KEEPING RESEARCH DATA SAFE

A COST MODEL AND GUIDANCE FOR UK UNIVERSITIES

Neil Beagrie, Julia Chruszcz, and Brian Lavoie

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

Final Report - April 2008

Prepared by:

Charles Beagrie Limited

www.beagrie.com

A study funded by

JISC

Contents

Acknowledgements 1

About the Authors 2

Executive Summary 3

1. Introduction 10

2. Methodology 13

3. Benefits of Research Data Preservation 16

4. Describing the Cost Framework and its Use 20

5. Key Cost Variables and Units 24

6. The Activity Model 36

7. Resources Template 47

8. Overviews of the Case Studies 49

9. Issues Universities Need to Consider 58

10. Different Service Models and Structures 67

11. Conclusions and Recommendations 76

References and Sources Consulted in Desk Research 81

Appendices 87

Appendix 1- Case Study : Archaeology Data Service Charging Policy 87

Appendix 2- Case Study : The University of Cambridge 95

Appendix 3 - Case Study: King's College London 108

Appendix 4- Case Study : The University of Southampton 118

Appendix 7: Extracts from Appendix D in Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century 164

Appendix 8: Interview Questionnaire 166

ACKNOWLEDGEMENTS

We would like to thank Neil Grindley, Joe Hutcheon, Neil Jacobs, and Alison Turner from JISC for their input to the study and suggestions for this report.

The study would not have been possible without input from the case study sites and their staff; particular thanks are due to: Patricia Killiard, Peter Morgan, and Elin Stangeland at the University of Cambridge; Sheila Anderson, Stephen Grace, and Gareth Knight at King's College London; and Simon Coles, Jeremy Frey, and Jessie Hey at Southampton University. We are also grateful to staff at the Archaeology Data Service for permission to include the ADS Charging Policy as an additional case study and commenting on the draft text for this.

Our other interviewees also made major contributions and gave their time and expertise to our work. Our thanks to: Kevin Ashley (University of London Computer Centre), Paul Ayris (University College London), Richard Davies (LIFE Project British Library), David Robey (Arts and Humanities Research Council and University of Reading), Kevin Schürer and Matthew Woollard (UK Data Archive and University of Essex), Allan Sudlow (Medical Research Council), Mark Thorley (Natural Environment Research Council), Heather Williams (Higher Education Funding Council for England), and Astrid Wissenburg (Economic and Social Research Council).

We are also grateful to Paul Wheatley and Richard Davies (LIFE project) and Kathy Fontaine (NASA) for commenting on drafts of Appendix 5 and 6 respectively; and to our peer reviewers appointed by JISC, and Kevin Ashley appointed by Charles Beagrie, for their comments on an earlier draft of this report.

Finally we gratefully acknowledge input from other Charles Beagrie staff including Daphne Charles and Robert Beagrie, for research, operational, and technical support over the course of the study.

About the Authors

Neil Beagrie is a leading expert on digital preservation and curation. He is a founding director of Charles Beagrie Ltd and has undertaken consultancy through the company for clients such as the Library of Congress, The National Archives, and the European Commission. His previous career spans a range of senior information and data management roles including Programme Director at the Joint Information Systems Committee, Director and Assistant Director of the Arts and Humanities Data Service, and Head of Archaeological Archives and Library at the Royal Commission on the Historical Monuments of England.

Julia Chruszcz has over 15 years senior managerial experience in UK HE academic computing services. By 1990 Julia had moved into computing services management at the University of Manchester, becoming the founder Director of MIMAS and had also worked with the Research Councils, principally EPSRC and NERC in establishing a national High Performance Computer service, CSAR. In October 2004, with the establishment of the new University of Manchester, Julia was appointed Deputy Director, Manchester Computing. As a member of the Manchester Computing Directorate, her primary areas of responsibility were the day to day management of Manchester Computing and the MIMAS, the JISC and ESRC supported data services to the UK academic community and beyond. Julia left Manchester University in October 2007 to focus on her consultancy work.

Brian Lavoie has a first degree and doctorate in economics and joined OCLC in 1996. He is a consulting research scientist in OCLC Research. His current research interests include analysis of aggregate collections, economic issues associated with information and the provision of information services, service models and frameworks for libraries, and digital preservation. He has written and presented extensively on many topics in digital preservation, such as the OAIS reference model, preservation metadata, costs, and economic sustainability. He is co-chair of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. This is charged with developing actionable recommendations on economic sustainability of digital information for the science and engineering, cultural heritage, academic, public, and private sectors.

EXECUTIVE SUMMARY

This study has investigated the medium to long term costs to Higher Education Institutions (HEIs) of the preservation of research data and developed guidance to HEFCE and institutions on these issues.

It has provided an essential methodological foundation on research data costs for the forthcoming HEFCE-sponsored feasibility study for a UK Research Data Service. It will also assist HEIs and funding bodies wishing to establish strategies and TRAC costings for long-term data management and archiving.

The rising tide of digital research data raises issues relating to access, curation and preservation for HEIs and within the UK a growing number of research funders are now implementing policies requiring researchers to submit data management, preservation or data sharing plans with their funding applications. This study provides:

- Brief overviews of the potential benefits to HEIs of preservation of research data; issues that HEIs will need to consider when determining the medium to long-term costs of data preservation; and different service models;
- A framework and guidance for determining costs consisting of:
    - A list of key cost variables and potential units of record;
    - An activity model divided into pre-archive, archive, and support services;
    - A resources template including major cost categories in TRAC; and divided into the major phases from our activity model and by duration of activity;
- A series of case studies from Cambridge University, King's College London, Southampton University, and the Archaeology Data Service at York University, illustrating different aspects of costs for research data within HEIs;
- Recommendations for future work and use/adaptation of software costing tools to assist implementation.

Overall our approach has focused on developing a framework for determining costs and the major deliverable from the study has been the costing framework.

In addition our case studies and specific work on costs provide valuable examples of research data costs. Given the emerging nature of the field, the limited time for the study, and sample size of case studies and interviews these must be regarded as illustrative examples of costs. However there are a number of emerging findings from them which are potentially very significant and which we have recommended should be explored and tested further in future work:

- **Institutional Data Repositories**. Our case studies suggest that the service requirements for data collections and the best structure for organising relevant services locally will be more complex than many have thought previously. Both Cambridge and KCL are developing central repositories to work with departmental facilities and discussing **federated local data repositories** for research data preservation combining services and skills from central and departmental repositories. Costs for the central data repository component at Cambridge and KCL are an order of magnitude greater than that suggested for a typical institutional repository focused on e-publications alone. These costs are discussed in greater detail in Chapter 10 and briefly summarised below:

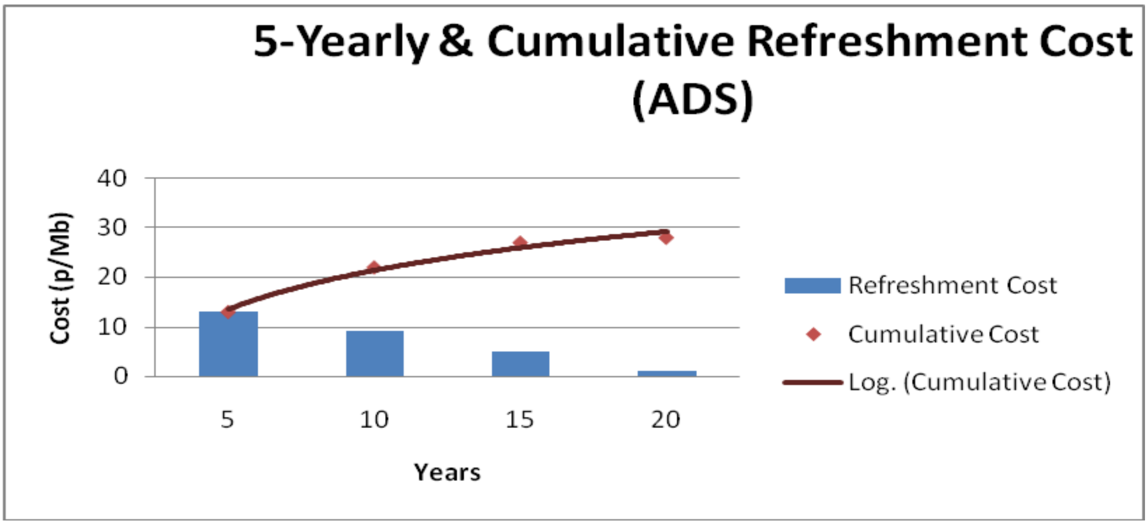
Institutional Repository (e-publications):	Staff	Equipment (capital depreciated over 3 years)
Annual recurrent costs	1 FTE	£1,300 pa

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

- **Long-Term Digital Preservation Costs**. The profile of costs across functions within the national data centres we interviewed appears to be very consistent. It was notable that they all believed their accessioning and ingest costs were higher than ongoing long-term preservation and archiving costs. For example the following approximate division of costs across high-level archive functions of our activity model were suggested for the UK Data Archive:

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

The implications of this for the cumulative long-term costs of archiving research data are particularly interesting and perhaps point to potentially effective management strategies (addressing issues early during acquisition and ingest) for managing longer-term costs. In a similar vein, the Archaeology Data Service (ADS) has been in operation for 10 years and provided an interesting projection of its long-term preservation costs for research data based on its costs to date and ongoing trends. This shows relatively high costs in the early years after accessioning but costs declining to a minimal level over 20 years as follows:



The ADS projection is a complex mix of underlying trends such as long-term declining data storage costs, costs for ongoing actions such as preservation interventions (file format migrations),and assumptions of archive growth which provide economies of scale. However, the implications of these factors and projection for sustainability of data archives e.g. via archive charges to project budgets, are notable and worthy of more extensive study and testing.

- **Archive economics**. We have observed and documented a number of significant issues for archives and preservation costs including:
    - **Timing**. Our activity model allows for consideration of relative costs arising from when activities are undertaken. We provide examples such as that from Digitale Bewaring Project which estimated costs c. 333 euros for the creation of a batch of 1000 records in the pre-archive phase. In contrast once 10 years have passed and material has been transferred to an archive it may cost 10,000 euros to 'repair' a batch of 1000 records with badly created metadata.
    - **Efficiency curve effects**. Our case studies illustrate a number of efficiency curve effects. The start-up phases of repositories reflect both the ramping-up of activities e.g. recruitment of staff and specific start-up activities such as developing new policies and procedures for the archive. The start-up costs particularly in terms of staff time can be substantial. The operational phases reflect increasing productivity and efficiency as procedures become established, tested and refined and the volume of users and deposits increases.
    - **Economy of scale effects**. We identify the importance of economies of scale and the impact this has on unit costs for digital preservation. As an example, the University of London Computer Centre (ULCC) which runs the National Digital Archive of Datasets, provided us with costs for accession rates of 10 or 60 data collections: a 600% increase in accessions only increases costs by 325% as a result of economy of scale effects.
- **"First- Mover Innovation" Costs**. Within our activity model we have identified digital preservation costs attributing to the traditional areas of archive storage, data management and preservation planning. However in addition we have identified activities and costs relating to the category of "First- Mover Innovation" Costs. Where

preservation functions and file formats are evolving a high-degree of R&D expenditure might be required in implementation phases and in developing the first tools, standards and best practices. Many of the disciplines and archives covered in this study have made considerable investments as communities in evolving shared standards, practices, and tools and we believe this could be making a significant impact on their long-term digital preservation costs.

- **The Cost Framework**. Our case study sites found the cost framework approach of value to their institutions and it will benefit from wider adoption, testing and evolution in other HEIs. Its particular strengths are:
  - It is based on Full Economic Costs (FEC) which are not in or partial in other models. We believe absence of FEC (a) can distort business cases and under-estimate cost benefits eg for automation, and also mean (b) HEIs cannot accurately compare in-house or out-source costs;
  - It can cost for in-house archive, full or partial shared service(s), or archive charges to projects and is implementation and technology-neutral. It is applicable in most digital preservation contexts, regardless of choices involving system architecture, preservation strategy, or service delivery;
  - It is tailored for research data by allowing for different data collection levels and preservation aims, and data-specific activities such as generating products from data.

Summary of Recommendations

This has been an intensive study over a period of 4 months focusing on the issue of the preservation costs of research data for UK HEIs. Our recommendations for future work to develop and implement outcomes from the study are discussed in detail in Chapter 11 and summarised below:

Recommendation 1: The outcomes of this study should be considered and utilised by the forthcoming JISC Data Audit Framework study.

Recommendation 2: Departments and Central Services within HEIs should utilise recurrent data audits to inform both their initial appraisal and development of data policies and future capacity planning for services.

Recommendation 3: HEIs should consider utilising the US National Science Board (the governing body for the National Science Foundation) long-lived data collection levels to aid understanding and categorisation of user requirements and costs over time.

Recommendation 4: HEIs should consider federated structures for local data storage within their institution comprising data stores at the departmental level and additional storage and services at the institutional level. These should be mixed with external shared services or national provision as required. HEIs should work with and utilise national and international disciplinary data archives where these exist. The hierarchy of data stores should reflect the detailed nature of the content, services required, and the changing nature of its importance over time.

Recommendation 5: We recommend consideration of the study and further work on development and implementation of relevant cost models and tools to HEIs, research funders, and service providers.

Recommendation 6: JISC should produce a short briefing paper or summary of this report and its findings aimed at senior managers including university academics, administrators and research support services.

Recommendation 7: JISC should consider developing project costing tools to build on and implement work within this study. These tools may be valuable for some of JISC's own projects and may also be of interest to other research funders and have potential for joint funding and development.

Recommendation 8: JISC should consider undertaking additional work to examine how the cost components and variables defined in our framework can be further quantified, and what additional data and data collection mechanisms are needed to support them.

Recommendation 9: JISC should consider further detailed study of longitudinal data for digital preservation costs and cost variables to extend the work of this study. Possibly this could be part of a UK based taskforce to feed into its joint international work on digital preservation costs.

Recommendation 10: JISC and/or other funders should consider funding further work on quantifying the benefits of research data preservation.

1. Introduction

"Because of the critical importance of data and information in the global scientific enterprise, the international research community must address a series of new challenges if it is to take full advantage of the data and information resources available for research today. Equally, if not more important than its own data and information needs, today's research community must also assume responsibility for building a robust data and information infrastructure for the future." (International Council for Science 2004).

There is a growing international understanding of the value of preserving research data and the necessity of developing an infrastructure to support this.

The UK Government's Ten Year Science and Innovation Investment Framework sees ready and efficient access to digital information of all kinds such as experimental data sets, journals, theses, conference proceedings and patents as the life blood of research and innovation (HMSO 2004). Within the UK a growing number of research funders are now implementing policies requiring researchers to submit data management, preservation and data sharing plans with their funding applications.

Some research councils fund national research data centres for their disciplines as part of this infrastructure. However not all research data within these disciplines will be housed there and there are also many other disciplines and datasets which need to be maintained within institutions.

As stated in the HEFCE strategic plan 2006-11(Updated April 2007):

> "The starting point is our role within the dual support system for the public funding of research, in supporting the core research infrastructure, underpinning work funded by other research funding bodies and enabling the sector to undertake curiosity-driven research. It also reflects the shared aim of the countries of the European Union to develop a knowledge-driven economy powered by a strong and innovative research

> base, and our belief that the UK is well-placed to play a leading role in achieving this." (HEFCE 2007,p30)

The strategy also states that HEFCE will:

> "...continue to encourage the effective sharing of research findings and outcomes, both to support research and teaching within HE and to inform the wider public. To achieve this we will work with partners to improve systems for researchers to share information and disseminate outputs as widely as possible, including through new technology." (HEFCE 2007,p34)

This in turn brings new challenges, and consideration now has to turn to the impact of born digital research data on UK Higher Education Institutions (HEIs). The rising tide of digital research data (Hey and Trefethen 2003) raises issues relating to access, curation and preservation in respect of how an institution might improve its international profile and support its primary purposes of research, learning and teaching by pursuing an effective digital preservation strategy for research data. In parallel with this there are often legal requirements and grant terms and conditions relating to the storage of and access to research data (Research Information Network 2007, Research Information Network 2008).

HEIs may host a wide range of different repository types with variable life-spans for preservation of data ranging from national or disciplinary data centres, research project databases, to institutional repositories. Data may also be deposited externally with other research support organisations such as the European Bio-informatics Institute or national research centres. In short there are different preservation, community and service options and requirements for data collections.

An approach which helps to predict demand, promotes institutional planning, and places costs in context of the benefits to institutions will be essential. HEIs need to be in a state of readiness to respond to Research Councils changing policy frameworks (e.g. on data sharing and data deposit; or events such as AHRC withdrawal of funding from AHDS and new expectations of institutions) and new requirements of researchers over time. HEIs need to understand current and emerging positions, and to gather information that will enable a gearing up for change in a timely way. This will be necessary in the context that investment in technical infrastructure, a skilled workforce, and the implementation of new systems can only take place with the support of a strong business case and a commitment by institutions to prioritise and commit funds.

This study is designed to provide some guidance to HEFCE and institutions on these issues and to inform their planning and future development by providing a flexible framework for costing the preservation of research data.

TERMS OF REFERENCE FOR THE STUDY

The study's principal aim was to investigate the medium to long term costs to Higher Education Institutions of the preservation of research data.

The Terms of Reference stated the study should:

1) investigate the costs (direct and indirect) of preserving research data, from an institution's point of view;
2) construct a list of issues which universities will need to consider when determining the medium to long-term costs of data preservation;
3) establish a methodology which will help institutions identify the cost elements, key variables and units of record needed for estimating the cost of preserving research data;
4) compare the costs of different models of preservation (eg. shared services, institutional repository, discipline- focused, centralised).

It was anticipated that the study would consider the direct and indirect costs of data preservation in the next 5-10 years and beyond (eg. re-skilling of staff; curation costs; disc space; transition cost; recurrent updating costs).

It was stated that the study should be undertaken within the context of other relevant reports and studies as well as other research which outline the benefits of the long-term preservation of research data. This particular study was required to focus clearly on the institutional costs of research data preservation.

2. METHODOLOGY

This study which began in December 2007 and completed at the end of March 2008 has developed a model of digital preservation costs for research data in HEIs.

Given the constrained timescale the project aims to achieve this by leveraging pre-existing investment by JISC and others in models and extend this through further desk research, interviews and analysis, input from the project team (Neil Beagrie, Julia Chruszcz, Brian Lavoie), and validation via involvement of three case study sites (Cambridge, Southampton, and KCL) to produce a cost model and guidance of direct relevance to UK HEIs.

The project has examined in detail two activity based models for costing and two broader models for context in terms of long-term digital archiving and full economic costing in UK Universities. We have considered the following two models as activity based cost models: the LIFE digital preservation cost model developed primarily for library materials, and NASA's Cost Estimation Tool (CET) developed for space and earth observation research data centres. We have mapped the Open Archival Information System (OAIS) reference model (an ISO standard for digital archives) against both of these and evaluated transferable practice and relative strengths and weaknesses for each. Our review of the LIFE cost models with OAIS is provided as Appendix 5 and our review of NASA CET with OAIS as Appendix 6 to this report.

Finally the study recognised the importance of aligning its model with existing costing systems in HEIs and has therefore also looked closely at the Transparent Approach to Costing (TRAC) model for assessing Full Economic Costs (FEC) in UK universities. TRAC employs the principles of activity based costing and the processes underlying TRAC allow all of the costs of the institution, direct and indirect, to be analysed and attached to activities in a fair and reasonable way. The TRAC methodology has been accepted and endorsed by the HE sector, by government, and by the principal funders of research and teaching. Assessment of TRAC and its application to the study has been based on our reading of the TRAC Manual (Joint Costing & Pricing Group 2005, 2008) and is being validated by finance staff with direct experience of its application in the case study sites and HEFCE.

In addition, desktop research has looked at the existing literature on preservation costs and a range of annual reports and documents from data services and research funders in the UK. Analysis of these has also fed into the model and final report. Particular documents of note were the Archaeology Data Service (ADS) Charging Model which also discusses costs (ADS 2007) and is the focus of an additional case study in this report (Appendix 1); and the classification of research data collections proposed by the US National Science Board in its study of long-lived data collections (NSB 2005). This identifies and characterises three research data collection types with different preservation and funding (and cost) requirements: research data collections; resource or community data collections; and reference data collections. These are being used extensively in the study. Summary extracts of these are provided in Appendix 7.

To supplement printed sources and validate information gathered through the desk research 12 interviews have been conducted using a pre-defined questionnaire (appendix 8) to collate information to feed into different sections of the final report. 10 interviews were conducted in person and two by telephone. Our interviewees were Kevin Ashley (University of London Computer Centre), Paul Ayris (University College London), Richard Davies (LIFE Project British Library), David Robey (Arts and Humanities and Humanities Research Council and University of Reading), Kevin Schurer and Matthew Woollard (UK Data Archive and University of Essex), Allan Sudlow (Medical Research Council), Mark Thorley (Natural Environment Research Council), Heather Williams (Higher Education Funding Council for England), and Astrid Wissenburg (Economic and Social Research Council) and staff from the 3 case study sites (Patricia Killiard, Peter Morgan, and Elin Stangeland at the University of Cambridge; Sheila Anderson and Gareth Knight at King’s College London; and Simon Coles, Jeremy Frey, and Jessie Hey at Southampton University).

Finally we have been working with three leading UK research universities, Cambridge, King’s College London and Southampton, who have acted as partners and contributors of case studies to the report. Our case study sites are helping us validate the methodology and illustrate the variety of costs and community and service requirements for research data across institutions. These universities generate significant quantities of research data and host distributed repositories holding important research datasets. We have discussed inclusion of research datasets from a range of disciplines with our academic partners and capturing their experience of different service models and costs. To help those contributing case studies, comprehensive guidance notes and a common template have been issued. Individual case studies illustrate institutional issues and costs. Following discussion with our case study sites and interviewees we have also included the Archaeology Data Service Charging Policy and its discussion of costs as an additional case study.

From these sources we have produced:

- Brief overviews of the potential benefits to HEIs of preservation of research data; issues that HEIs will need to consider when determining the medium to long-term costs of data preservation; and different service models;
- A framework and guidance for determining costs consisting of:
    - A list of key cost variables and potential units of record;
    - An activity model divided into pre-archive, archive, and support services;
    - A resources template including major cost categories in TRAC (staff, equipment, etc); and divided into the major phases from our activity model (pre-archive, archive, support services) and by duration of activity (year 1, year 2, etc);
- A series of case studies illustrating different aspects of costs for research data within HEIs;
- Recommendations for future work and use/adaptation of software costing tools to assist implementation.

Overall our approach focuses on the framework for determining costs, and illustrative case studies. We believe that at this stage it will be a major contribution to develop a general framework and examples that articulate the key cost categories of digital preservation for research data, as well as their relationship to each other, in particular, the trade-offs and opportunity costs involved.

3. Benefits of Research Data Preservation

The costs of securing the long-term persistence of research data must be balanced against the anticipated benefits. Broadly speaking, the benefits for higher education in preserving research data extend from the fact that the discoveries of the future rely on the work of the past. Maintenance of a complete and accurate scholarly record, including the portion in digital form, is essential for continued progress in research and learning. Yet this often requires a significant commitment of funds, equipment, and expertise; consequently, an appeal to the Newtonian vision of “standing on the shoulders of giants” may fall short of what is needed to make a persuasive case for adding these costs to already-strained budgets.

In this chapter, four categories of benefits associated with the long-term preservation of digital research data are discussed which strengthen the case for institutions to invest in this area. These categories emerged from a synthesis of institutional interviews conducted for this study, as well as the discussion which took place at the February 2008 JISC Research Data Seminar.

Protecting investment in research

Universities and other funding bodies have invested, and will continue to invest, enormous sums in research activities. Research in turn creates assets, in the form of new knowledge. Knowledge can be manifested in many types of outputs, including research data sets. New knowledge is valuable, both for its own sake, and for the opportunities it provides to encourage further research and learning. If the full value of the national investment in research is to be realised, the outputs from these investments need to be protected. Research data represents a category of research output of growing importance and value.

Protection of the fruits of the national research investment is achieved through implementation and maintenance of a reliable infrastructure to support the long-term retention of research data and other research outputs. Such an infrastructure would help reduce the loss of digital research outputs through accident, neglect, or even deliberate act.

While the costs of maintaining digital preservation capacity are not insignificant, the costs of the alternative are often greater. Re-creating research data sets can be prohibitively expensive; in the extreme, it may be impossible to re-create lost data, as researchers, test subjects, testing conditions, and so on disperse or disappear over time. For observational data such as atmospheric conditions over time or event episodes such as volcanic eruptions it may simply be impossible to recreate it once it is lost. In these circumstances, maintaining a reliable, managed environment for protecting the considerable institutional investment involved in creating research data would represent a comparatively small cost when placed against the prospect of the higher and perhaps prohibitive costs of re-creation later on or the complete and irretrievable loss of data.

Preserving opportunities for future research

Research data often has a value that extends beyond the work with which it was originally associated. For the researcher who originally created the data, as well as for other researchers in the scientific community, the ongoing availability of research data affords the opportunity both to validate existing results, and to build upon them. Access to research data can catalyze further work that creates value in a variety of ways, including reinforcing or corroborating earlier inferences; expanding on the foundations laid by earlier work; and even re-purposing the data in ways that could not have been foreseen. The ongoing persistence of research data serves the twin purposes of cultivating a deeper understanding of the historical development of a discipline and its ideas, and moving the frontiers of the discipline forward. Loss of the research data drastically reduces the opportunity for such work.

