of the activity model has been changed to a more user friendly format (see sections 5.3 and 5.4);
- We have recognised that the activity model is designed for costing preservation activities where there is a distinct archiving phase based on a designated archive centre or function. Although these exist within our case study sites and many institutions, we have also encountered specific research disciplines and sub-

disciplines where this is not the norm (see further discussion of implications and recommendations below).

This work has confirmed the strengths of the approaches underlying the original Keeping Research Data Safe report produced in 2008 but also allowed some limitations and areas needing further development to be defined.

The UKDA 2002-3 and 2005-6 Cost Datasets illustrated the inherent difficulties of retrospectively constructing a time series for digital preservation costs from historic data when survival of data is partial or it had been compiled for different purposes. Other more recent datasets (e.g. those from ADS) proved more amenable to analysis in KRDS2 format although they still required mapping and re-formatting (often a week of effort or more). These experiences suggest that opportunities for developing a longitudinal series of cost information to analyse digital preservation costs may be best developed prospectively rather than retrospectively. Consistent data collection and terminology could then be applied.

Recommendation 1: 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 (and future prospective evidence accumulated to it). This is more amenable to mapping into a consistent framework for analysis and often more complete than more historic cost evidence. A range of potential sources of such cost evidence are identified in our survey.

The costs survey shows a relatively limited number of institutions have information on digital preservation costs and few have information for all the activities. To be viable for more extensive research, the information base may need to be extended by using good international data sources and partnerships. For example two institutions contributed to the data survey from continental Europe and others were offered from USA.

Recommendation 2: The KRDS project team should seek future opportunities to extend the costs survey; raise awareness of KRDS internationally; and develop research partnerships on digital preservation costs.

Our cost case studies suggested there are some potential opportunities for cost savings and further automation of archive tasks which could be investigated further. Our work suggests the greatest potential areas for future tool development could be in ingest and access activities.

Recommendation 3: From KRDS2 outcomes, it is likely that the largest potential cost efficiencies will come from future tool development supporting ingest and access activities. Funders may wish to focus on investigating the potential benefits that could arise from further automation of these activities.

It is clear that the existing KRDS activity model is still heavily influenced by the OAIS reference model in its current presentation and its application is therefore ideal for those disciplines and preservation services focussed on data archives and other institutional, national or subject repositories. As illustrated by the University of Oxford case study, its presentation and application is perhaps less ideal for focussing on near-term (pre-Archive phase) preservation and curation work from a researcher perspective, or disciplines and institutions where long-term preservation remains focussed on small research groups or indeed single researchers. A need has been identified for a modified version or versions of the KRDS2 activity model for these audiences.

Within the current JISC Research Data Management Programme (http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx), projects are assessing current researcher workflows in different disciplines and the I2S2 (http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmi/i2s2.aspx) project in particular is seeking to extend the pre-archive phase of the KRDS2 activity model in light of its work on this. Over the next 12 months the Data Management programme may provide the ideal testbed for further developing the pre-archive phase of the KRDS2 activity model and producing versions of the model from a researcher's perspective.

Recommendation 4: Examine further development of the pre-archive phase of the KRDS2 activity model and produce versions of the model from a researcher's perspective.

The data survey confirmed that the best available sources of cost information currently are national services. Further cost information for other more distributed curation and

preservation in universities may also begin to be assembled in the JISC Research Data Management Programme and its costs/benefits support project. Some promising work has also begun within the UK Research Data Service (UKRDS - http://www.ukrds.ac.uk/) pathfinder projects on establishing costs and implementing a costs spreadsheet. There is also parallel work on developing a costing tool in LIFE3 and a number of European projects. KRDS2 is not currently implemented in spreadsheet form. Although significant further research may still be needed on KRDS2 metrics and variables for full implementation in a spreadsheet, initial efforts in this area may still be helpful to many HEIs.

Recommendation 5: Seek to implement KRDS2 in cost spreadsheets and continue research on implementation variables and metrics that could enhance them.

Keeping Research Data Safe has been implemented as two study reports supported by a project website with supplementary materials. Experience in KRDS2 has emphasised the importance of presenting the outputs in better ways for end-users and we have taken the opportunity of re-presenting the Activity Model in new ways as part of the project. We believe that it would be possible to continue this process to 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. Currently KRDS1 and KRDS2 are presented as research study reports. These may need to be integrated and combined with the future KRDS applications and news on the project website. This would be a relatively low-cost activity which could substantially help end-users utilise results of the study.