Preservation of research data can be viewed as a form of knowledge transfer, in the sense of passing the outputs of research across time and space. These transfers can also span boundaries between domains. For example, a scientific discovery holding the promise of commercial opportunity can be more easily transferred to corporate R&D laboratories if the primary data associated with the discovery are readily available. Facilitating the transfer of knowledge across organisational boundaries expands the opportunity for mutually beneficial partnerships between HEIs and other organisations.

Promoting the work of the institution and the researcher

Long-term preservation of research data confers benefits to both the researcher and the institution in terms of maximising research impact and cultivating reputation. Securing persistent access to an important research data set increases the likelihood that it will be utilized in future research, which in turn elevates the impact and visibility of the original research from which the data emerged. A single data set can result in a string of studies, each building on the findings of the preceding ones. Contribution of a widely-used research data set to the permanent scholarly record can have effects of equal and lasting importance similar to those of a seminal published study. These effects redound to the credit of the researcher or researchers responsible for creating the data set in the first place.

Similar benefits can accrue to the institution with which the researcher is affiliated. Maintaining a collection of widely-accessible data sets can help an institution cultivate a reputation as a centre of research in a particular discipline. This increases the appeal of the institution for prospective faculty and students, as well as to external research partners in academia, government, and private enterprise. This appeal will only increase over time as the research data continues to be used and re-used by researchers and students from the local institution and beyond.

SUPPORTING RESEARCH AND LEARNING WORKFLOWS

The benefits from preserving digital research data also extend to supporting research and learning workflows. Research data is an essential input to scholarly endeavour, whether that endeavour is focused on extending the frontiers of knowledge, or understanding the discoveries of the past. Research and learning are facilitated when the scholarly record is complete and easily accessible. Properly curated digital research data can be readily integrated into research and learning workflows now and in the future. The data is easily located when needed, and maintained in forms compatible with contemporary technology environments. This can lead to lower costs and higher productivity in research and learning, as less time is spent searching for needed data and converting it to usable forms. Digital preservation services are part of the general information infrastructure needed to support research and learning workflows at HEIs.

In addition to supporting the research and learning processes themselves, long-term preservation of digital research data can also help in the management of the outcomes of these processes. For example, preserving research data in managed environments can help in gathering and evaluating outputs for the Research Assessment Exercise, as well as other quality assurance activities. Long-term preservation of research data can also support validation or replication of controversial research results. Preservation of research data is an important facet of the development of university policies for the long-term disposition of research results in conjunction with the provision of institutional repository services.

Summary

It is impossible to speak of the costs associated with the long-term retention of research data without articulating a persuasive case about the value expected to materialise from incurring these costs. As the discussion in this chapter indicates, there are at least four categories of benefits that potentially flow from preserving research data. These benefits emerge at the level of the researcher, the institution, and the research community at large. They are both practical in nature – e.g., protecting existing research and managing reputation – and enlightened self-interest – e.g., preserving opportunities for future research and contributing to the permanent scholarly record. HEIs should consider the range of potential benefits across all of these dimensions when they weigh the pros and cons of investing in the infrastructure needed to support the long-term preservation of digital research data.

4. Describing the Cost Framework and its Use

Introduction

This section describes the cost framework suggested by the study for determining the costs of preserving research data and provides guidance on its use. The framework is derived from analysis of the LIFE, NASA CET, TRAC and OAIS models, the desk research, and input from the project team.

In developing the Framework we were guided by the following considerations which make it distinctive from other related work:

- It should support Full Economic Costing (FEC) for UK HEIs and integration with approaches in TRAC. FEC is either not in or partial in other UK digital preservation cost models or uses different methodologies derived from other sectors. However use of TRAC is a requirement for UK HEIs. Additionally using FEC in the Framework means (a) it does not distort business cases e.g. for automation by understating or omitting some costs (b) it allows more accurate comparison of in-house or out-sourced service costs and improves the quality of management information;
- It should be "application neutral" in terms of how preservation services are delivered and should not be restricted to costing an archive's internal activities. It should support costing for an in-house archive, full or partially outsourced shared preservation service(s), or inclusion of national/subject data centre archive charges in research proposals;
- It should be tailored for research data and include consideration of different data collection levels and their requirements, the need for relevant documentation and metadata, and distinctive activities such as generating products from data, etc;
- It should be flexible, providing a general framework of activities (Pre-archive, Archive, and Support Services), resources (staff, equipment, etc), and variables (economic adjustments such as inflation, depreciation, cost of capital; and service adjustments such as salary levels, volumes, formats etc) which can be developed and applied in cost models to suit local requirements and circumstances.

The Framework is described below. Illustrations of the issues and different aspects of costs it describes are provided in the case studies.

THE FRAMEWORK

The 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.

Economic adjustments cover inflation/deflation, depreciation, and cost of return for financing and investment which apply over a number of years to activities. Note the TRAC methodology includes two cost adjustments for infrastructure costs and the return for financing and investment. The infrastructure cost adjustment is applied to the institution's accounting treatment to depreciation of buildings to ensure they better reflect the full long-term costs of replacing that infrastructure. The return for financing and investment is intended to cover the cost of financing and to generate a minimum level of retained surplus to permit rationalisation, updating and development (Joint Costing & Pricing Group 2006).

Service adjustments cover other major variables affecting research data preservation costs over time for example the type of file format, volumes, or required metadata, documentation and IPR. Many of the service adjustments relate ultimately to the varying preservation aims and user requirements required for the different levels of research data collections, resource or community data collections, and reference data collections, described in chapter 9 and Appendix 7.

Service adjustments can apply to more than one activity and create cost dependencies where changes in one affect the other(s) e.g. if a required level of documentation and IPR clearance is not undertaken in the Pre-Archive phase there is a significant cost increase for these during the Archive phase.

An activity model 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. It is divided into separate templates for Pre-Archive, Archive, and Support Services in line with our activity model. Cost categories taken from TRAC are Staff, Equipment, Travel, Consumables, Estate Costs, and indirect costs. In addition we have added a cost category for archive outsourcing/archive charges for the specific needs of this study. It is a summary model as in practice the cost categories would be expanded to cover specific items e.g. individual members of staff and items of equipment, etc. In a full TRAC presentation staff costs would also be divided into direct or directly allocated costs, and economic adjustments (inflation/deflation, depreciation/infrastructure cost adjustment, cost of return for financing and investment) would be subsumed in calculations and applied as approved by the institution and funder to staff and other costs.

USING THE FRAMEWORK

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.

When using the Pre-Archive section of the resources template, costs should calculate solely the preservation related component of costs in this phase. Frequently cost categories for activities such as creating the submission package for the archive (e.g. staff time) will be direct costs in a research grant.

Where there is an expectation of data being retained and maintained beyond the life of the project, typically the archive function will be either internal to the institution or outsourced either to a national data service, subject repository, or other third-party service provider.

The Support Services template includes costs for Administration and Common Services in the activities model. In most cases these will form the major component of the institution’s shared services and indirect costs which will be applied either to the Pre-Archive or Archive costs based on an institutionally and funder approved formulae. In other cases though some of the activities under support services might be part of an archive’s or project’s direct costs for staff and equipment and can be accounted for separately within either the pre-archive or archive templates.

5. Key Cost Variables and Units

Introduction

The key cost variables and units have been derived from the desk research and analysis of practice in individual data archives including ULCC (Ashley 1999), NASA (Hunolt 2006a), and the Archaeology Data Service (ADS 2007); general studies or discussions of digital preservations costs and activities (Beagrie and Jones 2001, Chapman 2003, Erpanet 2003, Hendley 1998, Mcleod et al 2006, Nationaal Archief 2005a, Nationaal Archief 2005b, NSB 2005, Sanett 2003, Wheatley et al 2007, Woodyard-Robinson 2006), and our case studies and project team.

General Considerations

Collection Levels and Preservation Aims

Collection levels and preservation aims are discussed further in Chapter 9 and have a major overall influence on a number of key cost variables. The majority of data collections in HEIs are likely to be at the research collection level intended only for use of the project team and sometimes a very small number of external users. Retention periods and preservation requirements may be set by the funder’s grant terms and conditions or by legal requirements (e.g. for clinical trials). Note preservation costs may be highest in the early years and become less significant over time. Preservation requirements are likely to be at a basic “secure storage” level for a set number of years with sufficient description to allow retrieval over that period.

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

Controlling Future Costs

It is possible for institutions to control some of the complexity and unpredictability of future costs by limiting the future effect of some of the service adjustments listed below. For example by taking action to regulate variables such as file formats during acquisition and ingest. This can be seen in the practice of a number of research data archives in the case studies.

Timing

The timing of actions within the lifecycle has important implications for costs and is a significant dependency within the model. This is particularly true in relation to generating descriptive or preservation metadata and user documentation in the Pre-Archive phase rather than generating new/upgrading deficient metadata and documentation during ingest in the Archive phase. The Digitale Bewaring Project in the Netherlands which focused on government electronic records estimated it costs approximately 333 euros for the creation of a batch of 1000 records in an appropriate manner at creation i.e. in the Pre-Archive phase. Conversely once 10 years have passed since creation it may cost 10,000 euros to ‘repair’ a batch of 1000 records with badly created metadata (Nationaal Archief 2005a, 15). Similarly preservation action to address technology obsolescence may change from easily solvable and inexpensive while the technology is familiar and relevant staff and equipment are available, too expensive or even impossible once access to relevant staff and equipment are lost.

Cost Dependencies, Linkages and “Ripple Effects”

The above illustration of the effects of timing is one example of cost dependencies which exist and need to be captured within any model for preservation costs of research data. The NASA CET aims to capture ripple effects for costs from one function to another as variables change and allow “what if” scenarios to be constructed. Typical ripple effects are changes in volumes ingested on other archive functions, or changes in other archive activity on costs for support services such as software development and maintenance (Hunolt et al 2006).

Figures have been provided for the National Digital Archive of Datasets (NDAD) to the study showing the effect of changing workloads on staff and other costs based on either 10 accessions per year or 60 accessions per year. Total costs increase by 325% for a 600% increase in accessions (Kevin Ashley pers comm. 9/3/2008). The proportionate costs for different resources expended also change differentially by workload as follows:

	10 Accession pa	60 Accessions pa
Staff costs	64.1%[1]	75.5%[2]
Indirect costs (staff)	15.4%	18.2%
Equipment	13.9%[3]	4.3%[3]
Consumables	1.5%	0.4%
Other Estates[4]	4.3%	1.3%
Outsourcing[5]	0.9%	0.3%

Notes

1. Total staff costs are comprised of 59.1% direct staff costs; and 5% directly allocated staff costs.
2. Total staff costs are comprised of 69.7% direct staff costs; and 5.8% directly allocated staff costs.
3. 'Equipment' includes some charges levied by ULCC infrastructure group for storage and server admin which themselves include staff and indirect costs.
4. Paper Document Store
5. Offsite copy Archival Storage

Sensitivities to Workload and Process Time Scheduling

Staff resources are not easily or quickly adjusted to changes in overall volume of deposits, or short-term fluctuations in workload particularly if the archive has little control over when the deposits will arrive or has fixed requirements for the speed with which they must be processed. Sensitivity will be greatest for inherently labour intensive, un-automated functions (Hunolt et al 2006).

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

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

Economic Adjustments

Economic adjustments consist of inflation/deflation, depreciation, and cost of return for financing and investment. Inflation rates are typically agreed between the institution and funders and applied to cost categories such as staff. Deflation rates are typically applied to some equipment categories such as computer storage media with known long-term trends in price reduction. There are several methods for calculating depreciation generally based on either the passage of time or the level of activity (or use) of the asset, which attribute the historical or purchase cost of an asset, across its useful life. The cost of return for financing and investment covers the cost of financing and generating a minimum level of retained surplus to permit rationalisation, updating and development. The TRAC methodology includes two cost adjustments for infrastructure costs and the return for financing and investment. The infrastructure cost adjustment is applied to the institution’s approach to depreciation of major assets such as buildings to ensure they better reflect the full long-term costs of replacing that infrastructure. Procedures for applying inflation/deflation, depreciation, and cost of return for financing and investment and other adjustments will be available from Finance departments in institutions and the guidance in TRAC (Joint Costing & Pricing Group 2005, 2006, 2008).

Service Adjustments

Generic

Staff Costs and Labour Rates

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

Staff costs are likely to be the major cost in any preservation activity within an HEI. 70% or more of the costs of preservation services in the case studies relate to staff costs and historically these have always been seen as the major component of preservation costs (Ashley 1999).

Activity Duration

The duration of activities (year 1, year 2, etc) will need to be recorded so that costs and adjustments for inflation/deflation can be captured and modelled.

Start-up and Operational Phasing of Activity

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

Levels of Automation

Given the overall impact and significance of staff costs, levels of automation (or conversely the levels of manual intervention required per dataset) are a significant variable for overall costs. Levels of Automation are used as one of the recorded variables to adjust cost estimations in the NASA CET tool (Hunolt 2006a). The level of impact will be dependent on the economies of scale that can be achieved. In areas such as archive storage a high-level of automation e.g. robotic tape storage is widespread. In other areas such as ingest it will be most beneficial for high-volume accessions with relatively homogenous content.

Acquisition, Disposal and Ingest

Number of Depositors

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

Number, Mode and Frequency of Deposits

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

Number, Complexity and Type of File Formats

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

The complexity and type of file formats have similar issues. For example the ADS suggests images, text, simple 'flat' spreadsheets and tables cost less than CAD, GIS and relational databases. The difference in cost is a result of the migration and validation aspects of digital archiving. For example, to assess the success of the migration of an image file it only has to be looked at, whereas for a more complicated file type, such as a GIS, it is necessary to ensure that the full functionality of the file has been preserved during its migration. The latter process takes more time and hence costs more (ADS 2007).

Use of file formats which have been well documented, have undergone thorough testing and are non-proprietary and usable on different hardware and software platforms minimises the frequency of migration and reduces the risk and costs in their preservation. Similarly utilising formats which have been widely adopted minimises risk as it is more likely that migration paths will be provided by the manufacturers and a degree of "backward compatibility" will be available between versions of the file format as it evolves. Note there are often trade-offs here and one choice can conflict with the other i.e., a choice might need to be made between a non-proprietary format no-one uses and a proprietary yet widely adopted format.

It is advisable for institutions where possible to identify file formats which are preferred for archival storage and to seek deposits in that form wherever a choice of formats exist. Some institutions have also identified and distinguished between preferred, acceptable and unacceptable formats for transfer to the institution, for archival storage once in the institution's care, and formats which can be provided for users. Narrowing the range of file formats handled streamlines the management process and reduces preservation costs. It will also reduce the ongoing cost of software licences required by the institution (Jones and Beagrie 2001). This applies even if free open source software is being used, since most studies show that there is still an associated total cost of ownership and more applications will always cost more to support and maintain.

Although such non-proprietary formats can be selected for many resource types this is not universally the case. For many new areas and applications, e.g. Virtual Reality only proprietary formats are available. In such cases a crucial factor will be the export formats supported to allow data to be moved out of (or into) these proprietary environments. In research areas where data is intimately bound into custom software and migration options are limited, an additional issue and cost which need to be considered in preserving data is the issue of preserving the software required to keep the data accessible. This means that a decision needs to be made on the adequacy of the preservation (that is what significant properties are needed to be preserved - for the data and the software - and to what level of accuracy and tolerance) and the performance of the software required to maintain that performance, which sets the preservation strategy (essentially in practice whether an emulation, or migration approach to preserving software is undertaken). These choices will set the bounds where costs of preservation can be determined. Another observation which has come through strongly in the JISC study of Significant [Preservation] Properties of Software is that good software preservation at the Archive phase is made a lot easier if you have good software engineering at the Pre-Archive phase. So investment in good software engineering practice will pay off in the long term (Brian Matthews pers comm. 13/2/08).

Data Volumes

Data volumes need to be recorded. Typically these will be measured in Mb, Gb, Tb, or Pb volumes and the overall number of files. In general, higher volumes will lead to higher costs but the ratio of cost to volume is not a linear relationship as economies of scale and efficiency gains lower per unit costs (see pages 25-6 and 28-9). Some disciplines require petabyte stores. Institutions need to establish a policy that deals with both local demands of researchers together with a balancing of opportunity to effectively use shared national and subject repository services.

Metadata, Documentation, Ethics and IPR

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

Levels of Processing, Validation and Calibration

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

De-accessioning Costs

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

Archive Storage, Preservation Planning, Data Management

Retention Period

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

Management and Refreshment

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

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

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

Number of Versions and Copies

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

Storage Media (capacity, costs)

Storage media will be selected on the basis of service requirements e.g. data volumes, required speed of access, or archival properties, and cost. The selection of storage media will influence the frequency of future storage media migration and staff and equipment needed for this task. It is important to remember that the total cost of ownership of archive storage media and systems is substantially higher than the purchase cost alone. Research suggests that the initial capital costs are less than a third of the total costs of ownership (Linden et al 2005, p5).

Archive media monitoring

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

Access

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

Number of Users and User Communities

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

Standard or Custom Interfaces

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

Level of User Support

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

Access Control

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

Number and Volume of Accesses

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

Access/Distribution Method

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

Service Response Times

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

Processed Products

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

6. The Activity Model

Introduction

The activity model shows the full range of activities that are involved in and support preservation of research data. The Activity Model is derived from our analysis of the LIFE [LIFE], NASA CET [NASA CET], OAIS [OAIS RM] and TRAC [TRAC] models, modified and extended by the desk research and work with the case studies [Study Team]. Scope notes are provided to guide interpretation and use. Where models overlap the OAIS RM definitions have been used wherever possible. Principal sources are indicated in square brackets thus [OAIS RM].

ACTIVITY MODEL	
Attribute	**Scope Notes & *[source]***
Pre-Archive Phase	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. *[Study Team]*
Initiation	Included to note any significant implications for preservation costs downstream. *[Study Team]*
Project design	Take into account implications of any data creation or acquisition activity including data formats; metadata; volume and number of files, etc. *[Study Team]*
Data management plan	Should include plans for future preservation and data sharing. *[Study Team]*
Funding application	Include FEC elements including activity relevant to preparation for preservation where applicable. *[Study Team]*

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. *[NASA CET]*
Creation	Included to note any significant implications for preservation costs or archive access/use downstream. *[Study Team]*
negotiate IPR/licensing/ethics	These need to be dealt with at the earliest stages so that when data is accepted into an archive there are no residual issues around IPR, licensing, or ethics. These can be very difficult to resolve at a later stage. This is important because an archive, as custodian, will honour all applicable legal restrictions. An archive should understand the copyright concepts and applicable laws prior to accepting copyright materials into the archive. It can establish guidelines for ingestion of information and rules for dissemination and duplication of the information when necessary. *[OAIS RM]*
generate research data	Conceive and plan the creation of both raw and derived data created throughout the duration of the project, including capture method and storage options. *[Study Team]*
generate descriptive metadata	This function extracts Descriptive Information from the Archival Information Packages (AIPs) and collects Descriptive Information from other sources to provide to Coordinate Updates, and ultimately Data Management. This includes metadata to support searching and retrieving AIPs (e.g., who, what, when, where, why), and could also include special browse products (thumbnails, images) to be used by Finding Aids. *[OAIS RM]*
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. *[Study Team]*

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. *[Study team]*
Data management	Services and functions for populating, maintaining, and accessing a wide variety of data by the project. *[OAIS RM]*
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 OAIS for use in the construction of one or more Archival Information Packages. *[OAIS RM]*
Archive Phase	
Acquisition	In LIFE model but not in OAIS reference model, apart from negotiate submission agreement. *[Study Team]*
Selection	The development of the Selection Policy and its application. *[LIFE]*
negotiate submission agreement	The specification of submission requirements for producers/depositors together with communication and negotiation with producers/depositors. *[LIFE]*
outreach and depositor support	Support and training for researchers submitting funding proposals that include creating research data, and support and encouragement for researchers with data to deposit. *[Study Team].* N.B. Poorly captured in most other models – probably equivalent to technical co-ordination in NASA CET
Disposal	Poorly captured in most other models and added by the study team – destroy is also in draft DCC curation lifecycle model (Higgins 2007). *[Study team]*
transfer to another archive	Transfer material to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. *[Study Team].*

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. *[Study Team].*
Ingest	The ingest functional area includes receiving, reading, quality checking, cataloging, 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. *[NASA CET] & [OAIS]*
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.*[OAIS RM]*
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 *[OAIS RM]*. 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 (or 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, data representation conversions or reorganization of the content information. [*modified from OAIS RM]*

generate administrative metadata	Metadata about the preservation process: • pointers to earlier versions of the collection item • change history ***[OAIS RM]***
generate/upgrade descriptive metadata and user 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. ***[NASA CET]***
co-ordinate updates	Provides a mechanism for updating the contents of the archive. It receives *change requests*, *procedures* and *tools* from Manage System Configuration. ***[OAIS RM]***
reference linking	The linking of primary data to textual interpretations of that data. Pioneering projects such as JISC-funded eBank, have demonstrated that this is a very powerful and valuable feature. It is now being explored by a number of other JISC-funded repository projects such as SPECTRa[1], CLADDIER[2] and a joint follow-on project, StoreLink[3]. There is also some evidence that such virtual links may facilitate real connections between physical services i.e. between data centres and institutional repositories in libraries ***[Study Team].***
Archive Storage	Services and functions used for the storage and retrieval of Archival Information Packages (AIPs). ***[OAIS RM]***
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. ***[OAIS RM]***

[1] SPECTRa project: http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_spectra.aspx

[2] CLADDIER project: http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_claddier.aspx

[3] StoreLink project: http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/storelink.aspx

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*. ***[OAIS RM]***
replace media	This provides the capability to reproduce the Archival Information Packages (*AIPs)* over time. ***[OAIS RM]***
disaster recovery	Provides a mechanism for duplicating the digital contents of the archive collection and storing the duplicate in a physically separate facility. 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. ***[OAIS RM]***
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. ***[OAIS RM]***
provide copies to access	The archive design will reference the preservation strategy and policy, considering off-site copies and any disciple requirement for multiple versions or editions. The number of versions and copies affects storage and management costs. ***[Study Team]***
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. ***[Study Team modified from OAIS RM]***

monitor designated user community	The Monitor Designated 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. ***[OAIS RM]***
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 archives current holdings. ***[OAIS RM]***
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 Community service requirements or technology trends that would require migration of some current archive holdings or new submissions. ***[OAIS RM]***
develop packaging designs and migration plans	The Develop Packaging Designs and Migration Plans function develops new IP designs and detailed migration plans and prototypes. This activity also provides advice on the application of these IP designs and Migration plans to specific archive holdings and submissions. ***[OAIS RM]***
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 needs to be put in place and monitored. Not in other models. ***[Study Team]***
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 ***[LIFE].***Preservation has a feedback loop back into/through Ingest functions in activity model. ***[Study Team]***

generate preservation metadata	
First Mover Innovation	Where preservation functions and file formats are evolving a high-degree of R&D expenditure might be required in implementation phases and in 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. Not in other models – added as has significant implications for cost modelling or potential for use/re-use. ***[Study Team]***
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. ***[Study Team]***
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. ***[Study Team]***
engage with vendors	This might include beta-testing, participation in user groups, and development of commercial partnerships. ***[Study Team]***
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. ***[OAIS RM]***

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. ***[OAIS RM]***
perform queries	Receives a *query request* from Access and executes the query to generate a *result set* that is transmitted to the requester***. [OAIS RM]***
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***. [OAIS RM]***
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.* ***[OAIS RM]***
Access	Services and functions which make the archival information holdings and related services visible to Consumers***. [OAIS RM]***
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. ***[NASA CET]***
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 specialized processing. ***[OAIS RM]***

deliver response	The Deliver Response function handles both on-line and off-line deliveries of responses (Delivery Information Packages, result sets, reports and assistance) to Consumers. *[OAIS RM]*
user support	The user support functional area includes support provided in direct contact with users by user support staff, including 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 science expertise to assist users in selecting and using data and products. *[NASA CET].*
new product generation	Initial generation and reprocessing with quality checking of new data products produced from data or products previously ingested, or generated *[NASA CET].* Note that this has as a feedback loop back into/through Ingest functions.
Support Services	
Administration	Services and functions needed to control the operation of the other functional entities on a day-to-day basis. *[OAIS RM]*
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 science activities, information technology planning, and data stewardship. *[NASA CET]*
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. *[Study Team]*
Administrative support	Administrative support and control provided by office managers, personal assistants and secretaries. *[Study Team]*

Common Services	These are the other shared supporting services supplied by the institution or located within the archive. *[Study Team]*
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.*[OAIS RM]*
network services	These provide the capabilities and mechanisms to support distributed applications requiring data access and applications interoperability in heterogeneous, networked environments. *[OAIS RM]*
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 *[OAIS RM]*
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. *[Study Team]*
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. *[Study Team]*
utilities	Supply of uninterrupted power supply, air conditioning, water etc. *[Study Team]*
supplies inventory and logistics	Management of supply chain, movement of goods, and recording of purchases and deliveries. *[Study Team]*
Estates	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 *[Study Team].*

7. Resources Template

Introduction

The resources template is derived from our activity model divisions of Pre-Archive, Archive, and Support Services, and TRAC cost categories (Joint Costing & Pricing Group 2005) with specific additions for archive charges and outsourcing for the requirements of this study. It provides a framework to draw together other elements of activity model and cost variables. The template presents categories of cost (e.g. staff) and duration (year 1, year 2, etc) in a simplified, generic form closer to that used in the cost methodologies of UK HEIs based on TRAC. It is a summary model as in practice the cost categories would be expanded to cover specific items e.g. individual members of staff and items of equipment, etc. In a full TRAC presentation staff costs would also be divided into direct or directly allocated costs, and economic adjustments (inflation/deflation, depreciation/infrastructure cost adjustment, costs of return for financing and investment) would be subsumed in calculations and applied as approved by the institution and funder to staff and other costs. Typically the cost model will implement these as a spreadsheet, populated with data and adjustments agreed by the institution. For further information see *Using the Framework* page 22-3.

Pre - Archive	Duration (year 1, year 2, etc.)
Staff costs	
Equipment costs	
Travel	
Consumables	
Estate costs	
Indirect costs	
Outsourcing/Archive Charges	

Archive	Duration (years 1-5, 5-10,etc)
Staff costs	
Equipment costs	
Travel	
Consumables	
Estate costs	
Indirect costs	
Outsourcing	

Support Services	Duration (year 1, year 2, or 5-10, etc.)
Staff costs	
Equipment costs	
Travel	
Consumables	
Estate costs	
Outsourcing	

8. Overviews of the Case Studies

Introduction

Each Case study is 10-20 pages in length and they are included as appendices to this report. The case studies are intended to provide detailed descriptions of issues and costs to illustrate the study. They have also helped to develop and validate the approaches to costing the preservation of research data proposed in our costs framework. The following short commentaries are a guide to the coverage of each case study:

Appendix 1: Archaeology Data Service Charging Policy

The ADS Charging Policy has been developed over a 10 year period and provides a useful and still relatively rare historic perspective on preservation costs. It is influencing charging policies being developed by other research data repositories.

The ADS model is an excellent illustration of the levels of service required and potential costs that can be incurred for Resource or Community Data Collections in a national subject-based archive in a university. It is activity based and can be mapped into the activity model in this study (although the grouping and presentation of activities differ slightly). Economic Adjustments (inflation/deflation, depreciation and costs of capital) from our model are built into York University and ADS staff and refreshment figures. Service Adjustments from our model are either reflected directly in the ADS cost structure (e.g. salary levels, file format, volume) or controlled in terms of deposit (e.g. mode of deposit; metadata, documentation and IPR; de-accessioning costs; standard or custom interfaces).

Implicit in the charging policy is the business model for sustainability and the need for a sustained flow of future deposits and deposit charges (this effectively is similar to the business model for sustaining research within universities used in TRAC or state pensions through national insurance contributions). This ensures that there is a continued existence of the service into the period where refreshment takes place. The refreshment charge itself includes modelling of the size of the archive and its future growth through a sustained flow of future deposits.

Finally it should be noted that the ADS charging model is specific to the service at York and conditions within its discipline, so precise figures would not apply to another service or subject area. However its generic features suitably amended for different economic and service adjustments will be valuable elsewhere.

APPENDIX 2: UNIVERSITY OF CAMBRIDGE

The main contributions from the University of Cambridge have come from the DSpace@Cambridge institutional repository and the Department of Chemistry's Unilever Centre for Molecular Science Informatics. In addition information has been collected for the study from the following departments: Department of Social Anthropology, and The University Library and Scott Polar Research Institute.

DSpace@Cambridge is of particular interest to the study as the first institutional repository in the UK to have a major focus on research data from its institution in addition to e-publications. It has only been fully operational since August 2006 so it has less longitudinal information on costs compared to the other case studies.

DSpace@Cambridge is a service run by Cambridge University Library and the University of Cambridge Computing Service. It is intended to be a core repository within a federation of repositories in the University and envisages taking primary responsibility for preservation.

The repository originated as a collaborative project, funded by the Cambridge-MIT Institute from 2003 to 2006, between Cambridge University Library, the University Computing Service, and MIT Libraries. To ensure that DSpace@Cambridge could be maintained as a sustainable service when CMI funding ended, the project also developed a business model for recurrent funding: this became operational when the repository service was inaugurated in August 2006.

The case study describes the issues encountered by the service and maps its activities and funding to the Study’s activity model. It notes running preservation services centrally is cost effective and for many departments at the university the only available option. For example in the Department of Social Anthropology data management is a responsibility of each researcher, and they estimate that it would cost approximately £ 30,000 p.a. to hire supporting staff to manage data. The Scott Polar Research Institute made the same decision when deciding to use DSpace@Cambridge for the preservation and dissemination of digitised images created by its Freeze Frame project.

The Unilever Centre is also described. It is aiming to establish a local departmental facility to capture and disseminate crystallography data. This will happen in two steps; initial implementation is planned for the summer of 2008 where two Graduate students will work on developing an infrastructure for storage and dissemination, simultaneously they will be approaching the chemists to acquire permission to make the data openly available and then facilitate deposit of the data. Further they plan to employ a Graduate student in a 10% FTE position continuing the work of acquiring permissions and facilitating deposits to the repository, in addition they estimate that they will need 25% FTE of a Computer officer. Equipment costs and indirect costs will be covered thorough the budget of the Unilever Centre. For preservation they wish to use the DSpace@Cambridge service.

Appendix 3: King's College London

This Case Study was prepared by the Centre for e-Research. The Centre incorporates the former Arts and Humanities Data Service (AHDS) Executive and its staff and projects.

The Case Study is based on the experience of the AHDS ingesting and preserving complex research data collections over an 11 year period, and on the more recent experience of the Centre for e-Research as it works to establish a research data management and preservation infrastructure for King's College London. It is taking as its starting point the strategic decision by King's College London to support research practice by developing a virtual research environment, including a research data repository to support the creation, management and long term preservation of College research data assets.

Application of research data preservation cost model

The case study concentrates on an initial allocation of costs to the TRAC categories as a prelude to costing the management and preservation of research data as a Major Research Facility (MRF), or a Small Research Facility (SRF). The Case Study applies TRAC methodology to the lifecycle activity model and resources template in the study in order to demonstrate how institutions might allocate costs across the TRAC elements. The Case Study broke down the lifecycle model elements into the TRAC categories of directly incurred, directly allocated, and indirect, following the application of TRAC at King's College London.

The approach taken has been to regard as directly incurred all those costs that could be validated against a project – those costs for which a project would be able to provide an audit trail to indicate costs directly spent on the data aspects of a project. Directly allocated costs are based on the FEC costs of running the research data facility. In line with the TRAC Guidelines, all costs are included as directly allocated rather than indirect.

A data audit exercise is needed at the outset of scoping a digital archive. This will identify collections and their relative importance to the institution and wider community. The archive can plan for ingest of collections from the data audit. By repeating the audit each year, an archive will also get a forward view of collections in the process of being created. Working in collaboration with Research Support Offices will give a three or four year perspective, and so help to inform infrastructure replacement.

A cost model is often created based on a set of pre-defined criterion or presumptions for the service adjustments. In this case the criteria indicate the expected operation of the digital archive, considering factors such as the collection policy of the digital archive, the time and effort required to curate and preserve each data type and any activities necessary to tailor the research data for the Designated Community. However, there may be circumstances in which the digital archive is offered research data that is considered atypical, requiring the creation of a new costing model to finance additional work, and/or requires reconsideration of the organisation and technical infrastructure that is required to curate data.

Two internal spreadsheets: Archive Average Costs and Collection Costs, were developed for use in KCL. Archive Average Costs provided average costs for running the Research Archive Facility – all costs that would be allocated under the Directly Allocated category. It must be borne in mind that the allocations made were a first attempt and are not regarded as definitive.

Cost data

The case study considered three staff members to be essential for the establishment of a repository: an Archive Manager (salary £45,000) to co-ordinate activities; a half time System Administrator (FTE salary £24,000) to install and manage hardware and software; and a Collections Officer (salary £35,000) to develop and implement appropriate workflow and standards for the curation and preservation of research data. The values assigned included salary, indirect and estate costs for each staff member as at March 2008.

The costs allowed a twelve month period to build and establish a data archive. Thereafter, the staffing affords the ingest of about 30 collections each year, with an assumption that 10 will be simple collections (images or texts created to standard formats and metadata), and 20 complex collections, comprising more complex formats (such as 3D visualisation materials) and multi media. Once this limit is reached, a new Collections Officer (CO) is required to cope with the additional work. The second CO may not be fully used on repository work at first, so the archive may spend more time on research, advocacy, data audits, retrospective additions, metadata enhancement. The cost of any new CO post will be smoothed over a 3-5 year period.

The hardware costs were based on those purchased by the AHDS in 2005 and comprises 15TB of storage, a tape library, and a dissemination server to allow end user access. The infrastructure was designed to cope with a maximum storage size and bandwidth; if either of these reach capacity a new or significantly altered arrangement is required. Archives should plan to renew their infrastructure every 3-5 years, and so build replacement costs into the annual planning.

Projecting data preservation costs

The spreadsheets included: 10 year projection for research archive; equipment projections; staff cost projections; and contained a first attempt at projecting costs over a 10 year period. The projections were relatively simplistic and did not account for the potential of automation of some processes - the complexity of such an exercise was beyond the remit of the case study.

The projections were based upon the concept of ‘spikes’ in cost: as the archive expands, so there is a need for more equipment and storage capacity to manage the increasing volume of data, hence costs increase over time. In a similar fashion, as the archive expands, and the number of collections ingested each year increases, so additional staff will be needed. The staff costs were based upon a collections officer dealing with the ingest and preservation of c.30 collections per annum, with 10 ‘simple’ collections (that is, simple text or images created to the preferred archival standards) and 20 complex collections, for example a multi media collection with images, text, video and sound and interlinking documentation (created to archival preferred standards); and with 20% of their time spent on general support tasks e.g. reviewing and updating licence agreements, standards activities etc.

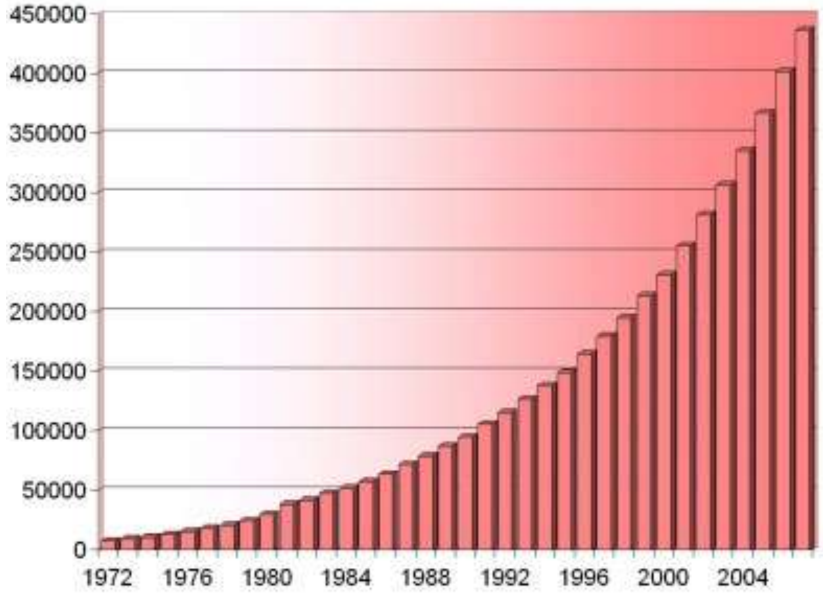
The figures demonstrated that at a steady state, costs rise slowly over time, but if the archive expands, then ‘spikes’ in cost must be planned for and incorporated into financial planning.

Appendix 4: University of Southampton

Information for the University of Southampton case study has been collected from the School of Chemistry and the National Oceanography Centre. The Chemistry case study is of particular interest in having a departmental perspective and for the very long time series of costs over a 19 year period assembled for the National Crystallography Service (NCS) based in the Chemistry department.

The NCS has operated a range of different instruments and has been operational for long enough to have seen, and be party to, a number of different techniques, processes, software programs, file formats and standards. Over the years this gives rise to very useful longitudinal data in the context of the acquisition of essential analytical chemistry data and so is an excellent case to inform this study. Initially policies on the archiving and storage of this digital data were scant due to a lack of knowledge or understanding of working with this medium. More recently it has become clear that a service that is operating on behalf of others must have a policy for the archiving of the data it generates so that the data can be provided on request, sometime after the original experiment(s) has been performed.

The rapid increase in crystallographic and computing hardware speed and capability over the last decade has resulted in a data deluge which is causing considerable problems for the management, archiving and publication of both raw and results data in the crystallographic field, as evidenced by the submission statistics to the CSD shown below:



For the costs study they have isolated one significant and representative activity undertaken within the School of Chemistry to highlight the production and curation of data. A major activity in Chemistry departments is the synthesis of a new compound, a fact that needs to be backed up and proven by structural studies. They have abstracted this process by considering both the typical synthetic task being undertaken in a synthetic organic chemistry group generating the samples (the project model), and the characterisation by the NCS (the sample model).

Based on the historic data available from the NCS a longitudinal study for the preservation of data in the sample model is presented. This is broken down into raw and results data and discussed alongside current innovations, which it is envisaged will provide a sound basis for preservation services in the next 10 years.

Raw data preservation costs per sample

1989-1996 Magnetic tapes	£21.95
1997-2003 Compact Discs	£6.00
2003-Present Outsourcing	£1.48

The cost of *archiving* per sample has roughly dropped by a quarter each time a new storage medium (and hence archival approach) has become widely available. It is important to note that this process is one of byte storage and very little, or no, preservation activity is performed – CD's were not periodically checked to ensure they were still readable and the outsourcing option merely ensures the retrieval of byte deposited. However, migration between media is often a problematic matter and is closely tied to the instrumentation – new instruments involve new software, formats and archival methods.

Results data preservation per sample

Per sample results data preservation costs is quite different to raw data in that its volume is considerably more manageable. Per sample costs were as follows:

1970-1990	Paper records	£30.00
1990-2000	Electronic copies on 3.25” floppy disks	£7.25
2000-present	Electronic copies on computer disks	£2.15

The real cost of archiving results data roughly drops by a quarter as new methods and media become available. The cost of migrations is extremely high, with paper to electronic being about £25 per structure and a large amount of data loss between spinning media and solid state. The cause of this high cost and large amount of loss is the large amount of time required to perform the process. Results can be regenerated if the raw data is preserved. However at modern day FEC, this would amount to between £50 and £400 (1-8 hours PDRA time) per structure. If raw data has not been preserved and results are lost then the cost of not preserving this data is enormous, as the compound generally cannot be re-synthesised and therefore the amount that might be attributed here would be the cost of generating a molecule from the macro study (£20,000).

Application of the Cost Framework

Southampton found the cost model is very thorough and maps reasonably well onto the the chemistry research data lifecycle. The activity model includes important elements that did not seem to be in other models and are vital to consider and quantify (particularly the initiation phase). On the whole the framework is very comprehensive and presented in an intuitive fashion with easy to comprehend terms which are well defined.

Looking forward

Current innovations that involve the NCS and School of Chemistry / University of Southampton are providing examples of best practice in the preservation of this data and therefore give indications of future costs.

Raw data: The archiving of large datasets is becoming much cheaper as mass storage solutions become commonplace. Research between eCrystals and ePrints at the University of Southampton is developing preservation services for such a system. The initial outlay is great, but will provide solutions for whole communities or disciplines. Hardware costs equate to approximately £1.60 per dataset for a fully redundant (RAID type system) that is automatically self-healing. However, this hardware solution only addresses the issue of bit-rot and full preservation services have yet to be developed (1 research assistant for a year, ca £80K at FEC rates), although once this is achieved the maintenance and migration work would be low (5% FTE).

Results data: The eCrystals project developed schema and repository software for the preservation of crystal structure data in its first two phases (£350,000), however it should be noted that this work was sufficiently generic not only to deploy for the whole crystallographic community, but potentially to act as a model for any experiment based science. The approximate year on year cost of running such a repository in the laboratory environment, with all the associated preservation administration and support would be £10 per crystal structure.

9. Issues Universities Need to Consider

Once the generic benefits in chapter 3 above and those specific to institutional goals and ambitions articulated within institutional strategies are well understood, it is important to define requirements. You can then apply a cost model to these requirements to estimate the level of investment needed for the preservation of research data. This will then contribute to building a business case, necessary to release funds or attract investment. When decisions and plans are being made to progress research data preservation initiatives within HEIs there are a number of areas that require careful thought. These will cover both the macro: the wider institutional considerations, and the micro: the detail of applying a cost model to the preservation of research data. One will inevitably have influence on the other.

Macro issues relating to costs emerging from the interviews and research that took place during the study can be categorised into five main areas: Strategic; Economic; Cultural; Operational; and Risk. The following highlights the issues in each area. This inevitably leads on to how certain key variables relate to the issues and how they are influenced. These are dealt with separately in previous chapters on the cost framework. In addition the case studies for Cambridge, KCL, and Southampton provide specific illustrations of both the macro and micro issues on costs for research data in individual institutions (see appendices 2-4).

Strategic

External Policy and National/International Context

External policy such as that of funders and developments in the international community are influencing the need for institutions to actively engage in the challenges of preserving research data. As well the policies of the UK funding councils, international agencies including the OECD and EU are now placing importance upon research data preservation.

In January 2004, the Organisation for Economic Co-operation and Development (OECD) published a Declaration on Access to Publicly Funded Research Data, to which the UK Government was a signatory. This proposed ten principles for open access to research data from public funding. The principles include openness, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality and security, efficiency, and accountability (OECD 2004). Sustainable preservation and archiving is seen as a key requirement for fulfilling the principles of formal responsibility, professionalism, quality and security, efficiency, and accountability (OECD 2007).

In November 2007 the European Union Council adopted "Council Conclusions on scientific information in the digital age: access, dissemination and preservation" during the Competitiveness Council meeting held in Brussels. Amongst other things the Member states are to be invited to ensure the long term preservation of scientific information - including publications and data, and pay due attention to scientific information in national information preservation strategies. In addition to the invitation to member states to do this, the Commission itself will encourage research into digital preservation, and deployment of cross-border data infrastructures; and seek to encourage policy co-ordination (European Commission 2007).

The Role and Responsibilities of HEIs (and that of Others)

With the "data deluge" there has come a rising tide of realisation and expectation that important research data resources will be preserved and remain accessible into the future. Most research funders have expectations that HEIs will take responsibility for data management during the lifetime of a project or programme. In some disciplines an offer to deposit with national or international repositories is then mandated for later long-term preservation and access. In others that expectation may rest on HEIs.

Not all data will have long-term value beyond the life of a project but a significant percentage does. Each institution should consider its strategic goals together with the potential value of its research data assets over time and plan accordingly.

Different Preservation Aims and Data Collection Levels

An understanding of different preservation aims and collection levels may assist in developing relevant data preservation and retention policies. It is very important to recognise that data collections vary substantially in terms of their anticipated user community and levels of use and therefore the associated preservation aims and costs. We have suggested earlier in the report that HEIs consider using the collection levels of research, resource or community, and reference data collections proposed for long-lived data collections by the National Science Board (NSB) in the USA (NSB 2005). In brief these are:

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

The full definitions of these collection levels provide indicators of likely number of users/user communities and levels of user support, periods of retention and preservation, and application of standards and quality control and validation of data and its accompanying metadata and documentation (see Appendix 7). These are significant cost factors so the collection levels and indicators for them may assist in identifying similar collections and cost estimation from "peer" collections with known cost data.

Note collection levels can change over time making possible for a collection and its preservation aims and intended user community to change. Such changes may be infrequent but incur significant preservation upgrade costs. The reasons for this are:

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

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

During the study we have noted a number of examples of this upgrading process for historic datasets. The Medical Research Council for example has proposed progressive development of bronze, silver, and gold service access levels to support wider research use of data collections at the Avon Longitudinal Study of Parents & Children (ALSPAC), the University of Bristol; and the National Survey of Health and Development (NSHD) held by the MRC Unit for Lifelong Health and Aging. These initiatives are essentially about enhancing a research collection so it can move towards becoming a reference collection supporting a much broader user community (Allan Sudlow pers comm. 24/1/08).

Preservation strategy and policy

Institutions should develop their own preservation strategy and policy with regard to research data, taking into consideration the influence of their institutional strategic plan, legal requirements, and the policy of, and support provided by, external agencies. Policies for research data needs to be embedded within institutional and departmental strategies. It is important to follow through and to develop the infrastructure that will deliver the strategy and policies.

An institution needs to ensure that its own research data retention or preservation policy is reviewed regularly to take into account external developments. Recent closure of AHDS by AHRC highlights some of the issues of sustainability and risk that may affect national services and the knock-on effects for institutions who have been expected to step into the void this has created.

Autonomy, Centralisation, or Federation

The institutional infrastructure for research data in HEIs is likely to be a mixture of central support services and departmental research support staff and facilities. Strategically, which balance of data management at subject discipline level or at institutional level will bring the most success? An understanding what researchers want and the delivery of this will determine whether the service(s) can be successful. Individual HEIs will need to consider the right mix of autonomy, centralisation, or federation to deliver the appropriate services and mixture of skills for management and preservation of research data in the institution.

In-house and/or External

A parallel consideration is whether this service(s) should be provided solely in-house or can be wholly or partially delivered by an external provider on behalf of the institution. The availability (now or in the future) of reliable external preservation service providers with adequate capacity, service models, and so on to meet the preservation requirements of HEIs will be a significant issue. This issue and costs of different service structures including shared services are considered further in chapter 10.

Economic

Funding Agencies and Funding

How to fund the development and maintenance of research data repositories within HEIs is an issue that can be understood better by applying a cost model. By identifying which elements of the data life cycle are important for archive and preservation and which of these lie within a research project's direct costs, and what can be considered through indirect and estate costs, provides a baseline for institutions to address funding and to assess the level of inward investment required. Clearly funding agencies expecting data plans in project proposals should expect to see evidence of them with the project's budget submission.

Institutional Investment

HEIs need to factor into their policy that the benefit of keeping research data sustainable would accrue over a potentially long time scale as it should be considered an archive and not a library. This should form part of a long term investment strategy. There is a need for selection and selection criteria – there will simply be too much data and cost will be too high without this. Given that project funding is usually short term in nature it is important to consider the mechanisms to facilitate preservation in the medium to long-term.

Sustainability - Minimum Investment and Ongoing Income and Projects

Without long term commitment similar to institutional investment in other key infrastructures any short term investment will be wasted. Depending upon the research profile of an HEI, both the start-up and maintenance of a research data repository may be substantial. Robust service and financial planning will be important to ensure both flexibility to deal with changing requirements and the demands of new projects. Service sustainability will come from inward investment; research grants; and externally funded development projects, together with the strategic drive and support from the institution.

Existing Infrastructure

HEIs may be able to leverage or augment parts of their existing support infrastructure in undertaking preservation of research data. This is particularly likely for most elements of IT and central support services in the activity model, data management in the Pre-Archive phase, and some elements of the Archive phase. The sharing of common services across a range of university functions not just preservation will bring economies of scale. An integrated approach may well influence start-up as well as recurrent costs.

TRAC

The institution will need to decide what is encompassed by a research data preservation infrastructure service and how to identify the full cost of the activity or service consistent with TRAC principles. Currently HEIs' costs associated with data preservation/archive activities are likely to form part of the support costs included in the indirect costs rates. To better identify the costs of these services in order to charge out their use would involve converting the costs of a support activity into a direct activity and then applying an appropriate share of indirect costs and estates costs. One possibility may be to consider whether you wish to identify it as a separate unit akin to a research support facility. This approach is being evaluated by one of our case study sites (see appendix 3). Care must be taken to avoid "double-counting" by adjusting existing TRAC calculations to reflect any changes made.

Start-up and recurrent

The start-up and implementation phase to establish a research data repository should be fully assessed. This will be separate from the operational phase and recurrent data lifecycle costs and usually funded through capital/infrastructure investment. The start-up and recurrent costs can be estimated using a cost model. You should factor in the cost of staff development as this is often overlooked. Training and developing capacity for people to do data curation and preservation work will be required.

CULTURAL

Disciplinary Differences

There are wide variations in the use and periods of retention of research data across disciplines. The inherent complexity and heterogeneity of research data sets brings with it challenging and conflicting requirements. Some disciples have established subject repositories and data centres providing solutions external to the institution for preservation, whereas others don't.

Consideration should be given to the cultural issues across different disciplines that might hinder the process of acquiring data, recording the materials and selecting the access rights.

Combining Subject Knowledge and Professional Expertise

In order for research data to be preserved cost effectively it is essential to implement appropriate standards and methodology throughout the life cycle. Both subject and preservation expertise are required to keep data long term and to keep it accessible.

Education and Training

Education and training is important for success and is not just a one-off cost at start-up but will also be recurrent. It is resource intensive but at the same time is important in order to ensure the early capture of data, related documentation, and software early in the lifecycle.

Operational

User Communities and User Expectations

As outlined in the summary description above of "Collection Levels", researchers in different subjects will be producing different types of data at different times and expectations as to how data should be treated will therefore change accordingly. Whether a data collection is a "research", "resource/community", or "reference" data collection will impact upon users' expectations of access and preservation for them. The difference in these requirements will feed into setting the variables in the cost model.

Repository Solution

This may well be influenced by the strategic decision made by the institution about the nature of the repository. The choice of repository solution whether it is open source or commercial can be challenging. There is an issue of maintenance and or adaptation of open source software systems and preservation tools when there is limited local developer time available. Where an institution has many repositories around the organisation that it intends to federate it is noted that this is an even more critical issue where typically a departmental repository may only have 1fte of staff resource for everything. Institutions will also need to consider carefully which repository solutions have the functionality and scalability to be adequate for their data requirements.