Similarly the KRDS2 research study report will not be suitable for disseminating the results and findings to all end users. Elements from KRDS2 and its findings should be considered by JISC for inclusion as appropriate in its Research 3.0 campaign (http://www.jisc.ac.uk/res3).

Recommendation 6: 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.

Recommendation 7: Elements from KRDS2 and its findings should be considered by JISC for inclusion in its Research 3.0 campaign to disseminate the results and findings to other end users.

Finally we believe the benefits taxonomy presented in KRDS2 has great potential for further development and implementation. Much of the literature addressing economic issues related to digital preservation focuses on the cost side of the cost/benefit equation. Comparatively little attention is paid to articulating the benefits to stakeholders arising from the preservation activity. We would encourage JISC and other funders to consider further work on identifying and quantifying the benefits of research data preservation.

Recommendation 8: JISC and other funders to consider further work on identifying and quantifying the benefits of research data preservation.

10. REFERENCES

Beagrie, N., Chruszcz, J. and Lavoie, B., 2008, *Keeping Research Data Safe: a cost model and guidance for UK Universities*, (Joint Information Systems Committee 2008).
http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx

Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008, *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation*, Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access December 2008.

http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010, *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access February 2010.

http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

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

Currall, J., and McKinney, P., 2007, *espida Handbook. Expressing project costs and benefits in a systematic way for investment in information and IT*.

http://hdl.handle.net/1905/691

Davies, R. (ed), 2008, *The LIFE2 Final Project Report*

http://eprints.ucl.ac.uk/11758/1/11758.pdf

East of England Regional Archive Council (EERAC), 2006, *Report of the East of England Digital Preservation Regional Pilot Project,* (MLA East of England and East of England Regional Archive Council June 2006).

http://www.data-archive.ac.uk/news/publications/darp2006.pdf

Economic And Social Data Service (ESDS), 2010, *Director's Evaluation Report* (March 2010). Internal document.

Fry, J., Houghton, J., Lockyer, S., Oppenheim, C., and Rasmussen, B., 2008, *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes* (JISC 2008)

http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/jiscdataproposal-public.pdf

Gerlach, J., Neumann, B., Moldauer, E., Argo, M., and Frisby, D., 2002, Determining the cost of IT services. *Commun. ACM* 45, 9 (Sep. 2002), 61-67. http://doi.acm.org/10.1145/567498.567500

Hunolt, G., Booth, B., Banks, M., 2008a, *Nasa Cost Estimation Toolkit*, Version 2.4 September 2008.

Hunolt, G., Booth, B., Banks, M., 2008b, *CET Version 2.4 Status and Results of Independent Testing Error Estimates and Progress Assessment*, 8 September 2008.

Hunolt, G., Booth, B., Banks, M., 2008c, *Technical Description Document Cost Estimation Toolkit (CET) Version 2.4, September 2008*.

Hunolt, G., Booth, B., Banks, M., 2008d, *Users' Guide Cost Estimation Toolkit (CET) Version 2.4, September 2008*.

Kejser, U., Nielsen, A., & Thirifays, A., 2009, *Cost Model for Digital Curation: Cost of Digital Migration*. Paper presented at The Sixth International Conference on Preservation of Digital Objects.

Macdonald, S. and Martinez-Uribe, L., 2009, "User Engagement in Research Data Curation", *Lecture Notes in Computer Science - Research in Advanced Technology for Digital Libraries*, Volume 5714, 2009. http://www.springerlink.com/content/7mnq13x34717p483/

National Academy of Sciences, Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*, (National Academies Press 2009).

Nationaal Archief, 2005a, *Costs of Digital Preservation version 1.0 May 2005* (Digital Preservation Testbed, The Hague, Netherlands).
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf

Serco Consulting, 2008a, *UKRDS Interim Report*, Version v0.1a.030708 7th July 2008.
http://www.ukrds.ac.uk/resources/download/id/17

Serco Consulting, 2008b, *The UK research data service feasibility study: Report and Recommendations to HEFCE*, 19 December 2008.
http://ukrds.ac.uk/resources/download/id/16

UK Parliament, 2001, *Hansard*, 3 May 2001, Column WA103. Written Answers.
http://www.publications.parliament.uk/pa/ld200001/ldhansrd/vo010503/text/10503w01.htm

Walker,A., et al, 2002, *Living in Britain. Results from the General Household Survey* (London: TSO, 2002).