Capacity Planning

The broader plans for research within the university need to be fed into the infrastructure planning for research data preservation. This should be a continuing process to provide sufficient lead in to any cases for growth and expansion. Consideration of the timescales of decision making processes to approve funds is important to ensure smooth planning and implementation cycles.

Multi-institutional Projects

Clarity is required around archiving for multi-institutional projects – they are increasingly important and becoming the norm rather than the exception. They risk, if no one takes responsibility, important research outputs being lost over time.

Risk

Data Audit

There is a risk that an institution is not fully aware of the research data within its faculties and departments, nor the extent of potential future growth. Institutions need to take intellectual and physical control over what they want to preserve. One of the hardest tasks for HEIs is determining what they have and what they should keep. Data audits can help with this. They provide an idea of how much there is and who it belongs to. So this can be a huge help even if it doesn't answer all the preservation questions. A rolling data audit either via repositories within the institution or by some other means would be very important. Results from the audit, particularly in relation to risk management and critical data, feed into the development of preservation policy. However, a major difficulty is the scale of most large research-led universities. It is hard to establish an institution-wide view or undertake an audit and almost universally this global view does not exist for research data in universities. Most HE institutions are heavily de-centralised around individual departments and centres. In recognition of a community requirement in this area, the JISC is funding work, including the development of a 'Data Audit Framework' to enable all universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation.

Legal Issues

There are perceived issues around the holding of sensitive data. Universities are risk averse organisations. There may be concerns over holding onto data that contains personal information, largely due to concern over confidentiality issues. This increases the cost of ingest for some disciplines as another set of questions have to be addressed when bringing in data or providing access to it. Risk can be managed if researchers and "experts" work together throughout the data life cycle, thereby applying guidance about longer term access and any necessary constraints and controls that need to be put in place. Advice from the beginning means that things like rights issues are dealt with from the start. Unless rights are clear and proper procedures and standards are adhered to there could be potential problems around access and IPR. In considering the impact on costs, HEIs need to take into account the time required to investigate legal issues and in keeping up to date with case law.

10. Different Service Models and Structures

Introduction

The terms of reference for the study include comparison of different service models and structures for preservation of research data. The case studies (appendices 1-4) provide detailed illustrations of indicative costs for different forms of provision. The purpose of the following discussion is to synthesise the relevant findings of the case studies and integrate into this additional information from the desk research and interviews.

The scope and timescale of the study and the evolving nature of service provision mean that very small samples are currently available in discussing costs. As noted in previous chapters preservation costs can be expected to vary and a wide-range of factors needed to be considered in order to make them comparable between different institutions and collections. Therefore costs given below must be regarded as broad illustrations only. It should be noted that there is an evolving landscape of service provision for research data and in several areas data for operational costs is not available yet as services are at the feasibility or piloting stage. In some disciplines national data centres have existed for many years and costs are well established for them. However they remain relatively localised to a few disciplines.

Discussion and comparison in this section has focused principally on HEIs particularly local repositories within institutions, and in addition discussion of national disciplinary or subject data centres, national shared services, and centralised repositories. We have defined centralised repositories for this study as encompassing national laboratories or research centres.

Local Repositories

Lyon notes that during the last three years, in the UK we have seen an increasing investment in institutional repositories (IR) though as yet, there are few examples of IRs containing research data, either raw or processed (Lyon 2007). The focus of the Institutional Repository movement has principally been on electronic publications particularly eprints and it has been possible to develop a single centralised open-access repository for the whole institution for these materials. The SHERPA project has suggested the following typical costs for a UK institutional repository (Pinfield and Hubbard 2004):

Installation costs:

Server £1,500

Software £0 [where open source software is used there is no purchase cost but the total costs of ownership over time will reflect other costs that will be specific to open source software]

Installation (5 days) £600

Customisation (15 days) £1,800

Total per institution £3,900

Ongoing maintenance costs:

Technical support Absorbed by institutional IT services

Supported archiving service £30,000 per year

Upgrades/migrations £3,900 every 3 years

Digital preservation Significant costs (applies to all digital objects)

Summary Table of Suggested Costs above for an Institutional Repository (e-publications):

	Staff	Equipment
Installation	(£2,400)	£1,500
Annual recurrent costs	1 FTE	£1,300

More recently The Repositories Support Project has provided typical hardware costs for four repositories. Excluding Cambridge (discussed further below), the hardware costs for the three other institutional repositories were given as £1945, £5000 and £18000, which on a straight-line depreciation over three years would be £648, £1666, and £6000 for annual recurrent costs respectively. The largest figure is for Southampton with a capacity for c.1,000,000 documents substantially larger than most IRs in the UK (Repositories Support Programme 2008).

The Digital Repositories Roadmap has noted that:

> "institutions need to invest in research data repositories..." and "...culturally, data repositories and have been the property of "scientists" and there is some tension between the data and information community. Institutional repositories could fill a gap where there is no data archive…"(Heery and Powell 2006).

However our case studies suggest that the service requirements for data collections and the best structure for organising relevant services locally will be more complex than this may suggest. It is notable that both Cambridge and KCL in our case studies are developing central repositories to work with departmental facilities and discussing **federated local data repositories** for research data preservation combining services and skills from central and departmental repositories with data distributed and located at different repositories in the institution. A similar discussion and scoping project is also currently underway at the University of Oxford (Martinez-Uribe 2008).

In the case of Cambridge the central repository (DSpace@Cambridge) is also an institutional repository for other research outputs and the central repository is specifically being developed to provide preservation expertise.

In the case of KCL the central preservation expertise and repository are based in the new Centre for E-Research established from the former Arts and Humanities Data Service Executive.

As stated in our Southampton case study a similar federated data structure could also emerge in Southampton involving departmental repositories, the central institutional repository, and probably use of a central mass Storage Area Network for bit-archiving.

As noted in the Southampton case study there are a number of reasons why these federated structures may be more appropriate for data management than a single institutional repository in HEIs:

> “It is however also important to consider the Department level in this landscape, in addition to the overall institutional level. It is an academics natural affiliation and an environment they understand and can often have an influence on, i.e. it is at this level where money can be raised and decisions surrounding ‘what is important’ can be made by the most appropriate people. Individual researchers are likely to feel alienated if archiving only occurs at an institutional level. At the individual school level this cost model remains highly pertinent when planning for preservation, however it is worthwhile noting that the School of Chemistry does not currently consider these issues, so advocacy and hands on help are required. A recommendation might be that a data store at the departmental level should be incorporated into the federation of institutional repositories and that the hierarchy of data stores reflects the detailed nature of the content and the changing nature of its importance over time.”

Services for the federated local repositories do not need to be all local to the institution. Potentially there is considerable scope for economies of scale and leveraging rare skills and experience across HEIs through either shared services, or where they exist disciplinary data centres or centralised repositories at national level. For example two of our case study sites (KCL and Southampton) currently out-source archival storage to the Atlas Data Store a central repository maintained by the Science and Technology Facilities Council (STFC). Researchers across all HEIs will also utilise disciplinary data centres where these exist and the centralisation of expertise and services and economies of scale that this may provide.

Finally it should be noted that based on our case studies costs for the central data repository component of federated local data repositories is likely to be an order of magnitude greater than that for a typical institutional repository focused on e-publications: this reflects the need for higher staffing levels for ingest and user support and much larger storage requirements as summarised below:

	Staffing	Data Storage and Equipment
Cambridge	4 FTE	£58,764 pa[1]
KCL	2.5 FTE	£27,546[2]

Notes

1. 150Tb mirrored storage. Capital purchase costs depreciated over 3 years.
2. Equipment purchase 2005. Capital purchase costs depreciated over 3 years.

Disciplinary Data Centres

National disciplinary data centres are currently funded by 3 of the research councils. The Natural Environment Research Council (NERC) funds 8 disciplinary data centres covering: atmospheric science; earth science; marine science; polar science; terrestrial and freshwater science; hydrology; and bioinformatics (environmental genomics). Two of these data centres are hosted by the Rutherford Appleton Laboratory within the Science and Technology Facilities Council. The Economic and Social Research Council (ESRC) funds the UK Data Archive at the University of Essex which hosts a number of services and projects and is the lead partner (with MIMAS and CCSR, School of Social Sciences at the University of Manchester) in the Economic and Social Research Council (ESRC) and JISC funded Economic and Social Data Service. The Arts and Humanities Research Council is funding the five service providers (archaeology, history, literature/languages and linguistics, performing arts, and visual arts) within the Arts and Humanities Data Service until 31 March 2008 – thereafter it will only fund archaeology.

The focus of the national disciplinary data centres are data collections at the resource/community and reference levels and for materials requiring long-term preservation and data sharing. The research councils involved mandate that research data from projects that they fund must offer data generated by the research to one of their national data centres. These data centres also tend to hold or act as brokers for their disciplines for datasets from outside HEIs and research from sources such as local and central government which are best managed nationally rather than in individual HEIs. The staffing and service levels of the national data centres reflect these requirements for community and reference collections and long-term preservation.

The Office of Science and Innovation (OSI) working group for preservation and curation found the running costs of the data centres across these three research councils to be remarkably consistent at between 1.4 and 1.5% of the total research expenditure of the research council (Beagrie 2006).

The profile of costs across functions within the data centres also appears to be very consistent. The following approximate division of costs across high-level archive functions of the activity model were suggested for the UK Data Archive (Matthew Woollard 29/1/2008 pers comm.):

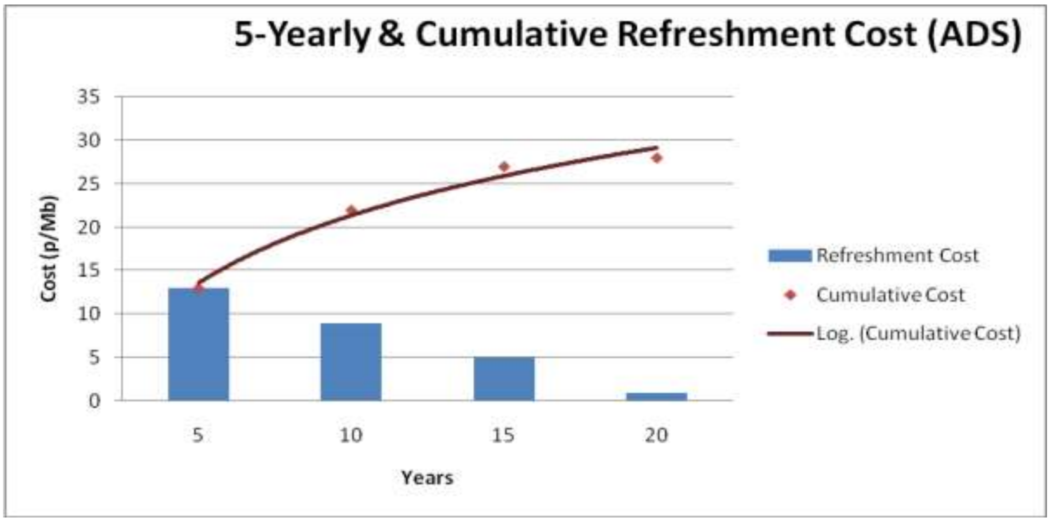
Acquisition and Ingest [1,4]	Archival Storage and Preservation [2,4]	Access [3,4]
c. 42%	c. 23%	c. 35%

Notes

1. includes preparation of metadata for resource discovery
2. includes preservation planning and data management
3. includes user-support both on finding and using.
4. Costs for administration and management functions are included (unevenly) within these headings and based on adjusted salaries.

It is interesting to note that NERC also believes that the major costs for its data centres are in accessioning rather than archival storage and preservation. Mark Thorley suggests separation between the relatively expensive accession cost – ensuring effective documentation and formats for re-use to the standard of " passing the 20 year test" i.e. is it capable of being understood 20 years from now without reference to the original PI and research team who may not be available - and the relatively less expensive long-term cost of maintaining the bits and bytes and metadata. Mark notes the one-off accession cost to appropriate standards for NERC data centres can be quite high. The fact that once you have paid the accession cost the rest is relatively small has implications for de-accessioning – even if use of a dataset falls, staff costs for appraisal and de-selection may outweigh any savings on future storage and maintenance (Mark Thorley pers comm. 5/2/2008).

Similar data can be seen in the Archaeology Data Service for costs as reflected in its charging model based on experience of costs over the past 10 years (case study - appendix 1). Early year costs for accessioning and ingest, preservation and archiving are highest. Note archival storage and preservation (“refreshment”) costs are seen as declining to minimal levels over 20 years as shown below.



The high percentage of expenditure by national data centres on acquisition and ingest represents both the level of validation and added-value needed for use of community and reference level data collections, and their procedures to reduce long-term preservation costs by early quality control of metadata, documentation and IPR and standardisation of file formats.

Finally it is worth noting a recent innovation at the UK Data Archive which intends to launch UKDA-Store a “self-archiving data repository”. It will enable researchers to submit a range of digital outputs with the right to set permissions for individual and group access, so that data can remain private (on embargo) although metadata continues to be searchable. Furthermore, data that is judged to meet the UKDA’s acquisition criteria can be formally lodged for long-term preservation within the UK Data Archive. This potentially extends the range of data collection levels at UKDA to the basic research data collections level and will allow review and investment to upgrade them to community collections if necessary.

National Shared Services

Within the study interviews we would suggest the national preservation services offered by the University of London Computer Centre (ULCC) might be considered under this category, particularly its service for the National Digital Archive of Datasets (NDAD). NDAD is in many ways similar to national disciplinary data centres in terms of its operations and costs and examples have been used as illustrations earlier in the study (pages 25-6).

RLUK (Research Libraries UK), and RUGIT (the Russell Group IT Directors Group) have just launched a UK Research Data Service (UKRDS) feasibility study. It is funded by HEFCE (the Higher Education Funding Council for England) under its Shared Services programme, with support from JISC (the Joint Information Systems Committee).The objective of the UKRDS study is to assess the feasibility and costs of developing and maintaining a national shared digital research data service for UK Higher Education sector. Such a research data service is seen by the project sponsors as forming a crucial component of the UK's e-infrastructure for research and innovation, and one which will add significantly to the UK's global competitiveness.

A range of options for national shared services have also been explored for preservation across institutional repositories in the JISC funded PRESERV and SherpaDP projects. Experience from these could be applicable for some file formats in data repositories although these are not the primary market or focus of the pilot projects. The pilots do not have any real costs data as yet for operational services, since they are still investigating a range of options (Hitchcock et al 2007, Steve Hitchcock pers comm. 13/3/08).

Centralised Repositories (national laboratory or research centre)

Centralised repositories are also comparatively rare. Within the case study sites they could include the National Crystallography Data Service at Southampton and utilisation of the Atlas Data Store (part of the national facilities maintained by the Science and Technology Facilities Council) by both NCS at Southampton and the AHDS at KCL. Outsourcing to Atlas has allowed the NCS at Southampton to reduce costs for archival storage by 41% between when this was an in-house and staff-intensive and when this was outsourced and highly automated.

National research centres which might be considered to be centralised repositories could be defined to include individual data collections in HEIs funded as national data resources. The funding by MRC of the Avon Longitudinal Study of Parents & Children (ALSPAC) based in the University of Bristol, and the National Survey of Health and Development (NSHD) based in University College London might fall in this category. Other national research centres are hosted by national laboratories: examples at Rutherford Appleton Laboratory include the Chemical Database Service and MRC Psycholinguistic Database.

11. Conclusions and Recommendations

This has been an intensive study over a period of 4 months focusing on the issue of the preservation costs of research data for UK HEIs. Our conclusions and recommendations from the study are provided below.

1. We believe the study has produced a robust framework for costing and detailed illustrative case studies which will assist and inform individual universities, departments and research groups needing to address the issue of maintaining research data over time within their institutions.
2. The study has identified the need for capacity planning and its potential link into recurrent data audits. It has also indentified a methodology which is suitable for assessing costs for different service delivery mechanisms for preservation of research data. JISC is initiating the development of a data audit framework and the outcomes of our study may be of interest to that initiative.

Recommendation 1: The outcomes of this study should be considered and utilised by the forthcoming JISC Data Audit Framework.

Recommendation 2: Departments and Central Services within HEIs should utilise recurrent data audits to inform both their initial appraisal and development of data policies and future capacity planning for services.

3. In considering the preservation of research data institutions will need to consider a wide range of disciplinary requirements. For different disciplines, research data management and preservation responsibilities may exist at local, national, and sometimes international level. However it is likely that some responsibility will always remain at the local level and therefore issues raised in this study should be of interest to all HEIs.
4. Service requirements for different data collections are also likely to vary considerable with data having different value and requirements for access over time. We suggest that institutions are likely to find the categories of research data collections, resource

or community data collections, and reference data collection levels proposed by the NSB long-lived data study to be of considerable value in understanding and categorising user requirements and costs over time. They should also note that significant costs are associated with moving data collections from one level to another over time. These collection enhancements may need to be funded separately on a case by case basis. We have discussed and provided examples in Chapter 9 and extracts of these data collection levels in Appendix 7.

Recommendation 3: HEIs should consider utilising the US National Science Board (the governing body for the National Science Foundation) long-lived data collection levels to aid understanding and categorisation of user requirements and costs over time.

5. Our case studies provide examples of a number of different service models and structures for research data preservation. HEIs will need to consider how best to achieve the right mix of skills and cost efficiencies for their own needs. We would note that research data is not as homogenous as research publications and is less likely to be available through a single institutional repository. We would note that subject knowledge, preservation and curation skills are needed for long-term management of research data and that the staffing and storage requirements will be more substantial than for eprint repositories. In some disciplines national and occasionally international data repositories will be available and can be utilised. Management of research data in institutions is therefore most likely to be federated with a mixture of skills and support from departmental and central services within the institution and/or mixed with external shared services or national provision.

Recommendation 4: HEIs should consider federated structures for local data storage within their institution comprising data stores at the departmental level and additional storage and services at the institutional level. These should be mixed with external shared services or national provision as required. HEIs should work with and utilise national and international disciplinary archives where these exist. The hierarchy of data stores should reflect the detailed nature of the content, services required, and the changing nature of its importance over time.

6. We have suggested mechanisms within the study for how HEIs could apply the framework and develop sustainable infrastructure to meet requirements for long-term management preservation of research data. HEIs should apply the framework with understanding of the national provision for archiving by research funders where this exists and funders' guidance on proposal costings and grant requirements.

7. Although this study focuses on the requirements of HEIs we believe its work will also be of interest to research funders and national data services amongst others. Costing and preservation of research data are complex subjects with a need for ongoing work and discussion between HEIs, funders and service providers.

Recommendation 5: We recommend consideration of the study and further work on development and implementation of relevant cost models and tools to HEIs, research funders, and service providers.

8. In addition to disseminating this report we believe there would be value in JISC producing a short summary of this report and its findings aimed at senior managers including university academics, administrators and research support services.

Recommendation 6: JISC should produce a short briefing paper or summary of this report and its findings aimed at senior managers including university academics, administrators and research support services.

9. There are a number of ways in which JISC could build on this study and assist institutions and individual researchers and research groups with implementation of its findings.

10. Project Costing Tools. We believe the framework should be implemented by means of automated interactive tools such as spreadsheets to build up estimates of costs. The future development of this study that may be of most value to Universities and to individual academics could be the development of tools for estimating costs, using the FEC model, for data management and archiving particular types of data or for particular disciplinary data collections over defined periods of time. The tools and figures provided can then be used: in proposals by applicants to secure funding to cover the cost of the long-term preservation of the data; in evaluation of proposals by

peer reviewers to provide a guideline of reasonable costs; by research funders in support of data management, data sharing, or preservation plans and strategies for research quality assurance, knowledge transfer, and demonstrating research impact and value.

Recommendation 7: JISC should consider developing project costing tools to build on and implement work within this study. These tools may be valuable for some of JISC's own projects and may also be of interest to other research funders and have potential for joint funding and development.

11. The duration of the study has allowed us to develop our approach and collect sample data based upon it. However the timescale has not allowed us to research the data we have begun to gather in any depth or across a larger sample of data collections. We believe both the approach and data that have emerged are important and should be further developed and researched. In particular the cost variables and dependencies outlined and the key variables and initial data for long-term digital preservation costs are potentially very significant additions to existing knowledge but need further work to quantify, validate, and operationalise them.

Recommendation 8: JISC should consider undertaking additional work to examine how the cost components and variables defined in our framework can be further quantified, and what additional data and data collection mechanisms are needed to support them.

12. JISC is participating in a two-year international taskforce examining digital preservation costs which commenced in January 2008. Digital preservation costs are notoriously difficult to address in part because of the absence of good case studies and longitudinal data for digital preservation costs or cost variables. We believe we have identified valuable data both within our case study sites and in a number of other national data centres, services and projects which would re-pay further detailed study over a longer timescale. In particular we would point to possible data within the UK Data Archive, University of London Computer Centre, the NERC Data Centres, and long-term projects in some universities which could contribute to the taskforce findings.

Recommendation 9: JISC should consider further detailed study of longitudinal data for digital preservation costs and cost variables to extend the work of this study. Possibly this could be part of a UK based taskforce to feed into its joint international work on digital preservation costs.

13. The study has included a brief overview of the benefits of preservation of research data in addition to our consideration of costs. This is an addition to requirements in the original invitation to tender but was considered important additional context. We would note that relatively little work has been done on quantifying the benefits of research data preservation and that further work in this area would be desirable.

Recommendation 10: JISC and /or other funders should consider funding further work on quantifying the benefits of research data preservation.

References and Sources Consulted in Desk Research

Archaeology Data Service (ADS), 2007, *Charging Policy 4th Edition November 2007*. Retrieved 7/2/08 from http://ads.ahds.ac.uk/project/userinfo/charging.html

Ashley, K., 1999, *Digital Archive Costs: Facts and Fallacies*. Retrieved 14/1/08 from http://ec.europa.eu/archives/ISPO/dlm/fulltext/full_ashl_en.htm

Beagrie, N., and Jones, M., 2001, *Preservation Management of Digital Materials: a Handbook*, (British Library).

Beagrie, N., 2006, *e-Infrastructure Strategy for Research: Final Report from the OSI Preservation and Curation Working Group November 2006*, (National e-Science Centre). Retrieved 10/12/07 from http://www.nesc.ac.uk/documents/OSI/preservation.pdf

Booth, B., Banks, M., and Hunolt, G., 2006, *Cost Estimation Tool Enhanced Operational Comparables Database.* Retrieved 3/1/08, from http://opensource.gsfc.nasa.gov/projects/CET/Doc.zip

Chapman, S., 2003, Counting the Costs of Digital Preservation: Is Repository Storage Affordable? *Journal of Digital Information , Volume 4* Issue 2 Article No. 178, 2003-05-07

Consultative Committee for Space Data Systems (CCSDS), 2002, *Recommendation for Space Data System Standards: Reference Model for an Open Archival Information System (OAIS).*

Consultative Committee for Space Data Systems (CCSDS), 2004, *Producer-Archive Interface Methodology Abstract Standard,* CCSDS 651.0-B-1 BLUE BOOK May 2004. Retrieved 14/3/08 from http://public.ccsds.org/publications/archive/651x0b1.pdf

Currall, J., & McKinney, P., 2006, Investing in Value: a Perspective on Digital Preservation. *D-Lib Magazine*, 12 (4) April 2006. Retrieved 3/1/08 from http://www.dlib.org/dlib/april06/mckinney/04mckinney.html

ERPANET, 2003, *Cost Orientation Tool.* Retrieved 3/1/08 from http://www.erpanet.org/guidance/docs/ERPANETCostingTool.pdf

European Commission, 2007, Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on Scientific Information in the Digital Age: Access, Dissemination and Preservation (Commission of the European Communities).

Fontaine, K., Hunolt, G., Booth, A., and Banks, M., 2007, *Observations on Cost Modeling and Performance Measurement of Long Term Archives* in PV2007 Conference Proceedings retrieved 3/1/08 from http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf

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

Gray, J., Szalay, A. S., Thakar, A. R., Stoughton, C., vandenBerg, J., 2002, *Online Scientific Data Curation, Publication, and Archiving*, Microsoft Research Technical Report MSR-TR-2002-74. Retrieved 3/1/08 from http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-2002-74

Heery, R. and Powell, A., 2006, *Digital Repositories Roadmap: looking forward* (UKOLN, University of Bath; Eduserv Foundation) retrieved 1/3/08 from http://www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc

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

Hendley, T. (1998). *Comparison of Methods & Costs of Digital Preservation.* (British Library Research and Innovation Centre).

Her Majesty's Stationary Office (HMSO), 2004, *Science and innovation investment framework 2004-2014*, (Her Majesty's Stationary Office, London).

Hey, T., and Trefethen, A., 2003, “The Data Deluge: an e-science Perspective” in: Berman, Fran (Ed.) et al, 2003, *Grid Computing: Making the Global Infrastructure a Reality*, (John Wiley and Sons).

Higgins, S., 2007, Draft DCC Curation Lifecycle Model, *The International Journal of Digital Curation*, Issue 2, Volume 2, 82-6

Higher Education Funding Council for England (HEFCE), 2007, *HEFCE strategic plan 2006-11(Updated April 2007)*. Retrieved from http://www.hefce.ac.uk/pubs/hefce/2007/07_09/

Hitchcock, S., Brody, T., Hey, J., and Carr, L., Digital Preservation Service Provider Models for Institutional Repositories: Towards Distributed Services, *D-Lib Magazine* Volume 13 Number 5/6 May/June 2007. Retrieved 13/3/08 from http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html

Hunolt, G., 2006a, *Users' Guide Cost Estimation Toolkit (CET)*, version 2.1 September 2006. Retrieved 3/1/08, from http://opensource.gsfc.nasa.gov/projects/CET/Doc.zip

Hunolt, G., 2006b, *Technical Description Document Cost Estimation Toolkit (CET)*, version 2.1 September 2006. Retrieved 3/1/08, from http://opensource.gsfc.nasa.gov/projects/CET/Doc.zip

Hunolt, G., Booth, B., Banks, M., 2006, *Cost Estimation Toolkit (CET)*, version 2.1 September 2006. Retrieved 3/1/08, from http://opensource.gsfc.nasa.gov/projects/CET/CET%20V2p1.xls

Hunter, L., 2006, *Investment in an Intangible Asset.* Retrieved 3/1/08 from DCC Digital Curation Manual at http://www.dcc.ac.uk/resource/curation-manual/chapters/intangible-asset/intangible-asset.pdf

International Council for Science. 2004. *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information* (International Council for Science).

Joint Costing & Pricing Group, 2005, Transparent Approach to Costing (TRAC) Manual. Retrieved 3/1/08 from http://www.jcpsg.ac.uk/guidance/downloads/TRACmanual.zip

Joint Costing & Pricing Group, 2006, Annex 15 Cost Adjustments. Issued July 2006. Retrieved 25/4/08 from http://www.jcpsg.ac.uk/guidance/revisions/annex15.doc

Joint Costing & Pricing Group, 2008, Revisions to TRAC guidance (webpage listing documents issued at various dates 2006-2007). Retrieved 25/4/08 from http://www.jcpsg.ac.uk/guidance/revisions/

Joint Information Systems Committee, 2007, JISC Circular 05/07: Digital Repositories. Retrieved 3/1/08 from http://www.jisc.ac.uk/media/documents/funding/2007/11/jisccircular507digitalrepositories.doc

Lavoie, B., 2003, *The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making.* Dublin, Ohio: OCLC Research.

Lavoie, B., 2004, *The Open Archival Information System Reference Model: Introductory Guide*, DPC Technology Watch Series Report 04-01 January 2004. Retrieved 3/1/08 from http://www.dpconline.org/docs/lavoie_OAIS.pdf

Linden,J.,Martin,S., Masters,R., and Parker,R., 2005, *The large-scale archival storage of digital objects*, DPC Technology Watch Series Report 04-03 February 2005. Retrieved 3/1/08 from http://www.dpconline.org/docs/dpctw04-03.pdf

Lord, P., and Macdonald, A., 2003, *e-Science curation report* (Joint Information Systems Committee)

Lyon, E., 2007, Dealing with Data: Roles, Rights, Responsibilities and Relationships (UKOLN University of Bath). Retrieved 3/1/08 from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

Martinez-Uribe, L., 2008, *Scoping Digital Repository Services for Research Data Management: Project Plan*, v2.2 date 27/2/08 (University of Oxford). Retrieved 20/4/08 from http://www.ict.ox.ac.uk/odit/projects/digitalrepository/docs/DigRepoProjectPlan.pdf

Mcleod, R., Wheatley, P., & Ayris, P. (2006). *Lifecycle information for e-literature: full report from the LIFE project.* (LIFE Project, London, UK). Retrieved 10/12/07 from http://eprints.ucl.ac.uk/archive/00001854/01/LifeProjMaster.pdf

Nationaal Archief, 2005a, *Costs of Digital Preservation* version 1.0 May 2005 (Digital Preservation Testbed, The Hague, Netherlands). Retrieved 3/1/08 from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf

Nationaal Archief, 2005b, *Digital Preservation Costmodel* version 1.0 20 April 2005 (Digital Preservation Testbed, The Hague, Netherlands). Retrieved 3/1/08 from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Kostenmodel_in_Excel_versie_1.0_(final).xls

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

OECD, 2004, *Declaration on Access to Research Data from Public Funding*. (Organisation for Economic Co-operation and Development, Paris). Retrieved 20/12/07 from http://www.codataweb.org/UNESCOmtg/dryden-declaration.pdf

OECD, 2007, *Principles and Guidelines for Access to Research Data from Public Funding*. (Organisation for Economic Co-operation and Development, Paris). Retrieved 20/12/07 from http://www.oecd.org/dataoecd/9/61/38500813.pdf

Pinfield, S., and Hubbard, B., 2004, *Establishing a National Network of Repositories - Supplementary Evidence for the Parliamentary Inquiry into Scientific Communication*. Retrieved 20/2/08 from http://www.sherpa.ac.uk/documents/SHERPA_Supplementary_Evidence.pdf

Repositories Support Programme, 2008, *Typical Hardware costs* in “Making a Business Case”. Webpage last reviewed 29 January 2008. Retrieved 25/04/08 from http://www.rsp.ac.uk/repos/business

Research Information Network, 2007, *Research Funders’ Policies for the management of information outputs*. Retrived 23/4/08 from http://www.rin.ac.uk/files/Funders'%20Policy%20&%20Practice%20-%20Final%20Report.pdf

Research Information Network, 2008, Stewardship of digital research data: a framework of principles and guidelines. Retrieved 23/4/08 from http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf

Sanett, S., 2003, Cost to Preserve Authentic Electronic Records in Perpetuity: Comparing Costs Across Cost Models and Cost Frameworks, *RLG DigiNews ,* volume 7 No.4 15 August 2003. Retrieved 3/1/08 from http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070511:000006283731&reqid=7171#feature2

Tessella, 2006, *Mind the Gap: Assessing Digital Preservation Needs in the UK* (Digital Preservation Coalition, York).

Watson, J.,2005, *LIFE Project Research Review : Mapping the Landscape, Riding a Life Cycle,* (LIFE Project, London, UK). Retrieved 20/12/07 from http://eprints.ucl.ac.uk/1856/1/review.pdf

Wheatley, P., Ayris, P., Davies, R., Mcleod, R., and Shenton, H., 2007, *The LIFE Model v1.1.* Discussion paper, (LIFE Project, London, UK). Retrieved 20/12/07 from http://eprints.ucl.ac.uk/archive/00004831/

Woodyard-Robinson, D., 2006, 3.7 Costs and business modelling in Beagrie, N., and Jones M., *Preservation Management of Digital Materials.* DPC online edition retrieved 3/1/08 from http://www.dpconline.org/graphics/inststrat/costs.html

APPENDICES

APPENDIX 1- CASE STUDY : ARCHAEOLOGY DATA SERVICE CHARGING POLICY

PREAMBLE

This case study text is based on edited extracts from the Archaeology Data Service (ADS) Charging Policy (ADS 2007) with some re-arrangement and additional material and commentary by the study team. The ADS Charging Policy can be consulted in full at http://ads.ahds.ac.uk/project/userinfo/charging.html.

The ADS Charging Policy has been included as an additional case study for a number of key reasons. It has been mentioned by several of our interviewees and main case study sites as an example of evaluation of costs and therefore charges; it has been developed over a 10 year period and provides a useful and still relatively rare historic perspective on preservation costs; and finally it is influencing charging policies being developed by other research data repositories for example the History Data Service (part of the UK Data Archive) at Essex University and the Oxford Text Archive at Oxford University and may therefore be of wider interest.

INTRODUCTION

The ADS was established in 1996 and is part of a long-established tradition for archiving data within Archaeology. Professional ethics within the archaeological community require that access to primary data should be free at the point of use. This approach has been extended to digital archives, although it is accepted that in order to recoup the ongoing costs of digital preservation, some means of cost recovery is essential. Within the Archaeology Data Service (ADS) this led, in the late 1990s, to the introduction of a charging policy. The central tenets of this policy remain that:

- ADS resources will be freely accessible;
- archiving costs should be recovered from the body funding the archaeological investigation, or research;

- a one-off payment collected at the time of deposit will be used to safeguard the long-term future of digital data.

The latest edition of the charging policy (ADS 2007) reviews the categories of depositors for which charging has been implemented and defines the new level of charges in operation. A refined level of charging has been introduced in order to reflect the increasing volume (both in file size and number of files) of an average deposit and the related storage and refreshment costs for digital data. The ADS currently receives some core funding from the Arts and Humanities Research Council (AHRC). The AHRC have indicated that the ADS should investigate a move toward a responsive mode funding for archives created by AHRC funded projects in the long term.

The purpose of the ADS Charging Policy is to make the scale of charges explicit and open so that those preparing project applications are able to allow for appropriate additional costs to cover digital archiving. Note ADS still requires all potential depositors to contact the ADS Collections Development Manager prior to the submission of project designs including ADS costs.

THE CHARGING MODEL

The cost is calculated on the basis of four elements:

a. Management and administration: The cost reflects the time spent in processing the deposit, including negotiation with the depositor, dealing with rights management issues and deposit licences and issuing invoices. For most straightforward archives this will amount to one or two days of the Collections Development Manager (CDM) and a day of the Administrator, costed at current daily rates plus overheads, but more substantial projects of long duration may typically require in excess of ten days of CDM time.

b. Ingest: The costs reflect the number of ADS staff days necessary to migrate the data to ADS preferred formats; the harmonisation of filenames, the creation of delivery and preservation formats and their transfer to offline storage, checksum procedures, and creation of file level and project level metadata, and its entry within the ADS Collections Management System. The time required will be dependent on the number and complexity of files deposited.

For simplification files are banded according to number and complexity of format.

Images, text, simple 'flat' spreadsheets and tables cost less than CAD, GIS and relational databases, for example. The difference in cost is a result of the migration and validation aspects of digital archiving. For example, to assess the success of the migration of an image file it only has to be looked at, whereas for a more complicated file type, such as a GIS, it is necessary to ensure that the full functionality of the file has been preserved during its migration. The latter process takes more time and hence costs more.

Deposit type (download archive only)	**Minimum charge in days**
For deposits of text and image files only:	
1 - 10 files	1 day
11 - 100 files	2 days
100+ files	4 days
For deposits of mixed files including GIS, CAD, Geophysics, Databases etc.	
1 - 10 files	2 days
11 - 100 files	3 days
100+ files	6 days
Archives of over 1,000 files	By arrangement

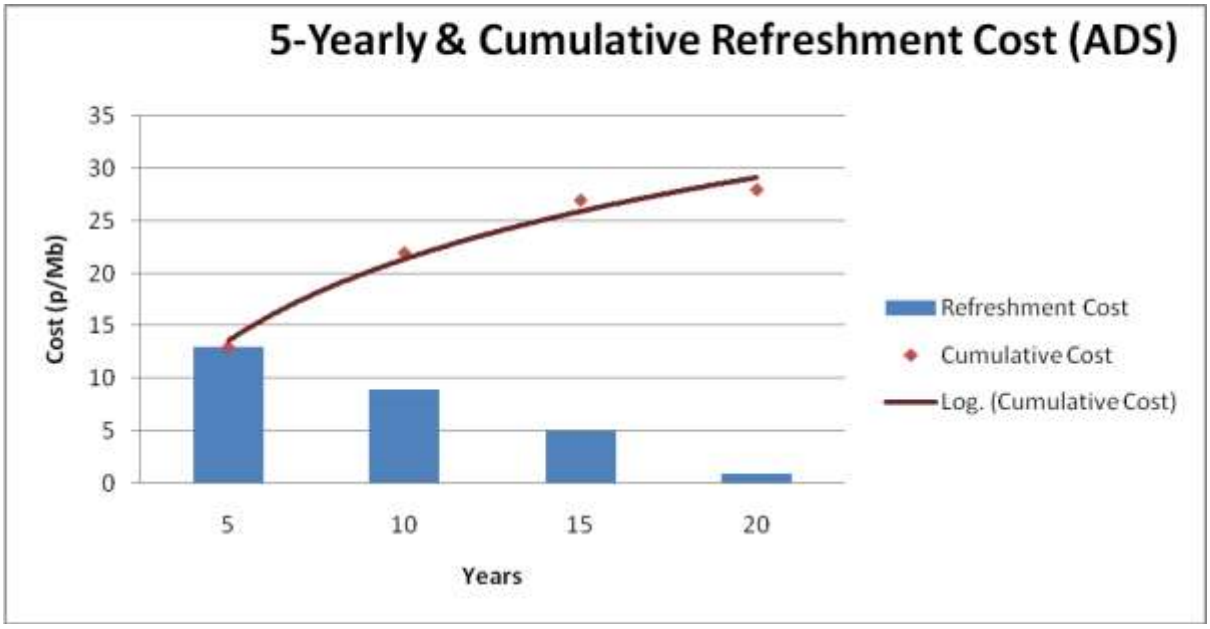
c. Dissemination: The file-based ingest charges itemised above include an allowance to cover the creation of a basic archive delivery web page, within the ADS catalogue, and the delivery of data via simple file download. Should the depositor require a special interface (on-line searchable database, interactive map interface etc) then such services will be charged at the current ADS day rates, plus university overheads on staff costs. As the interface requirements for each project will be unique these must be subject to price on request, but as an approximate indicator, an online queriable database may generally cost c. £1000-£5000, whilst a fully-functional GIS interface may cost as much as £10,000.

d. Storage and refreshment. The term 'storage' encompasses the ongoing periodic process of data refreshment. In order to take advantage of technological advances and hardware and software changes, archives have to periodically upgrade systems or parts thereof. As an example, during 10 years of existence, the ADS has progressed through three generations of equipment. Thus it is operating on a five year upgrade cycle. This is expensive both in terms of equipment and staff time. The long term cost of storage is often difficult to conceptualize but a dataset maintained for 100 years would go through 20 refreshments based on the five year cycle.

From ADS experience the cost of refreshment for a given resource decreases with time as archival systems become more sophisticated and a given archive becomes an increasingly smaller part (presuming archival growth) of a periodic refreshment. Thus there is a gradual decrease in the cost of refreshing a given ADS resource although this is partially offset by the increasing cost in terms of human resource (i.e. increasing wages). Between refreshments the ongoing management and administration within an OAIS framework is proactive and similarly subject to increasing costs in terms of human resource.

In contrast the cost of physical disc storage and back up media such as tape decreases rapidly. Currently the cost of a gigabyte of disc storage can be as low as 7p. Analysis of past and current trends suggests this will be 1p in five years time and so negligible not long after that to be considered as zero cost. However, the capital cost of the systems associated with such storage can be substantial as can ongoing maintenance, backup and insurance costs. Like disc storage systems they consistently fall in price but still remain a significant cost over time.

The test of time suggests that so far the one off 50p per megabyte charge in the previous ADS charging policy is near the mark for an earlier archival tradition. Recent developments, however, in terms of systems upgrades suggests the 50p charge can be reduced significantly. The 'per megabyte' charge is shorthand for what has been described above which might be better described today as 'ongoing management and refreshment'. The following is simplistic but attempts represent more accurately the current situation of lifecycle management with its associated retention and discard policies.



Retention period	Cost for refreshment
5 years	R + E
10 years	R - DR + E - DE
15 years	R - 2DR + E - 2DE
20 years	R - 3DR + E - 3DE
25 years	R - 4DR + E - 4DE

Where R = refreshment cost = 9p per megabyte DR = decreasing cost of refreshment = 3p

E = cost of physical equipment = 4p DE = decreasing cost of equipment = 1p

As an example, if R = 9p, DR = 3p, E = 4p and DE = 1p (all pence per megabyte charges - please note these figures should be close to a final policy but are subject to an ongoing examination of past processes) then:

Retention period	Cost for refreshment	Cumulative total (pence)
5 years	9 + 4 = 13	13
10 years	9 - 3 + 4 - 1 = 9	22
15 years	9 - 6 + 4 - 2 = 5	27
20 years	9 - 9 + 4 - 3 = [1]	28
ongoing		30

The above one off costs suggests that preservation costs become negligible after 20 years. This is, to a degree, a product of the simplicity of the model as clearly there will be ongoing costs beyond this point in terms of the refreshment, management and administration of a resource should a retention policy dictate it. Thus a one off charge of 30p per megabyte would cover ongoing preservation beyond 20 years. ADS policy is currently based on the assumption that 'best efforts' will be used to preserve all data deposited with ADS into perpetuity (i.e., following the 20-year cost-model above).

ADS Requirements and Additional Cost Adjustments

- Published charges apply where data sets are supplied via the download archive in ADS delivery formats, with accompanying documentation, as stipulated in the Guidelines for Depositors. Where other formats are involved, or where the data as supplied to the ADS requires the attention of a member of ADS staff, prices are supplied on request;
- Data must be supplied with either an OASIS record id number or a metadata record for the project. A template for the metadata record is supplied;
- Relevant documentation, depending on the file type deposited. See http://ads.ahds.ac.uk/project/userinfo/deposit.html documenting for more details;
- Introductory and Overview texts describing the project and the dataset deposited. The Introductory text will summarise the aims and objectives of the initial research project, the data collected, and any special characteristics of the digital archive. The Overview text will provide an outline contents list of the archive, along with a guide to the documentation required for its reuse;
- Special interfaces (on-line searchable database, interactive map interface etc) will be charged at the current ADS day rates, plus university overheads on staff costs.
- All charges are subject to University of York overheads on staff costs and VAT at the standard rate;
- All depositors are required to sign the ADS Standard Deposit Licence [with the user licence this forms part of the rights management framework for the ADS];

- In cases where a depositor subsequently withdraws their deposit, the ADS reserves its right to charge a withdrawal fee to recover value added to data through the archival process, management and dissemination by the Service. Withdrawal fees may include all costs incurred by the Service up to withdrawal minus any deposit charges where these have been levied at deposit..

COMMENTARY

The ADS model is an excellent illustration of the levels of service required and potential costs that can be incurred for Resource or Community Data Collections in a national subject-based archive in a university.

The ADS Charging Model is activity based and can be mapped into the activity model in this study (although the grouping and presentation of activities differ slightly). ADS Management and Administration maps to our activities in Acquisition and customer accounts in Administration; ADS Ingest to Ingest; ADS Dissemination to Access and interface design in Common Services; ADS Storage and Refreshment maps to Archive Storage, Data Management and Preservation Planning.

Economic Adjustments (inflation/deflation, depreciation and costs of capital) from our model will be built into York University and ADS staff and refreshment figures as they are core elements of UK university accounting.

Service Adjustments from our model are either reflected directly in the ADS cost structure (e.g. salary levels, file format, volume) or controlled in the heading ADS Requirements and Other Cost Adjustments above (e.g. mode of deposit; metadata, documentation and IPR; de-accessioning costs; standard or custom interfaces).

Implicit in the charging policy are the business model for sustainability and the need for a sustained flow of future deposits and deposit charges (this effectively is similar to the business model for sustaining research within universities in TRAC or state pensions through national insurance contributions). This ensures that there is a continued existence of the service into the period where refreshment takes place. The refreshment charge itself includes modelling of the size of the archive and its future growth through a sustained flow of future deposits.

Costs for refreshment exclude our First Mover Innovation costs for major development of community standards, best practices and tools affecting preservation. An implicit assumption would need to be that major changes over timescales of 10 years or more in these costs would be funded from other sources.

Finally it should be noted that the ADS charging model is specific to the service at York and conditions within its discipline, so precise figures would not apply to another service or subject area. However its generic features suitably amended for different economic and service adjustments will be valuable elsewhere.

APPENDIX 2- CASE STUDY : THE UNIVERSITY OF CAMBRIDGE

BACKGROUND

The main contributions from the University of Cambridge have come from DSpace@Cambridge and the Department of Chemistry's Unilever Centre for Molecular Science Informatics. In addition we have collected information for the study from the following departments: Department of Social Anthropology, The University Library and Scott Polar Research Institute.

DSpace@Cambridge is a service run by Cambridge University Library and the University of Cambridge Computing Service. It is intended to be the core repository within a federation of repositories in the University and envisages taking primary responsibility for preservation.

The repository originated as a collaborative project, funded by the Cambridge-MIT Institute from 2003 to 2006, between Cambridge University Library, the University Computing Service, and MIT Libraries. The project's main objective was to establish the DSpace software platform as an institutional repository for the University of Cambridge, with a remit to consider a wide variety of materials and file formats, to explore the policy and management issues that would arise, and to pay particular attention to digital preservation functionality. To ensure that DSpace@Cambridge could be maintained as a sustainable service when CMI funding ended, the project also developed a business model for recurrent funding: this became operational when the repository service was inaugurated in August 2006.

DSpace@Cambridge is not currently applying any cost models for its preservation but various models were taken into account in the business planning process for the service. These resulted in the staff and hardware provision currently in place but the source of funding differs, i.e. the university meets the cost of the service because of its significance for the university. If other expenses are incurred, these will have to be covered elsewhere, for example, DSpace@Cambridge will be charging for some of the services supplied to the Cambridge University Colleges. The circumstances of their association with the University vary from college to college and all of them are financially and organisationally independent of the University as such, and therefore it has been decided that storage, preservation and dissemination of any content besides scholarly papers will be charged for. This is mainly done to cover the actual costs of managing deposited items; however there is an option of using surplus revenue from these charges to cover costs incurred by for example necessary preservation activities.

The business planning process produced a report, written in 2004, which remains useful from a number of viewpoints:

- In establishing the perceived benefits to deposit by the Cambridge academic community and the potential demand for services
- As a survey of the data types and formats which are held by potential depositors
- As a quantitative survey of the same community
- In underlining the importance of digital preservation as a key motivator in deposit
- In looking at the impact of different types of user on the cost of deposit

The report identified three broad user types, based on expertise in working with data, requirements for IT support, and availability of that support, e.g. in using the DSpace submission interface, reformatting data prior to submission, if required, and metadata creation. It anticipated that the cost of ingest would vary between user types since those without access to IT specialists and with few skills of their own would either require substantial technical help or training. The DSpace@Cambridge support structure anticipates that most users will need assistance in depositing data and other items. However, experience already suggests that this will vary considerably between communities. For example, within the Chemistry department, capture of research data in high volumes requires automated ingest procedures that greatly reduce the need for human intervention. However, this may generate additional set-up costs for software development in creating the necessary ingest tools.

Unlike many institutional repositories in the UK, DSpace@Cambridge accepts digital content in a wide range of formats. As long as the content has a scholarly or heritage focus the repository is in principle mandated by the university to accept it, and this has led to DSpace@Cambridge mainly containing images and research data with additional small collections of traditional research publications such as articles and theses. Part of the

DSpace@Cambridge service mandate includes preserving deposited content from the University of Cambridge. In order to achieve this, the team and the departments responsible for the service will adopt a lifecycle management approach, will create preservation plans and make use of available shared tools from the community. The JISC cost study will provide DSpace@Cambridge with potentially important tools for estimating costs related to the preservation of research data, a type of content that it is expected DSpace@Cambridge will receive more of in the future.

The Unilever Centre is aiming to establish a local facility to capture and disseminate crystallography data. This will happen in two steps; initial implementation is planned for the summer of 2008 where two Graduate students will work on developing an infrastructure for storage and dissemination, simultaneously they will be approaching the chemists to acquire permission to make the data openly available and then facilitate deposit of the data. Further they plan to employ a Graduate student in a 10% FTE position continuing the work of acquiring permissions and facilitating deposits to the repository, in addition they estimate that they will need 25% FTE of a Computer officer. Equipment costs and indirect costs will be covered thorough the budget of the Unilever Centre. For preservation they wish to use the DSpace@Cambridge service.

Analysis of institutional issues

Traditionally Cambridge University Library has, in addition to providing library services for the University, assumed a national and international role in supporting research. For this Cambridge University Library receives additional funding. While it is not envisaged that DSpace@Cambridge will preserve data for external institutions, much of the data deposited will have national and international significance. Examples from the existing collection include chemical informatics, data resulting from archaeological and anthropological fieldwork, and unique image collections. Undertaking the preservation of this data will have a significant cost. By facilitating preservation the University of Cambridge makes a contribution to the international research community, which one can assume that other institutions will be making as well.

Benefits of digital preservation

- Sharing and re-use of data, thereby deriving new data contributing to the creation of new knowledge. Examples of this can be data mining and re-engineering of original research leading to potentially new research fields. In Cambridge chemistry data is shared with Southampton University and Imperial College.
- By preserving links between published research and the processed data the research process in itself becomes transparent and thereby accountable. If we look at the chemistry example again, for any research grant commercial partners or research councils will have reporting requirements and making the original data available contributes to the reporting responsibilities. Another increasingly important aspect worth mentioning here is the possibilities open data gives to retest data against what has been published either for verification purposes or to avoid falsified research.
- Running preservation services centrally is cost effective and for many departments at the university the only available option. For example in the Department of Social Anthropology data management is a responsibility of each researcher, and they estimate that it would cost approximately £ 30,000 p.a. to hire supporting staff to manage data. The Scott Polar Research Institute made the same decision when deciding to use DSpace@Cambridge for the preservation and dissemination of digitised images created by the Freeze Frame project[4].
- A data preservation strategy is expected to form part of the university's overall information strategy. As part of fulfilling the preservation strategy the preservation capabilities of DSpace@Cambridge will be important, e.g. the University Streaming Media Service is planning to deposit all of its files in DSpace.

[4] www.spri.cam.ac.uk/resources/freezeframe/

Important issues to review when determining medium to long-term costs of data preservation in Cambridge:

Issues	Other information	How to resolve
Selection and/or appraisal processes	For DSpace@Cambridge it is currently the departments who decide what content to deposit based on recommendations from the DSpace@Cambridge team. In the longer term the scale of preservation activities required will determine possible selection processes for preservation activities.	Create policies on data deposit in collaboration with the research communities
Creation of adequate preservation metadata	Metadata generation is expensive and mainly a staff cost.	Development of automation routes and shared tools for metadata extraction across repositories extract metadata as well. Responsibilities of Support and Liaison Officer for DSpace include 30% metadata creation/standards support. Costs given below.
Supported formats	How many can we support? Do we store multiple versions?	Follow community standards and recommend preservation standards, for example DSpace@Cambridge has a

		web page describing formats issues[5].
Preservation methods	Develop preservation plans for each format based on community standards. Recommend open and supported formats.	This is largely a staff cost. The cost of a 0.5 FTE post for preservation planning is given below. A proportion of the system manager's costs (20%?) could be attributed to specific preservation tasks, e.g. migration on ingest to open formats, Word to ODF, identifying file formats, authenticity checks on existing content.
Authenticity and usability requirements	Do we aim for full usability of the content and new uses, or do we only wish to make an authentic file available for digital archaeologists? The repository also needs to consider to what degree it can support every reuse.	In general by supporting open formats and continued development of open standards. For DSpace@Cambridge a pragmatic approach will probably have to be adopted where the material is preserved to be accessible and usable, but without necessarily maintaining all properties of a file. To

[5] http://docs.repository.cam.ac.uk//index.php?option=com_content&task=view&id=41&Itemid=90

		achieve this it is important that the preservation activities are coordinated with each research community's requirements.
Sustainability	As mentioned above DSpace@Cambridge is currently supported by the University of Cambridge. Continued funding depends upon the achievements of the DSpace@Cambridge team, and the perceived value of the content stored versus expected costs to maintain it.	Several approaches are taken to ascertain that DSpace@Cambridge is sustainable in the long term; most importantly by offering departments managed space to deposit research data. In addition the repository will function as a core digital preservation service for research data by accepting preservation deposits from department repositories. Also, as mentioned above a charging policy for Colleges will be implemented.

Apart from the preservation support functions in DSpace and format guidelines, little work has yet been done to prepare the DSpace@Cambridge service for the preservation requirements to come. The intention is to implement a life cycle approach to preservation; evaluating each stage of the current process and where necessary change this to include appropriate preservation activities. An important contributor in this process will be the new digitisation and digital preservation specialist that the University Library is in the process of hiring (costs for this position is described below). Half of this person's time will be dedicated to facilitating preservation activities for both the University Library and university departments, ensuring that digitisation projects both inside and outside the University Library are coordinated and that preservation requirements are taken into account across the institution.

APPLICATION OF RESEARCH DATA PRESERVATION COST MODEL

Attribute	Who	Time
Initiation – Most research projects will be required to store the initial project documents and data. Main costs from a preservation perspective will be incurred when personnel are needed to advise and negotiate with the project on storage requirements, schedule, retention, formats and deposit methods.	Project staff in consultation with repository manager and system manager	Varies from 2-5 days per project for 2 FTEs at Grade 8, depending on factors such as whether format and project staff is new to DSpace.
Creation – Rights negotiation needs to facilitate usage through the entire content lifecycle although new usage modes will be difficult to predict. The generation of descriptive metadata and user documentation are resource intensive tasks and researchers will benefit from guidelines and expertise, e.g. in Cambridge metadata mapping has been essential for deposit of legacy data. The effort should decrease as depositors standardise on metadata.	University Research Services Division legal team. Attributable to indirect costs. DSpace@Cambridge Support and liaison officer (Grade 6) for metadata mapping.	One-off effort but may need to be revisited. 1-5 days at outset of projects. Grade 8 and above 5% of Grade 6 post. Recurrent.

Acquisition – Selection can require extensive staff resources but the DSpace team itself does not select content – this is carried out by submitters. Submission agreements are being standardised and should therefore be manageable. Outreach support is resource intensive but at the same time is important in order to ensure the early capture of data, related documentation, and software early in the lifecycle.	University RSD legal team advised on submission terms. Repository manager (Grade 8) & Support and liaison officer (Grade 6)	5 days for standard terms (one-off) + additional negotiations for specific content and usage (recurrent). >20% of ongoing effort
Transfer (e.g. on closure of service) – between repositories, either within university or external. Requires co-ordination and standardisation between repositories both on a technical and legal level. The costs related to these may depend on how well other parts of the process, e.g. rights or metadata generation, are managed.	Repository manager and System manager to create plan for transfer in case of closure, plus annual maintenance of plan.	Initially one-off plus maintenance on format or project basis
Ingest – the most cost-intensive stage but also most amenable to reduction through automation. Costs will depend on ingest methods chosen, and whether it is possible to develop effective automated methods, e.g. once metadata has been mapped for the Scott Polar images and a trial upload undertaken, the remaining image and	Most data is ingested in bulk by DSpace System Manager (Grade 8).	Recurrent 10% of ongoing effort.

metadata can be uploaded by Scott Polar project staff using secure FTP. Similar arrangements likely for output of Streaming Media Service. Integration with other campus systems, e.g. Sakai for VRE/VLE, content management and archival systems will potentially reduce effort and costs.	Integration with Content management system for images by DSpace System Manager and Developer -Development of integration carried out as part of JISC-funded project in CARET (VRE/VLE support team)	20% of 2 x Grade 8 posts for 3 months £300,000 project funding over 3 years.
Archive storage – Cost information on storage is readily available along with depreciation/replacement formulae. Technical developments may have a major impact on costs. See section on costs in section 1.4 Additional cost of off-site storage.		
Preservation planning - DSpace@Cambridge has a varied designated community and monitoring the various groups and their needs will be challenging. Raises questions about the long-term value of particular types of content and the cost of recreating it. In	To be undertaken by holder of new post of Digitisation and Digital Preservation Specialist (Grade 8) in conjunction with	Recurrent 0.5 FTE at Grade 8

Archaeology or Anthropology, for example, it may not be possible to repeat the research. • Tasks for the digital preservation planner include: Implementing a life-cycle management approach to digital materials, continuously assessing collections, their long-term value and formats, and making recommendations for action needed to ensure long-term usability. • Regular audits and preservation surveys of the Library's digital assets, evaluating their volume, formats, and state of risk. • research into preservation methodologies. • ensure that preservation actions are carried out on digital assets at risk of loss by Dspace team • formulate and publicise advice to data creators	DSpace team.	
First Mover Innovation: DSpace Federation; DPC, Research Data User Forum, JISC projects	Repository manager DSpace developer	Recurrent Up to 5% of 2 x Grade 8

COST DATA

Costs for storage and staff for DSpace@Cambridge

DSpace@Cambridge team (3 x FTE posts at Grade 8, point 45 on national scale, and one FTE post at Grade 6, point 37 on national scale) *

They cover the following activities:

- System management (installation, upgrades and testing, security, account administration and user authentication)
- Ingest, including bulk upload of files
- Support for metadata creation; mapping of existing metadata
- Training for users
- Development of new services, including features to support Cambridge users such as web services, Shibboleth-compliance
- Code fixes
- Database optimisation

Digitisation and digital preservation specialist (0.5 FTE post at Grade 8, point 45 on national scale).*

- Preservation strategy and planning
- Metadata management
- Technology watch and research on preservation methods
- Advice to data creators on formats and preservation issues
- Lifecycle management

**staff costs summarised on separate spreadsheet*

Storage solutions

For DSpace@Cambridge we have recently invested in the hardware described in the below, giving us around 150 TB of mirrored storage.

Description	Total (incl. VAT)
Sun Fire X4500 x64 Server	69,171.12
DELL/EMC CX3-20c FC4 SPE DAE4P-OS for CX3-20	107,122.70

DAE DAE4P FC4 for CX3-10/20/40/80	

Unilever Centre repository expected costs

Costs	**Initial phase**	**Production**
Staff costs	2 FTE Graduate students Grade 6, point 37 on national scale (3 months) £ 16,655.50	1 10 % FTE Graduate student Grade 6, point 37 on national scale £ 3331.1 1 25 % FTE Computer Officer Grade 8, point 45 on national scale, £ 12338.75
Other costs are covered on Unilever annual budget		

APPENDIX 3 - CASE STUDY: KING'S COLLEGE LONDON

BACKGROUND

This Case Study was prepared by Sheila Anderson (Director), Stephen Grace (Preservation Services and Projects Manager) and Gareth Knight (Digital Preservation Officer) at the Centre for e-Research. The Centre incorporates the AHDS Executive and its staff and projects.

The Case Study is based on the experience of the AHDS ingesting and preserving complex research data collections over an 11 year period, and on the more recent experience of the Centre for e-Research as it works to establish a research data management and preservation infrastructure for King's College London. It is taking as its starting point the strategic decision by King's College London to support research practice by developing a virtual research environment, including a research data repository to support the creation, management and long term preservation of College research data assets. The Centre for e-Research is applying the expertise gained managing the national Arts and Humanities Data Service to an institutional setting. This includes the work undertaken on the SHERPA DP and SHERPA DP2 projects, and AHDS contributions to the Life projects.

ANALYSIS OF INSTITUTIONAL ISSUES

The KCL research data infrastructure is taking advantage of the expertise and experience of the AHDS Executive, now integrated into the KCL Centre for e-Research (CeRch). To that end, King's is perhaps not facing such a daunting task to establish an infrastructure, policies and processes to support the life cycle of the research data produced by its faculty members.

The establishment of data management facilities is taking place under a wider project to develop an integrated virtual research environment to support the practices and processes of e-research. To start this process three key pieces of work are underway to scope:

1. the extent and quality of King's research data;

2. legacy systems that may need to be incorporated into the VRE, and systems that it will need to interact with;

3. the needs of the users of the VRE.

At the end of this process we will have produced: a data audit that identifies the range, extent and quality of data produced by King's academics, and the legal and ethical issues that need to be considered; an overall service oriented architecture within which the VRE will be placed, including a review of legacy systems that may need to integrated or upgraded as part of the process; an understanding of user requirements.

Some of the early issues that are arising from this work are as follows:

- What is the appropriate level of responsibility the College should take for King's research data beyond that necessary for supporting its own research? Should it incorporate an open access approach? What is its responsibility to the wider research community? How might the additional costs associated with an open access approach be met?
- What should be the appropriate sharing of responsibility for research data arising from collaborative projects that are across institutions? The increasingly national and international collaborative nature of research across all disciplines makes this a pressing question.
- How should the costs of managing and preserving King's research data be met, and does TRAC provide a potential solution?

This case study concentrated on an initial allocation of costs to the TRAC categories as a prelude to costing the management and preservation of research data as a Major Research Facility (MRF), or a Small Research Facility (SRF).

Application of research data preservation cost model

The Case Study is based on work that sought to apply TRAC methodology to the lifecycle activity model and resources template in order to demonstrate how institutions might allocate costs across the TRAC elements. The Case Study breaks down the lifecycle model elements into the TRAC categories of directly incurred, directly allocated, and indirect, following the application of TRAC at King's College London. The approach taken was to regard as directly incurred all those costs that could be validated against a project – those costs for which a project would be able to provide an audit trail to indicate costs directly spent on the data aspects of a project.

Directly allocated costs are based on the FEC costs of running the research data facility. In line with the TRAC Guidelines, all costs are included as directly allocated rather than indirect.

Activity Model	TRAC
Pre-Archive Phase	
Initiation	
project design	Directly Incurred
data management plan	Directly Incurred
Funding application	Directly Incurred
project implementation	Directly Incurred
Creation	
negotiate IPR/licensing/ethics	Directly Incurred
generate research data	Directly Incurred
generate descriptive metadata	Directly Incurred
generate user documentation	Directly Incurred
generate customised software	Directly Incurred
data management	Directly Incurred
create submission package for archive	Directly Incurred

Archive Phase	
Acquisition	
selection	Directly Allocated
negotiate submission agreement	Directly Incurred
outreach and depositor support	Directly Incurred (where directly related to a project) Directly Allocated (for general outreach and support)
Disposal	
transfer to another archive	Directly Allocated
Destroy	Directly Allocated
Ingest	
Receive submission	Directly Incurred
quality assurance	Directly Incurred
generate Information Package for Archive	Directly Incurred
generate administrative metadata	Directly Incurred
generate/upgrade descriptive metadata and user documentation	Directly Incurred
co-ordinate updates	Directly Incurred
reference linking	Directly Incurred

Archive Storage	
Receive data from ingest	Directly Incurred
manage storage hierarchy	Directly Allocated
Replace media	Directly Allocated
disaster recovery	Directly Allocated
error checking	Directly Allocated
Provide copies to access	Directly Incurred
Preservation Planning	
Monitor designated user community	Directly Allocated
Monitor technology	Directly Allocated
develop preservation strategies and standards	Directly Allocated
develop packaging designs and migration plans	Directly Allocated
develop and monitor SLAs for outsourced preservation	Directly Allocated
preservation action	Directly Allocated
generate preservation metadata	Directly Incurred
First Mover Innovation	
develop community data standards and best practice	Directly Allocated
share development of preservation systems	Directly Allocated

and tools	
Engage with vendors	Directly Allocated
Data Management	
administer database	Directly Allocated
Perform queries	Directly Allocated
generate report	Directly Allocated
receive database updates	Directly Allocated
Access	
search and ordering	Directly Allocated
generate information package for dissemination to user	Directly Incurred
deliver response	Directly Allocated
user support	Directly Allocated
new product generation	Directly Allocated
Support Services	
Administration	
general management	Directly Allocated
customer accounts	Directly Allocated
Administrative support	Directly Allocated
Common Services	
operating system services	Directly Allocated

Network services	Directly Allocated
Network security services	Directly Allocated
software licences and hardware maintenance	Directly Allocated
physical security	Directly Allocated
Logistics	Directly Allocated
Utilities	Directly Allocated
Supplies inventory	Directly Allocated
Estates	Directly Allocated
Consumables	Directly Allocated
Travel and Subsistence	Directly Incurred

Two spreadsheets, Archive Average Costs and Collection Costs, were developed for use within KCL. Archive Average Costs provided average costs for running the Research Archive Facility – all costs that would be allocated under the Directly Allocated category. It must be borne in mind that the allocations provided were a first attempt and should not be regarded as definitive.

Cost data

The case study considered three staff members to be essential for the establishment of a repository: an Archive Manager (salary £45,000) to co-ordinate activities; a half time System Administrator (FTE salary £24,000) to install and manage hardware and software; and a Collections Officer (salary £35,000) to develop and implement appropriate workflow and standards for the curation and preservation of research data. The costs allow a twelve month period to build and establish a data archive. Thereafter, the staffing affords the ingest of about 30 collections each year, with an assumption that 10 will be simple collections (images or texts created to standard formats and metadata), and 20 complex collections, comprising more complex formats (such as 3D visualisation materials) and multi media. Once this limit is reached, a new Collections Officer (CO) is required to cope with the additional work. The second CO may not be fully used on repository work at first, so the archive may spend more time on research, advocacy, data audits, retrospective additions, metadata enhancement. The cost of any new CO post will be smoothed over a 3-5 year period.

The hardware costs were based on those purchased by the AHDS in 2005 and comprised 15TB of storage, a tape library, and a dissemination server to allow end user access. The infrastructure was designed to cope with a maximum storage size and bandwidth; if either of these reach capacity a new or significantly altered arrangement is required. Archives should plan to renew their infrastructure every 3-5 years, and so build replacement costs into the annual planning.

A data audit exercise is needed at the outset of scoping a digital archive. This will identify collections and their relative importance to the institution and wider community. The archive can plan for ingest of collections from the data audit. By repeating the audit each year, an archive will also get a forward view of collections in the process of being created. Working in collaboration with Research Support Offices will give a three or four year perspective, and so help to inform infrastructure replacement.

A cost model is often created based on a set of pre-defined criterion or presumptions for the service adjustments. In this case the criteria indicate the expected operation of the digital archive, considering factors such as the collection policy of the digital archive, the time and effort required to curate and preserve each data type, and any activities necessary to tailor the research data for the Designated Community. However, there may be circumstances in which the digital archive is offered research data that is considered atypical, requiring the creation of a new costing model to finance additional work, and/or requires reconsideration of the organisation and technical infrastructure that is required to curate data. Specific events in the archive that may require action include:

- Capacity management: At its current rate of expansion, the digital archive will reach or exceed its storage capacity in the near future and must upgrade its infrastructure.

- Organisational IT environment: Develop in the IT infrastructure of the institution requires the digital archive to invest in making changes or upgrades to their underlying architecture.
- Development of preservation standards: The preservation practices of the digital archive may change as a result of new developments, e.g. creation of emulation tools.
- Capabilities of the repository architecture: The capabilities of the repository architecture may be assessed and a decision made that the digital archive upgrades to alternative repository software.
- Expectations of the Designated Community: The expectations of the Designated Community may change, in terms of the quality of information that is provided and the access method.
- Requirements of a funder: A funder may require that the research data is made available using particular methods or that access is limited to specific users. E.g. the publication of commercial data may have specific criteria that must be met.

The type of research data being submitted to the digital archive may also change over time as a result of newly developed requirements of the management board, archive manager, or Designated Community Research data may change, in terms of its:

- Size: the average size of a research collection may grow exponentially, due to the expanding scope of research projects (e.g. a project may digitise 100,000 images, as opposed to the 1000 – 10,000 images that earlier projects produced).
- Complexity: The type of research data being produced may increase in its complexity, consisting of several types of resource that are stored locally and/or remotely. To maintain the intended meaning the relational structure must be mapped.
- Type: New types of research data may be created, due to new development in the technical environment, improved processing and storage capabilities of modern hardware, and greater knowledge by resource creators. For example, an increasing number of collections deposited with the AHDS in recent years consist of moving image resources created by performance artists.

PROJECTING DATA PRESERVATION COSTS

The spreadsheets contained a first attempt at projecting costs over a 10 year period. The projections were relatively simplistic and did not account for the potential of automation of some processes - the complexity of such an exercise was beyond the remit of the case study.

The projections were based upon the concept of ‘spikes’ in cost: as the archive expands, so there is a need for more equipment and storage capacity to manage the increasing volume of data, hence costs increase over time. In a similar fashion, as the archive expands, and the number of collections ingested each year increases, so additional staff will be needed. The staff costs are based upon a collections officer dealing with the ingest and preservation of c.30 collections per annum, with 10 ‘simple’ collections (that is, simple text or images created to the preferred archival standards) and 20 complex collections, for example a multi media collection with images, text, video and sound and interlinking documentation (created to archival preferred standards); and with 20% of their time spent on general support tasks e.g. reviewing and updating licence agreements, standards activities etc.

The figures demonstrated that at a steady state, costs rise slowly over time, but if the archive expands, then ‘spikes’ in cost must be planned for and incorporated into financial planning.

APPENDIX 4- CASE STUDY : THE UNIVERSITY OF SOUTHAMPTON

BACKGROUND

Information to be considered as a case study for the cost model will be collected from the UK National Crystallography Service (NCS), The School of Chemistry and the National Oceanography Centre, Southampton (NOCS).

The National Crystallography Service

History

The National Crystallography Service has been operation since 1981, firstly at Queen Mary College London, then at the University of Wales College of Cardiff and since 1998 at the University of Southampton. For the whole period of its existence the NCS has always been housed within the respective Schools or Departments of Chemistry at these institutions and has been funded by a succession of research grants under both the rolling grant and responsive mode funding schemes of the EPSRC. The NCS provides an analytical service for UK chemists, based on state of the art experimental data collection facilities. This service includes the provision of raw data for those 'skilled in the art', who wish to work up a crystal structure themselves but don't have experimental facilities available to them or the provision of fully analysed crystal structures for chemists who do not have the necessary training or facilities to conduct these experiments.

The Service has operated a range of different instruments and has been operational for long enough to have seen, and be party to, a number of different techniques, processes, software programs, file formats and standards. Over the years this gives rise to very useful longitudinal data in the context of the acquisition of essential analytical chemistry data and so is an excellent case to inform this study. Whilst the technique of crystallography has always been reliant on computational power to develop its models, it is really since the mid-late 1980's that the entire process, from data collection to publication, has been digitally underpinned. This increase in power of the equipment and data analysis – deriving from improved instrumentation, increased computer power and much improved algorithms, has led to an explosion in the need for x-ray crystallography. Initially a technique that was used when absolutely necessary, it has moved to the routine structural technique of choice when a suitable crystalline sample is available.

Initially policies on the archiving and storage of this digital data were scant due to a lack of knowledge or understanding of working with this medium. More recently it has become clear that a service that is operating on behalf of others must have a policy for the archiving of the data it generates so that the data can be provided on request, some time after the original experiment(s) has been performed.

Data formats and standards

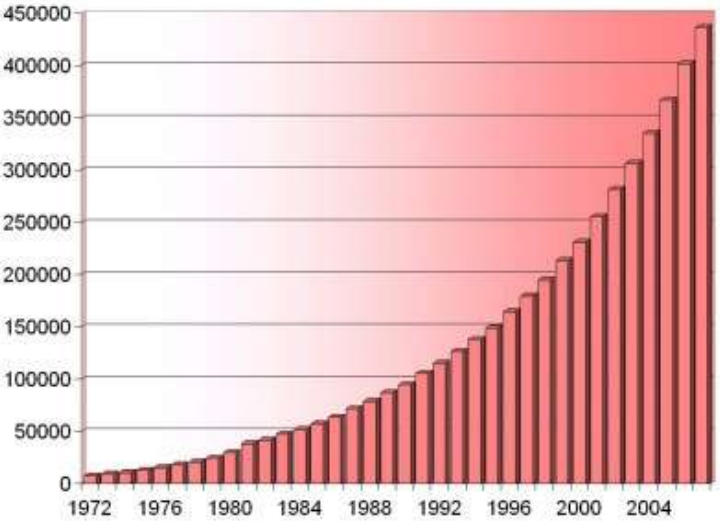
It is also recognised that there is almost unmeasurable value in some of this data, as the samples are prepared by highly skilled researchers in purpose built laboratories, either of which will most likely not be to hand after several years. Additionally there are patterns or events observed in some of this data that cannot be analysed using current techniques, but most likely will be resolvable in the future. There is also the matter of time related discoveries, where serendipitous measurements made today may be incorporated into entirely different studies at a later stage (e.g. cosmic rays are observed in some of these measurements as a side effect and considered to be a nuisance, however these observations could be of use in other fields in years to come).

As with many instrument based scientific techniques, crystallographers make a definite distinction between raw and analysed data. In this case raw data consists of about a gigabyte of ‘image’ files that are recorded in a proprietary binary format, depending on the manufacturer of the instrument being operated. This raw data is then processed into a condensed form (order of megabytes in ASCII text format) that can be read and worked up by a range of open source software than easily runs on an average desktop computer.

At the dawn of the digital era in the early 1990’s the crystallographic community saw the need, and invested heavily in, a common interchangeable standard for the exchange of the final result of a structure determination. This format, known as the Crystallographic Information Framework (CIF: http://www.iucr.org/iucr-top/cif/index.html) has been universally adopted by the crystallographic and chemistry communities for publication purposes and is maintained by the International Union of Crystallography. However, this format predominantly applies to the analysed result and there is no such universal acceptance of a standard format for the raw data, which remains a considerable problem.

Data Publication

Crystal structures are generally published in the journal article that describes the synthesis of the compound, although there are now an increasing amount of independent structural chemistry studies. The crystallographic data presented in these articles is collected by the Cambridge Crystallographic Data Centre (CCDC), which has been forming a collection of all the published crystal structure data (the Crystal Structure Database, CSD) for the last 40 years and is considered by the community as some form of subject repository. The rapid increase in crystallographic and computing hardware speed and capability over the last decade has resulted in a data deluge which is causing considerable problems for the management, archiving and publication of both raw and results data in the crystallographic field, as evidenced by the submission statistics to the CSD shown below.



As a result the NCS has developed a publication policy (http://www.ncs.chem.soton.ac.uk/pub_res.htm), which has arisen in conjunction with the development of a data repository for its outputs, that attempts to address these archiving, management and publication problems. This innovation is now being used as the basis for the JISC funded eCrystals Federation (http://wiki.ecrystals.chem.soton.ac.uk), which intends to build a network of such data repositories for the crystallographic community.

The National Oceanography Centre, Southampton

The centre is the UK's focus for oceanography and represents an unparalleled investment in marine and earth sciences and technology in the country. The centre opened in 1995 in a purpose-built, £50 million waterfront campus on the city's Empress Dock. It is a collaboration between the Natural Environment Research Council (NERC) and the University of Southampton. The Centre houses some 520 research scientists, lecturing support and seagoing staff as well over 700 undergraduate and postgraduate students.

The NOCS, as a partnership between the NERC and the University of Southampton, is guided by the NERC data policy, policy requirements of the University and the requirements of various collaborative projects.

The British Oceanographic data Centre (BODC) is NERC's designated centre for Marine Sciences. BODC is the primary repository for oceanographic data collected by researchers at the NOCS. Within NOCS, BODC activities are supported by the Scientific Data Management Group. In addition, NOCS is home to the British Ocean Sediment Core Research Facility (BOSCORF) and maintains part of the Discovery Collections, as well as ocean model data, some marine geophysics data and raw datasets.

NOCS is actively involved in the NERC DataGrid, a project of the UK's e-Science programme, involving atmospheric, oceanographic and geophysical sciences, with the aim of making its data more widely available to the scientific community. The NERC Data Discovery Service allows the searching of data resources held in the NERC DataGrid (NDG) catalogue. The catalogue makes data discovery easier as it connects data held in managed archives and other initiatives. It is populated with 'discovery' metadata (information about datasets) harvested on a regular basis from the NERC Data Centres and other providers in the UK and worldwide.

NOCS was a key player in the development of the University of Southampton's Research Repository with JISC project funding via the TARDis project. The British Atmospheric Data Centre (BADC) is the Natural Environment Research Council's (NERC) Designated Data Centre for the Atmospheric Sciences and its role is to assist UK atmospheric researchers to locate, access and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by NERC projects. Through the CLADDIER project, the University also explored the innovative use of a discovery interface linking publications in the research repository with the BADC Datasets. Funding for a part time person was provided by JISC.

To accommodate smaller scale local holdings, the RODIN database has been created and developed in-house at NOCS. RODIN (Repository of Oceanographic Data and Information) has been designed for compatibility and interfacing with the DataGrid and future e-science protocols. It allows data managers to archive and retrieve data files associated with metadata records. They can input, edit and copy metadata by means of the metadata editor and easily manage associated datasets.

While the emphasis in the past has been on external NERC repositories, it is early days in discussions of research preservation and these costs will need to be more specifically identified as part of future discussions of this local repository management. What is important at this time is that significant work has been done in identifying the research data created by NOCS researchers which needs to be preserved and, for example, attention can be turned at a later stage to data housed in paper archives which are perhaps at less immediate risk.

The University of Southampton Environment

The School of Chemistry at Southampton is one of the leading Chemistry research centres in the UK with an international reputation in several areas of Chemistry encompassing many aspects the wide spread of research in the Chemistry with research output in the top international chemistry journals. Funding for research is obtained from several of the UK research Councils (EPSRC, NERC, BBSRC), government, charities and industry. The research School consists of just under 30 academics, about 100 post-doctoral Research Fellows and 150 post-graduate students. The School is very forward looking and embraced the opportunity to investigate the applications of e-Science to Chemical research.

The University of Southampton is at the forefront of the development and implementation of the Institutional Repository model for the capture, storage and dissemination of its digital research outputs. Traditionally these outputs have been considered to be the journal articles, which for many represent the culmination of a piece of research work and the University of Southampton operates a highly successful Institutional Repository which mandates deposit of such articles by its research staff. We are now beginning to see research data, as opposed to these reports on schemes of research, becoming considered as equally important outputs of research in their own right and the University of Southampton has been involved in several highly innovative projects probing the issues around this area (CombeChem, eBank, R4L, eMalaria). The eCrystals project is very interested in this study, as during the scale up phase we will need to inform adopters of the costs inherent with best practice in preservation. Additionally academic research in the modern age is generating massive amounts of digital information with the accompanying requirement for mass storage and archiving facilities. To this end the University has convened a working group to consider the issues of quantity, diversity and cost of managing ALL its digital research output and hence this institution is extremely interested in the findings of this Research Data Preservation Costs study. This working group forms a discussion between the Library and Information Systems Services and academic staff.

ANALYSIS OF INSTITUTIONAL ISSUES

A primary issue from an institutional level is that it is highly important to improve the management of individuals research data by handling this issue more centrally. This is currently done very poorly in Chemistry, with data resident and isolated on computers attached to instruments and it is not unusual for many resort to hardcopy paper versions of data to ensure they have a version readily available (or at all) in the future. Moreover it is very difficult to publish a research study once an individual involved has left the research group or institution, which is often the case as a study cannot generally be published until it is finished, which often coincides with the conclusion of the short term employment of most research workers. In this context the ability to gain access to data at any point in time is highly desirable. The data does however have to be understandable, which can be a problem when accessing old (or even recent, but poorly curated) data sets.

It is however also important to consider the Department level in this landscape, in addition to the overall institutional level. It is an academics natural affiliation and an environment they understand and can often have an influence on, i.e. it is at this level where money can be raised and decisions surrounding ‘what is important’ can be made by the most appropriate people. Individual researchers are likely to feel alienated if archiving only occurs at an institutional level. At the individual school level this cost model remains highly pertinent when planning for preservation, however it is worthwhile noting that the School of Chemistry does not currently consider these issues, so advocacy and hands on help are required. A recommendation might be that a data store at the departmental level should be incorporated into the federation of institutional repositories and that the hierarchy of data stores reflects the detailed nature of the content and the changing nature of its importance over time.

An unexpected consequence of adequate data curation is the ability to expose the data with little extra work. This has been demonstrated in Southampton, by means of an institutional repository and the research work of the EPrints.org team, to lead to greater visibility of research output for individuals who have engaged with the data curation exercises.

In areas where regulation or patent issues are important then a second driving force comes into play, which is the ability of the archiving and preservation processes to underpin a whole study and provide a ‘provenance trail’. This trail acts as evidence for the thorough conduction of the experiment and avoids falsification of results. With increased public accountability and stories of fraud increasing, this is becoming a more important agenda. For the most part in Chemistry we do not have to deal with issues of confidentiality of personal data of the type that arise in studies involving human data.

The CombeChem and eBank projects, run in the School of Chemistry, were concerned with the capture, management and dissemination of chemistry research data and the following data preservation issues were raised and considered. These are issues that apply at the institutional level and frequently do not have clear answers, but considering them does give clarity to the expectations of the data providers and their view on potential consumers of their data.

Issues	**Comments**
What parts of the whole data set are actually worth storing?	This requires the intimate knowledge of the domain experts who generated the data
How much is generated?	Not just a matter of data volume, but also how many separate files, associated physical materials as well as data files and complex and highly specific metadata.

How diverse is this data?	The details of our expectations for preservation and curation of data vary with the degrees of refinement of the data. We expect that raw data will largely be used only by those generating it and within a few years of its initial production (i.e. a typical PhD project lifetime). However we are used to the fact that the reduced and analysed data used to support chemical theories is available for the long term (i.e. for ever or until supersede by more accurate data).
What are the cultural issues across different disciplines that might hinder the process of recording and acquiring data assigning access rights?	Case studies from numerous disciplines, such as those involved in this study will bring to light cultural differences.
What level of support should the institution be providing for this kind of service?	At one level, support is assumed in the continued existence of network services. It is the detailed support of hardware and software for the curated store that might be expected. The support for data preparation (and ingest) is much less likely to be supported at an institutional level as it may well be seen as too subject specific.
What is required to develop a policy to underpin institution-wide research data preservation and can a policy such as this work across different disciplines?	
How will this affect organisational structure in Library and Information Science?	
How will the curation effort be funded in the long and short term and who should be concerned about the costs?	
How is IPR handled?	Some data arising from an investigation may have significant attached financial value. Issues of prior disclosure may arise for data as well as ideas in papers.
Access rights and Embargoes?	For reasons similar to those outlined above, it may be necessary to restrict the access to the material at least for sometime. There is no concerned around who is accessing the data once it is made public and there are no systems in place to monitor this in detail.

APPLICATION OF RESEARCH DATA PRESERVATION COST MODEL

We have isolated one significant and representative activity undertaken within the School of Chemistry to highlight the production and curation of data. A major activity in Chemistry departments is the synthesis of a new compound, a fact that needs to be backed up and proven by structural studies. Here we have abstracted this process by considering both the typical synthetic task being undertaken in a synthetic organic chemistry group generating the samples (the project model), and the characterisation by the NCS (the sample model).

The Sample Model

In this case the centralised analytical characterisation service is considered. This is represented by the NCS, where a very expensively equipped laboratory is manned by dedicated personnel to provide a service, which once set up, operates according to a reasonably formulaic model. The service has a director, three research assistants and an administrator, is housed in the School of Chemistry and must bid for funding in three year cycles for both personnel and equipment. The NCS therefore uses numerous university facilities, such as estates human resources, finance, library and information systems services. The micro model is based on full economic costs for the staff, cost, maintenance and depreciation of equipment and consumables (including data storage). The currency for comparison is a crystal structure and the cost of operating the laboratory is divided by the average output to generate a unit cost.

The Project Model

In this model we are considering the new molecules being produced by the collaborative efforts of the Principal Investigator (PI), a Post-Doctoral Research Associate (PDRA) and a PhD student in a typical synthesis research project. This team would be occupying a laboratory module highly equipped to perform specialist experiments and using several of the School and University ‘centralised’ facilities. These centralised facilities would normally be analytical characterisation, estates support, human resources, finance, library and information systems services. A typical project would be active over a 3 year funding cycle, with 10 new compounds being generated per year. The cost for conducting this work is estimated from the full economic costs for these people in addition to consumables and equipment costs (which would include chemicals and other necessary reagents etc).

Application to the Activity Model

From a consideration of the preservation activities involved in the project and sample models the following points arise with respect to the components of the proposed activity model:

Activity Model	Sample Model	Project Model
Initiation	Data management plan is policy for NCS operation as a whole and defined at outset. NCS is a 'rolling' operation therefore no implementation. Funded by salary costs on grant.	A considerable amount of time is required in design and application by the PI. Data management plans are not usual in EPSRC grants, however this is changing as the attitude of the research councils change. (e.g. Wellcome Trust). That being said, the EPSRC does require a statement of how the research outputs will be disseminated and the dissemination pathways do have an impact on the curation that is implied or required. Implementation – recruitment and initial training can be a very lengthy exercise. Funded by salary costs on grant.
Creation	Negotiate IPR through agreement by user to adhere to publication policy at application stage. Considerable amount of data management undertaken and descriptive metadata generated due to implementation of repository [NB this is not common across the discipline and the current metadata schema is aimed at dissemination rather than preservation].	IPR issues depend on the project and the possible involvement of commercial and industrial partners. For most projects however these aspects are not usually discussed until latter stages. Researchers have little interest in generating preservation metadata or documentation, but could be persuaded via demonstration of the benefits of management and submission packages [NB metadata generation should only need to be done once so that the same information does not have to be generated again at a later stage]. Would have to be funded by indirect costs at institutional level.

Activity Model	Sample Model	Project Model
Acquisition	Selection requires experts to define 'what is worth keeping' and the structure of LIS doesn't currently provide for outreach or support. Currently the nature of what is kept replies on experience and past practice, it is rarely considered from scratch. The realisation of what can be kept and made searchable is changing attitudes to planning this activity. It is mostly still centred on a research group rather than a department or institution. Bibliographic data is stored on a University wide scale.	
Disposal & Transfer	Very costly to transfer from one archive to another (see cost data), despite having some standards and metadata schema. Potentially this can now be done automatically to some extent, but in the past this has been extremely human intensive work transferring between media etc. COSHH Book example – these safety books are required to be kept for several years but are hard to search and take up valuable space. The School has to destroy the books after the legal retention period, as it can't justify the storage costs. Would have to be funded by indirect costs at institutional/school level.	
Ingest	**Give**n the formulaic nature of a crystal structure study the QA, Archival Information Package, administrative metadata etc can be generated almost automatically and impact little on the depositor	This stage would have to be heavily assisted by preservation experts. Currently most researchers would have little idea or even interest when it comes to Archival Information Packages, QA or administrative metadata when it comes to preservation. They do realize the potential value of proper record keeping but it is not clear that the benefit outweighs the time needed. This is an area that needs considerable clarity and advocacy.

Activity Model	Sample Model	Project Model
Archive storage	Currently outsourced at reasonable cost, but this is simply for bit storage. Current hardware solutions being investigated will make this achievable by the institution itself (see section 1.7) with the additional advantage that bespoke preservation services can be designed on top of in-house hardware (although this is costly in terms of the input required by domain and software engineering experts). Would have to be funded by indirect costs at department or institution level.	
Preservation planning	Well understood in the crystallographic community and use of standards assists here. Media migration and avoiding file format obsolescence (of raw data) needs heavy financial investment (cost data in section 1.7).	Human labour intensive to monitor communities, technologies and standards. Some research areas do not lend themselves easily to developing or adopting standards. If possible it is desireable for generation of preservation metadata to be automatic.
First Mover Innovation	Very important and costly as LIS will have to restructure and spend a lot of time in this area. Vendors will be reluctant to engage unless a new business model evolves, as adoption of common interchangeable formats threatens their exclusivity stranglehold. Shared services, systems and tools are very important as they help assist in developing standards. Would have to be funded by indirect costs at institutional level.	
Data Management	A very labour intensive stage (see cost data), despite adoption of standards and falling costs of storage. Generally needs to be conducted by the characterisation service provider and therefore funded by direct costs in a research grant.	Doesn’t happen currently (as outlined above), but will be more costly than the sample model due to the variety of different types of data arising from the numerous different characterisation services used. Would require domain experts to conduct this stage.

Activity Model	Sample Model	Project Model
Access	Solved by development of a repository for results data, but for raw data this is managed locally and manually by the NCS.	Currently a big problem, due to a lack of standards or infrastructure and the diversity of the different experiment types. A considerable investment in infrastructure and advocacy would be required under indirect costs.
Administration & common services	A considerable amount of NCS staff time is devoted to administration, despite infrastructure developments.	This does not currently occur and would have to be implemented at the School or Institution level as part of indirect costs.

Cost data

Cost data for the generation of a crystal structure under the sample model and the creation of a new characterised molecule under the project model are provided below. Additionally some costs are also provided for the storage solutions used by the NCS. A full breakdown of the cost data is provided as supplementary information and further cost data are provided as part of the projection costs in section 1.7. Figures are given under the Full Economic Costing model, which includes both direct (salary) and indirect costs [NB These FEC costs will look large to anyone undertaking the project model, as much of the costs are not transparent to the academic researchers, where marginal costs as opposed to average costs are more obvious to them].

1) The annual costs for the sample model are:

Staffing	NCS (4 RA's)	£332000
	Department Service (Experimental Officer)	£90000
	Department Self Service (1 RA)	£83000
	Research Students (3PhD's)	£90000
Lab	Instrumentation capital cost (10%)	£45000

Maintenance		£2000
Repair (averaged over 10 years)		£10000
Raw data storage		£1200
Consumables		£4000
	Total	£657200

The laboratory collects approximately 2000 datasets per annum and therefore the cost per crystal structure (or sample in this model) is £328.60.

Notes

1. Research Associates in the Chemistry discipline are always employed at the Post Doctoral level.

2. Instrumentation cost is averaged over a 10-year period and is assumed to have depreciated to zero worth at the end of that period.

3. Maintenance costs refer to a regular (annual) amount that is routinely required (akin to 'having a service', whilst repair costs cover unforeseen breakdowns. The latter is averaged over the 10-year lifespan of the scientific instrument as these costs are initially low, but become considerably more significant with time.

2) The annual costs for the macro model are:

PI (10%)		£30000
RA		£100000
PhD Student		£40000
Consumables		£20000
Departmental Characterisation Services		£10000
	Total	£200000
Cost per molecule		£20000

The typical work and cost involved in the project study would involve: academic working up idea; application for funding; advertising & recruitment; student stipend (FEC); literature search & project plan; synthesis (Chemicals and Lab set up costs); characterisation (X-ray, Mass Spectrometry, NMR); property characterisation; publication. The costs associated with these activities are either direct staffing costs (eg for PI and PhD student etc) or indirect costs associated with an FTE for departmental or institutional central services.

3) The NCS costs for storage solutions are:

The averaged cost for preserving a results dataset in the NCS eCrystals repository is £2.15. This does not include the research work undertaken to produce a metadata schema, the software etc. The cost is calculated for the repository to be run and maintained in the laboratory, by a systems administrator who is also a crystallographer. Further assumptions are that the deposit process is performed by researchers (average 5 minutes each dataset) who have been provided with advocacy training. Additional costs incurred are the maintenance of metadata registries and assignment of persistent identifiers.

The averaged cost for outsourcing the archiving of a raw dataset with the Atlas Datastore is £1.48. This is based on an annual storage fee (TB/year) and time taken by a member of NCS staff to manage the deposit and retrieval processes. As part of a research project we are investigating the potential of an institutional solution (large object store) to store raw data, which would have a unit cost of approximately £1.60. This system would provide redundancy and self-healing, but requires bespoke development of extensive preservation services.

Different Models of Preservation and Associated Costs

A number of considerations of differing preservation models arise from this case study:

- Figures are provided for both outsourcing and an institutional solution for the storage of large raw data files. The cost for these two options is roughly the same, however the institutional solution on one hand offers complete flexibility for developing bespoke preservation services, whilst on the other it will require a skilled systems administrator to ensure reliable service provision (this would be funded from indirect costs at the institutional level).
- Crystallographers are used to depositing data with a centralised subject repository (CSD), but there is not total acceptance of the business model of licence charging and the fact that there is a requirement for the ownership of the data to be signed over to CCDC. The eCrystals Institutional Data Repository offers an alternative

solution, whilst also leaving open the option of the CSD automatically harvesting from it so that the ownership of the data is retained by the institution at the same time as the crystal structure being incorporated into the subject repository.

- There are some good organisational and cost arguments for using a departmental level repository that may be controlled and administered at the school level, whilst still being incorporated into the federation of institutional repositories. Additionally there could be some sharing of administration and management by both school and institution levels.
- Institutional Repositories will be seen by many as cumbersome. The Southampton experience is successful (and indeed necessary for the RAE) but has many issues. It is seen as difficult to use, and contains multiple entries for the same item of work. Many users have found it extremely valuable as a resource and it has increased exposure for those that use it well. So Institutional Repositories can and do exist, but the development of thin client tools will be necessary for researchers to embrace and adopt this approach. These clients need to work seamlessly with the usual word processing, web and database tools. A key gain is to demonstrate the value, at the laboratory level, of being able to easily get your data at any point in time and with minimal effort to deposit.

Units and Key Variables

Some general comments on the applicability of the key data categories and variables in the cost model to the case study presented here follow:

- Collections can be considered on a number of different levels in the chemistry/crystallography field: whilst a crystal structure is only relevant in the project model as a single type of characterisation, when assembled with other crystal structures in a community collection this single observation assumes much greater value and meaning. This has an impact on preservation aims as at the outset you may not know if this dataset is destined only for a project collection or whether at some point it will become part of a community collection.

- Preservation aims require a lot of interaction at both project and community levels to define – in the project model discussed here a researcher will not want to be spending their time doing this.
- Controlling future costs through the regulation of file formats is a very valuable exercise as it also has a number of implications for the ability to automatically disseminate content so that those outside the institution / project may reuse it in the future
- Timing and dependencies are vital to consider – the longitudinal study presented here shows the cost of having to take remedial action to repair ones data.
- This study presents costs arising from first mover innovations and community standards developments, which illustrate requirement for a significant initial outlay before new approaches to preservation can become routine (and hence cost effective).
- Staff costs are by far the greatest cost in this study – there is also a requirement for a lot of domain or community expert staff time in systems development and community interaction or engagement.
- The return in investing in automation is enormous and can reduce staff costs significantly.
- Effective community standards development requires critical mass and coordination in the community, which is a rarity and often taken on by a select few 'in their spare time'. De-facto standards, e.g. from popular software are common and may have some worth if interconversion software is available. Outreach, such as that by the DCC, is highly important to demonstrate the value of standards (drivers are generally publication and reuse, but effective preservation is a fortunate side effect). Some disciplines don't lend themselves easily to standards development or use.
- Most solutions scale well for the number of depositors, frequency, amount etc, but complexity, type and new formats are likely to be a problem as a considerable amount of research is conducted using bespoke or proprietary software).

- Migration is historically very costly, but likely to be vastly reduced as most data is now on spinning disks and more often in a standard or well described format. However periodic checking should be performed and this is costly if not automatable.

Projecting data preservation costs

The key costs for data preservation are going to be the development of infrastructure, advocacy and administration. These are generally very labour intensive exercises and therefore the real costs (direct and indirect) will be perceived as high.

The University of Southampton currently employs 1 FTE as administrator and developer of the institutional repository, however the initial development of the infrastructure and the advocacy required to get researchers to deposit ePrints would be estimated at 5 times this amount of input. The University has convened a working group to investigate and assess the development and administrative input required to cope with the preservation of ALL its digital research data. A conservative estimate at this very early stage is that this will require AT LEAST an order of magnitude more effort than the institutional repository and will require considerably more interaction with the active research community.

The School of Chemistry doesn't currently have any plans for data preservation services. The nature of chemistry research data is very diverse and some will not easily lend itself to preservation under a generic model. However the Repository for the Laboratory (R4L: http://r4l.eprints.org) that was run in the school developed a generic repository for data deposition. The cost of running preservation services in this way would require an initial outlay for development, deployment and advocacy but thereafter the costs would be similar to those incurred for the running of an institutional repository – i.e. the primary cost would be for an FTE to perform administrative duties.

Based on the historic data available from the NCS a longitudinal study for the preservation of data in the sample model is presented. This is broken down into raw and results data below and discussed alongside current innovations, which it is envisaged will provide a sound basis for preservation services in the next 10 years.

Raw data preservation per sample:

1989-1996 Magnetic tapes	£21.95
1997-2003 Compact Discs	£6.00
2003-Present Outsourcing	£1.48

The cost of *archiving* has roughly dropped by a quarter each time a new storage medium (and hence archival approach) has become widely available. It is important to note that this process is one of byte storage and very little, or no, preservation activity is performed – CD's were not periodically checked to ensure they were still readable and the outsourcing option merely ensures the retrieval of byte deposited. However, migration between media is often a problematic matter and is closely tied to the instrumentation – new instruments involve new software, formats and archival methods e.g. it was not possible or sensible to migrate ANY data from magnetic tapes to CD's, due to a new instrument, but the format was maintained for the CD to outsourcing migration and it was therefore deemed worthwhile to perform. The cost of the latter migration was approximately £0.75 per dataset, which was predominantly labour. There was a 7% loss of data in this process due to CD's being corrupt or unreadable. The cost of this data loss is doing all the chemistry and analysis again for any affected datasets.

Results data preservation per sample:

Results data preservation is quite different to raw data in that its volume is considerably more manageable.

1970-1990 Paper records	£30.00
1990-2000 Electronic copies on 3.25" floppy disks	£7.25
2000-present Electronic copies on computer disks	£2.15

The real cost of archiving results data roughly drops by a quarter as new methods and media become available. The cost of migrations is extremely high, with paper to electronic being about £25 per structure and a large amount of loss between spinning media and solid state. The cause of this high cost is the large amount of time required to perform the process.

Results can be regenerated if the raw data is preserved. However at modern day fEC, this would amount to between £50 and £400 (1-8 hours PDRA time) per structure. If raw data has not been preserved and results are lost then the cost of not preserving this data is enormous, as the compound generally cannot be resynthesised and therefore the amount that might be attributed here would be the cost of generating a molecule from the macro study (£20,000).

Looking forward:

Current innovations that involve the NCS and School of Chemistry / University of Southampton are providing examples of best practice in the preservation of this data and therefore give indications of future costs.

Raw data: The archiving of large datasets is becoming much cheaper as mass storage solutions become commonplace. Research between eCrystals and ePrints at the University of Southampton is developing preservation services for such a system. The initial outlay is great, but will provide solutions for whole communities or disciplines. Hardware costs equate to approximately £1.60 per dataset for a fully redundant (RAID type system) that is automatically self-healing. However, this hardware solution only addresses the issue of bit-rot and full preservation services have yet to be developed (1 research assistant for a year, ca £80K), although once this is achieved the maintenance and migration work would be low (5% FTE).

Results data: The eCrystals project developed schema and repository software for the preservation of crystal structure data in its first two phases (£350K), however it should be noted that this work was sufficiently generic not only to deploy for the whole crystallographic community, but potentially to act as a model for any experiment based science. The approximate year on year cost of running such a repository in the laboratory environment, with all the associated preservation administration and support would be £10 per crystal structure.

Appendices and Supplementary Information

Historic Study:

a) Raw data preservation costs over time

1989-1996

DEC VAX	Cost	£20000
	Service	£21000
50% for preservation purposes		£20500
600 datasets per annum		
Hardware annual cost		£2929
Magnetic tape media Unit cost £4 (10 datasets per unit)		£240
30% technician hours to write to tape & manage storage / retrieval		£10000
Annual raw data preservation cost		£13169
Cost /dataset/year		£21.95

Not possible to migrate across media. Tape reader kept going for about 2 years, software rapidly obsolete: 100% loss. Tapes need to be reread every year to maintain them, considerable space required for storage.

1997-2003

Linux PC (2 in time period)	£4000
1000 datasets per annum	
Hardware annual cost	£667
CD's	£85
15% technician hours to write to CD & manage storage / retrieval	£5250
Annual raw data preservation cost	£6002
Cost /dataset/year	£6.00

Migration CD to USB/Atlas £2500+£1000	(7% Loss)	£3500

Data loss dependent on method (press vs burn) and speed of writing – generally accepted that data will be lost after 5 years if written at highest speeds. Would require re-reading periodically.

2003-Present

Atlas DS fee (per TB/year)	£1200
5% technician hours to write, manage storage / retrieval	£1750
Annual raw data preservation cost	£2950
Cost /dataset/year	£1.48

b) Results data preservation costs over time

1970-1990 Paper records			
	Printing		£400
	Space		£300
	Management time		£8300
		Total	£9000
		300 datasets pa	£30.00
1990-2000 Electronic copies on 3.25" floppy disks			
	Floppy disk cost		£200
	Writing & Management time		£4150
		Total	£4350
		600 datasets pa	£7.25
2000-present Electronic copies on computer disks, mirrored			
	Disk cost		£500
	Writing & Management time		£2075
		Total	£2575
		1200 datasets pa	£2.15

'Sample' study:

The following is based on current fEC costs and an annual estimate for instrumentation is taken to be 10% of the purchase cost over 10 years (+upgrades).

Staffing	NCS (4 PDRA's)		£332000
	Department Service (Experimental Officer)		£90000
	Department Self Service (1 PDRA)		£83000
	Research Students (3PhD's)		£90000
Lab	Instrumentation capital cost (10%)	£45000	
	Maintenance		£2000
	Repair (averaged over 10 years)		£1000

Raw data storage	£1200	
Consumables		£4000
	Total	£657200

The laboratory collects approximately 2000 datasets per annum and therefore the cost per crystal structure is £328.60.

‘Project’ Study

Costs per year (FEC) based on a typical PhD student study in a well-supported laboratory and research group.

PI (10%)	£30k
PDRA	£100k
PhD Student	£40k
Consumables	£20k
Services (analytical characterisation)	£10k
Total	£200k p.a

Infrastructure development costs

Initial investment:

Grant to develop eCrystals (first 2 phases)	£350000

Annual operation and deposit time for a maintained laboratory based repository:

Time for self deposit of results data = 5 minutes each structure

For 2000 structures = 167 hours		£7614
Raw data deposit cost (from above)		£2950
Hardware cost (server)		£1000
Sys Admin (5% time)		£4150
Persistent identifier cost		£250
Maintenance of metadata registries etc	(5% time)	£4150
	Total	£20114

2000 structures per year: unit cost £10.06

APPENDIX 5: REVIEW OF LIFE COST MODELS

Introduction

This review analyses components of the LIFE 1.0 (Mcleod et al 2006) and LIFE 1.1 (Wheatley et al 2007) costing models against each other and against the Open Archival Information Systems (OAIS) Reference Model (CCSDS 2002). The comparison with OAIS was undertaken because as an ISO standard it supplies a generally-accepted high-level view of the key functional components associated with long-term digital asset management and a neutral mechanism to compare different systems and terminology. OAIS is also accompanied by well-defined definitions for each component that facilitate comparison with other models and implementations.

LIFE 1.0 was the final project report from the original LIFE project. LIFE 1.1 represents an interim output for discussion from the LIFE 2 project which concludes in August 2008. Further work is planned in LIFE 2 and a new version of the cost model will be published at the end of the project. Proposed next steps include: (i) performing detailed mappings to related standards or work e.g. OAIS; (ii) Application to LIFE 2 case studies; (iii) continuing to gather feedback (Wheatley et al 2007,14).

The OAIS Reference Model is a conceptual framework that provides an implementation-neutral description of the high-level functions of an archival system. It also provides a high-level information model describing archived information and its associated descriptive information as it is submitted to, stored, and disseminated by the archive. The standard itself is relatively long and detailed but there are good short overviews to the standard available for those who are unfamiliar with it and want a short introduction to its key principles (Lavoie 2004). It is currently undergoing its five-year review as an ISO standard and some minor changes are expected but have not yet been published.

Strengths of the LIFE Model

An analysis of the strengths and weaknesses of the LIFE models from the perspective of the requirements of the Research Data Preservation Costs study is provided below:

(1) The LIFE projects have adopted a lifecycle approach to digital preservation costs. This has a long pedigree in cost modelling in other sectors; has been applied to costing in traditional library collection management; and has been advocated as an approach to digital collection management and costing digital preservation (Watson 2005).

(2) Preservation is normally a component function for most organisations: a means of achieving other key objectives such as current and future access or re-use of data and information rather than an end in itself. Preservation costs therefore can be heavily influenced by or be difficult to separate out completely from, other functional costs. Use of a lifecycle model helps to address these challenges.

(3) LIFE 1.1 has added a stage for Creation or Purchase. This is a helpful development and could allow a model to reflect dependencies and implications for costs between pre-archive and post-archive phases (note this stage in LIFE 1.1 is in development and has no elements assigned as yet). Arguably the purchase component could be seen as distinct and separate from creation and absorbed into ordering under acquisition in the LIFE model.

(4) LIFE draws heavily on the experience of the British Library which has preservation as a core function. This experience is supplemented by experience from other institutions including the UCL university library, which although not heavily involved in preservation provides a different Higher Education perspective.

(5) LIFE allows for more pro-active collection development processes than the OAIS model. In particular it has an Acquisition stage with a selection element which will be more appropriate for some research archives with a degree of choice over acceptance of data collections offered for ingest.

Weaknesses of the LIFE Model

(1) HEIs need to consider Full Economic Costs (FEC) and the LIFE model focuses mainly on lifecycle episodes rather than associated ongoing and support infrastructure costs. Treatment of some direct and most indirect costs is very weak or

excluded in LIFE 1.0 and still poorly developed at this stage in LIFE 1.1. Typical indirect costs such as estates and utilities are not included. The new functions of management and administration may address some indirect support costs such as finance and human resource functions and administrative support (note these stages in LIFE 1.1 are in development and have no elements assigned as yet). General information technology and service functions and costs are also weakly developed at present but a new systems infrastructure stage has been introduced into LIFE 1.1. Economic Adjustments for inflation and discounting (i.e. depreciation) are now included in Life 1.1.

(2) Legal Deposit libraries have a set remit and mandate for preservation in perpetuity and the model tends towards an implicit assumption of a uniform preservation aim or outcome and therefore cost. Research data collections will be much more variable in terms of preservation outcomes and timescale and therefore cost could be highly variable. This will need to be factored into any new models for research data preservation costs.

(3) LIFE focuses on the library sector and library materials. Many high-level stages and elements may be identical but careful review and adjustments are needed for application to other areas and data types.

(4) Metadata has been given a separate stage to itself in LIFE 1.1 to emphasise the importance of metadata. However this draws metadata out of context from different stages and costs in the lifecycle. Compared to the OAIS model LIFE is also narrower as it does not have both metadata and documentation. For research data often both metadata and documentation (research designs, instrumentation, lab books etc) will be essential to preservation and use.

(5) The model currently treats costs in each stage as independent elements. In practice in many cases choices made at one point in the lifecycle could ripple across to other stages and costs. Greater sensitivity may need to be built in to the model to these choices, linkage between different elements, and modelling how costs are affected.

Comparison of the LIFE Costing Models and OAIS

The LIFE models consist of “stages” representing high-level processes with the lifecycle. A number of distinct “elements” are grouped under each stage and represent specific high-level processes within it. Definitions are provided for stages and elements in the project documentation. Within each element definition sub-elements which may be specific components are suggested for guidance only. Changes between LIFE 1.0 and 1.1 can be identified in the table below. They are discussed in detail by the project team (Wheatley et al 2007, 13-14) and are not repeated here.

The OAIS reference model consists of six high-level “functional entities” with a number of distinct functions grouped under each. Definitions are provided in the standard for functional entities, specific functions and a range of associated issues. As concepts they are normally exactly or very closely comparable to stages and elements within LIFE.

Occasionally an element may appear in different stages of the lifecycle between LIFE 1.0 and LIFE 1.1 or have an exact/close match to, or be subsumed within, a function in different functional entities in the OAIS model. In such cases the match will be shown in grey type.

Table. The Life 1.0 and Life 1.1 cost models compared against each other and the OAIS Reference Model.

Key to Table

xxxx LIFE Stage or OAIS Functional entity

xxxx LIFE element or OAIS function

xxxx equivalent located elsewhere in LIFE or OAIS model

LIFE 1.0	LIFE 1.1	OAIS	Comments
	Creation or Purchase		– new stage added in LIFE1.1 optional and separated from archiving lifecycle
Acquisition	Acquisition		
selection	selection		
IPR	IPR and licensing	negotiate submission agreement	
licensing			
ordering and invoicing	ordering and invoicing	customer service	
obtaining	obtaining		
check-in	check-in	receive submission	See OAIS supplement (CCSDS 2004)
	submission agreement	negotiate submission agreement	See OAIS supplement (CCSDS 2004)
Ingest	Ingest	Ingest	
quality assurance	quality assurance	quality assurance	

LIFE 1.0	LIFE 1.1	OAIS	Comments
deposit	deposit	generate AIP	
holdings update	holdings update	co-ordinate updates	
	reference linking		This has moved stage in LIFE 1.1
descriptive metadata	metadata creation	generate descriptive information	
Metadata	Metadata creation		Given separate stage in LIFE to emphasis its importance. Not directly comparable to OAIS. In OAIS metadata is mostly subsumed in ingest activities
characterisation			
descriptive		generate descriptive information	
administrative			

LIFE 1.0	LIFE 1.1	OAIS	Comments
	re-use existing metadata		
	metadata creation		
	metadata extraction		
Storage	Bit-stream Preservation	Archive Storage	
	repository administration	receive data from ingest provide copies to access	
bit-stream storage	storage provision	manage storage hierarchy	
	refreshment	replace media	
	backup	disaster recovery	
	inspection	error checking	
Preservation	Content Preservation	Preservation Planning	
technology watch	preservation watch	monitor technology	
	preservation planning	develop preservation strategies and standards	

LIFE 1.0	LIFE 1.1	OAIS	Comments
		develop packaging designs and migration plans	
preservation action	preservation action	archival information update	
	re-ingest	generate AIP	
preservation tool cost			
preservation metadata			
quality assurance		audit submission	
	preservation watch	monitor designated community	
Access	Access	Access	
access mechanism	access provision	co-ordinate access activities generate DIP deliver response	
user support	user support	co-ordinate access activities	

LIFE 1.0	LIFE 1.1	OAIS	Comments
	access control	security services	
reference linking	reference linking		
	Management and Administration	Administration	new stage added in LIFE1.1. Part of non-lifecycle costs. In development – a definition for this LIFE stage needed prior to mapping elements in detail to OAIS
	management		
	administration		
		negotiate submission agreement	
		manage system configuration	
		archival information update	
	access control	physical access control	

LIFE 1.0	LIFE 1.1	OAIS	Comments
		establish standards and policies	
		audit submission	
	access mechanism	activate request	
	user support	customer service	
	Systems/Infrastructure		new stage added in LIFE1.1. Part of non-lifecycle costs. In development – a definition for this LIFE stage needed prior to mapping elements in detail to OAIS
	repository software		
	Economic Adjustments		new stage added in LIFE1.1. Part of non-lifecycle costs. Only applicable to cost models so not in OAIS
	inflation		

LIFE 1.0	LIFE 1.1	OAIS	Comments
	discounting		i.e. depreciation
	repository administration	Data Management	
		administer database	
		perform queries	
		generate report	
		receive database updates	
		Common Services	
		operating system services	
		network services	
		security services	

Conclusions

We found the lifecycle model in LIFE and the treatment of preservation within this to be very valuable input for the development of our research data cost model for HEIs. It is derived from a library context and needs adaptation though for this new purpose. As additions, we are looking at changes required for research data including costing a range of preservation aims and retention periods, closer alignment with HEI's use of TRAC for Full Economic

Costs, and greater flexibility and sensitivity in the model to variables and choices which will influence the costs of preservation.

In comparing the LIFE and OAIS models we found:

- LIFE has incorporated an optional pre-archive stage for creation or purchase which we believe is a helpful addition;
- LIFE has a distinct stage for acquisition which we would also support. Elements of this are only partially represented in the OAIS model;
- LIFE has a separate stage for Metadata – we prefer the OAIS emphasis on both documentation and metadata (descriptive information) and leaving metadata in situ within the lifecycle;
- There is broadly a close match between the Ingest, Archive Storage, Preservation Planning, and Access functions in LIFE and OAIS;
- The Data Management, Administration and Common Services functions are more developed in OAIS than LIFE;
- LIFE adds a number of elements under Economic Adjustments appropriate to a cost model.

Appendix 6: Review of NASA Cost Model

Introduction

This review analyses components of the NASA Cost Estimation Tool (Booth et al 2006, Fontaine et al 2007, Hunolt 2006a, Hunolt 2006b, Hunolt et al 2006) against the Open Archival Information Systems (OAIS) Reference Model (CCSDS 2002). The comparison with OAIS was undertaken because as an ISO standard it supplies a generally-accepted high-level view of the key functional components associated with long-term digital asset management and a neutral mechanism to compare different systems and terminology. OAIS is also accompanied by well-defined definitions for each component that facilitate comparison with other models and implementations.

The NASA Cost Estimation Tool (CET) was developed for estimating the lifecycle costs from implementation through a time-limited period of operations (currently up to 12 years) for NASA science data activities and projects. It does not currently address long-term archiving requirements but these are under consideration as potential extensions to the tool and its underlying data activity reference model. The model has been developed from comparison of costs in 29 operational data centres for NASA earth and space science data activities and related international partners and initiatives. These centres vary in size and staffing ranging from 2FTEs to 66FTEs (Fontaine et al 2007).

The OAIS Reference Model is a conceptual framework that provides an implementation-neutral description of the high-level functions of an archival system. It also provides a high-level information model describing archived information and its associated descriptive information as it is submitted to, stored, and disseminated by the archive. The standard itself is relatively long and detailed but there are good short overviews to the standard available for those who are unfamiliar with it and want a short introduction to its key principles (Lavoie 2004). It is currently undergoing its five-year review as an ISO standard and some minor changes are expected but have not yet been published.

Strengths of the NASA CET Model

An analysis of the strengths and weaknesses of the NASA CET model from the perspective of the requirements of the Research Data Preservation Costs study is provided below:

(1) The NASA CET has adopted a lifecycle approach to costs and it can be mapped relatively easily into the LIFE and OAIS models.

(2) The model is derived from experience with 29 operational data centres from space and earth observation. This gives a strong empirical underpinning to the cost model and a strong degree of confidence in the statistical validity of its cost data for NASA activities. It has a shared focus on research data with our study and is particularly useful for the early stages of our cost model even if it currently excludes costs for long-term preservation.

(3) The NASA CET reference model has particularly good description of functions with definitions for Information Technology and systems costs associated with projects. Several of these relate to "Creation" phase activities excluded from the OAIS reference model.

(4) The CET has a set of 94 metadata fields (descriptors) with accompanying definitions which are used to describe specific functions. A number of these have menu options to capture key cost variables e.g. level of service or automation levels. There are sensitivity adjustments and linkages within the CET which inter-link components of the model and allow "what if" scenarios and ripple effects from changes in different elements to be modelled.

(5) The model distinguishes between "operational" and "support" functions. Support activities are essentially overhead that is distributed across one or more of the operating functions. Support activities may suggest the contours of a general institutional infrastructure that underpins a network of preservation activities.

(6) The reference model is supported by a prototype suite of Excel-based tools and a database of comparable costs from 29 projects/activities for estimating lifecycle costs. These may provide a model of how to extend and support future implementation of our model in UK HEIs through development/adaptation of appropriate software tools and comparators.

(7) The cost estimation process currently has an overall average absolute error of 22.9%. It could be argued that knowing how accurate your predictions are can be a strength (even if it is a broad margin), gained in this case from lots of longitudinal data. Comparative data and estimating techniques have been refined over time and it is believed the cost-estimation performance of the CET may now be as good as it can be (Fontaine et al 2007).

Weaknesses of the NASA CET Model

(1) The CET model currently has no provision for long-term digital preservation costs or functions although there is ongoing discussion on how these could be added in future (Fontaine et al 2007).

(2) It is based primarily on NASA projects and costs in space and earth observation research. It will not always extend or fully cover requirements for research data in other disciplines.

Comparison of the NASA CET Costing Model and OAIS

The NASA CET reference model consists of two parts representing operating and support functions. These contain a set of specified functions representing high-level processes within the lifecycle of activities/projects. Definitions and examples of sub-processes are provided for functions in the project documentation.

The OAIS reference model consists of six high-level “functional entities” with a number of distinct functions grouped under each. Definitions are provided in the standard for functional entities, specific functions and a range of associated issues.

A mapping between OAIS functional entities and functions and functions in NASA CET reference model has been published by the NASA team (Fontaine et al 2007). This is mainly a high-level mapping of CET functions to OIS functional entities. An independent mapping has been undertaken for this study mapping at a more detailed level to OAIS functions using the CET function definitions and sub-processes mentioned within these. It is broadly comparable with the NASA team mapping apart from this additional level of detail which helps identify differences and potential areas for future development. Occasionally there are also some differences in interpretation of “best match” between the function definitions.

Occasionally a function in the NASA CET model will have a partial match or an exact/close match/ be subsumed within, a function in different functional entities in the OAIS model, or vice versa. In such cases the match will be shown in grey type.

Table.The NASA CET reference model compared against the OAIS Reference Model.

Key to Table

xxxx NASA CET Function or OAIS Functional entity

xxxx OAIS function or sub-processes in NASA CET function definition

xxxx equivalent located elsewhere in NASA CET or OAIS model

Mapping against NASA CET Operational Functions		
NASA CET	**OAIS**	**Comments**
Technical Co-ordination		Probably not in OAIS directly but within supplementary guidance on OAIS archive –producer interface (CCSDS 2004). CET definition “co-ordination on programme level...on data management, data stewardship, standards and best practices, interfaces, common metrics and interoperability as needed to support.”
Ingest	Ingest	
receiving	receive submission	
reading	quality assurance	
quality checking	quality assurance	

cataloguing of data/metadata	generate descriptive information co-ordinate updates	
	generate AIP	
Product Generation		Not in OAIS? CET definition “initial generation and reprocessing with quality checking of new data or products from data or products previously ingested or generated”
Archive data stewardship	Archive Storage Preservation Planning Data Management	Archiving is much more extensive within the OAIS model and maps onto 3 OAIS functional entities.
	Archive Storage	
insert into archive	receive data from ingest	
	manage storage hierarchy	
	replace media	
	error checking	
	disaster recovery	
	provide copies to access	

	Preservation Planning	
	monitor designated community	
	monitor technology	
	develop preservation strategies and standards	
	develop packaging designs and migration plans	
Search and Order	Access	Search and Order combined with Access and Distribution in CET database
access to catalogue	co-ordinate access activities	
search and order for user capability	activate request customer service	
Receiving requests	activate request	
Access and Distribution		Search and Order combined with Access and Distribution in CET database
retrieval for requests	generate DIP	
subsetting /format conversion / packaging /reprojection	generate DIP	

providing to end user	deliver response	
User Support		User support is not well developed in OAIS: it is subsumed in OAIS function co-ordinate access activities
response to queries	co-ordinate access activities	
taking orders	co-ordinate access activities	
help desk	co-ordinate access activities	

Mapping against NASA CET Support Functions

NASA CET	OAIS	Comments
Implementation		Not in OAIS?: a pre-archive creator activity? CET definition “development of data and information system capabilities including design and implementation of the data system (hardware and system software) and applications software.”
Sustaining Engineering		Not in OAIS?: CET definition “maintenance and enhancement of custom applications software...”

Engineering Support	Common Services	Matched but some significant variations in level of detail and coverage.
system engineering	operating system services	
test engineering	operating system services	
configuration management	operating system services	
COTS procurement/upgrades		
system administration	operating system services	
database administration	administer database	
network engineering	network services	
network security	security services	
Management	Administration	Matched but some significant variations in level of detail and coverage.
management /administration at data activity level		
management of functional areas		
administrative support		

	negotiate submission agreement	
system administration	manage system configuration	
	archival information update	
	physical access control	
	establish standards and policies	
	audit submission	
	activate request	
	customer service	
	Data Management	
	administer database	
	perform queries	
	generate report	
	receive database updates	
Facilities/Infrastructure		Staff costs merged with engineering support in CET database.
resource planning		
logistics		

supplies inventory/acquisition		
facility management		
maintenance of system and site security		
Non-staff costs items such as supplies, facility lease, utility and other overhead costs, hardware maintenance, COTS licences, etc.		

Conclusions

We found the NASA CET model and the treatment of cost and cost estimation in this to be a particularly valuable input for the development of our research data cost model for HEIs. Its strengths were in being derived from the research process and data handling. It used a lifecycle approach derived from experience with research data applications and Excel-based tools that were widely useable by non-specialists. Its weaknesses are that it is currently designed for activities of no more than 12 years duration and is less well-developed for assessing long-term preservation costs and functions (although it is easily extensible for this); Also it is based largely around 'big science' so it needs some adaptations for small-scale research data. The study had found the model to be very flexible and modular (each function can be costed independently) and very strong on technical development and identifying different cost variables and sensitivities. The inclusion of documentation, user support, and technical co-ordination for data generators and depositors were also very relevant for research data. The study was in addition looking at factoring in long-term preservation functions and varying preservation aims, collection/project scales and requirements.

In comparing the NASA CET and OAIS models we found:

- NASA CET includes functions for Implementation, Sustaining Engineering, Technical Co-ordination, and Product Generation which are largely absent from OAIS;
- NSA CET has a more developed view of User Support than is present is OAIS;
- Archive and preservation activities are more developed in OAIS than NASA CET;
- There are relatively close matches between Ingest and Access/Distribution in both models;
- Engineering Support and Management/Administration match but have significant variations in level of detail and coverage;
- Data Management is treated as a separate function in OAIS but not in NASA CET.

APPENDIX 7: EXTRACTS FROM APPENDIX D IN LONG-LIVED DIGITAL DATA COLLECTIONS: ENABLING RESEARCH AND EDUCATION IN THE 21ST CENTURY

DIGITAL DATA COLLECTIONS BY CATEGORIES

Introduction

Digital data collections vary greatly in size, scope, usage, planned duration, and other dimensions. We distinguish between three functional categories of data collections:

(1) research database collections, which are specific to a single investigator or research project;

(2) resource or community database collections, which are intermediate in duration, standardization, and community of users; and

(3) reference collections, which are managed for long-term use by many users.

The following sections provide descriptions and examples of each of these types of digital data collections.

It should be noted that there are not always clear distinctions between these categories: data collections for large research projects overlap with community database collections, and many community data collections transition to become reference data collections. These categories are based on functional attributes of the collection rather than location or size of the data set, and some data centers support all three kinds of collections.

Research Database Collections

Description

Research database collections are the products of one or a few focused research projects. The collections may vary greatly in size, but are intended to serve a specific group, often limited to immediate participants. These collections have relatively small budgets and may be supported directly or indirectly, often through the research grants supporting the project that they serve. Funding is assured for only a short period of time. They typically contain data that is subject to limited processing or curation, and may or may not conform to community standards (e.g. standards for file formats, metadata structure and content, access policies, etc.). Often, applicable standards may be limited or rudimentary as the data types may be novel and the size of the user community may be small. The collection may not be intended to persist beyond the end of the project. Some research collections are accessible to the public through the Web, but many are not, and many of the Web links to research collections are ephemeral.

RESOURCE OR COMMUNITY DATA COLLECTIONS

Description

Resource or community data collections serve a specific science and engineering community. They are typically between research and reference data collections in size, scale, funding, community of users, and duration. They typically conform to community standards, where such standards exist. Often these digital collections can play key roles in bringing communities together to develop appropriate standards where a need exists. In many cases community database collections migrate to reference collections. In some fields, such as biology, resource data collections are often separate, directly funded projects. In other areas, such as the earth and environmental sciences, resource database collections are often managed under the umbrella of a data center that also supports research and reference databases.

REFERENCE COLLECTIONS

Description

Reference collections are intended to serve large segments of the general scientific and education community. Conformance to robust and comprehensive standards is essential to provide the diverse user access and impact that are the mission of these collections. Adoption of standards by reference collections often ‘sets the bar’ for a large segment of the community, effectively creating a ‘universal’ standard. Budgets are often large, reflecting the scope of the collection and breadth of impact, and are typically provided by long term, direct support from one or more funding sources.

Appendix 8: Interview Questionnaire

JISC Research Data Preservation Costs Study

The study contract was awarded to Charles Beagrie Limited and will report at the end of March. We are arranging interviews during January with key people to share what we are doing and to ascertain views and comment to feed into the report. Very briefly, the JISC is expecting the study should:

1. Investigate the costs (direct and indirect) of preserving research data, from an institution's point of view;

2. Construct a list of issues which universities will need to consider when determining the medium to long-term costs of data preservation;

3. Attempt to establish a methodology which will help institutions estimate the cost of research data preservation;

4. Compare the costs of each different model of preservation (eg. shared services, institutional repository, discipline focused, centralised);

5. Consider the direct and indirect costs of data preservation in the next 5-10 years and beyond.

Below is a list of issues and areas we are interested in exploring during our interviews. Not all may be applicable to/or answerable by every interviewee but these issues and questions should provide a useful checklist and framework for the interview:

Benefits of preservation of research data

How would you “sell it” within your institution?

Can you cite any good examples/case studies of benefits?

Do researchers recognise the need for preservation and if so do they look to the institution and/or to outside agencies?

Costs/Funding information/components

Do you have any in-house information on costs – current provision for research data storage within institution and at what level of the organisation e.g. School; Faculty; or Institution?

What are the main challenges for institutions in forecasting and planning current and future scale of research data preservation requirements?

Preservation costs in TRAC and funding streams

Views on “units of research data” – what are the key elements to consider in defining this?

Future demands and trend

Any longitudinal data on past and current data volumes/deposit/use/cost

Views on future demands and trends

Issues universities need to consider

Your suggestions

Do current systems take into account and deal with Digital Preservation to any extent?

Investing in infrastructure – timescales, capital planning and implementation at institutional level

Our Cost Model

How can we make it most useful for you?

Have you used/looked at any existing preservation cost models –what did you think of them?

Relationships and Models for preservation

HEIs and research councils/RC national services

How much is DIY by researchers within your HEI?

Any global or departmental services in your HEI?

What % is offered/accepted (question for national services)

Views on issues/cost components for different preservation models (shared service, discipline-based, centralised, institutional repository)

Would you be willing to comment on any draft outputs from study?

Any other comments/additional thoughts?