Charles Beagrie

Keeping Research Data Safe 2: The identification of long-lived digital datasets for the purposes of cost analysis

Project Plan

Version 1.0 Date 29 April 2009

A JISC-funded Project

contact details:
Name: **Neil Beagrie**
Position: **Director of Consultancy**
Email: **neil@beagrie.com**
Address: **2 Helena Terrace, College Street, Salisbury, Wiltshire, SP1 3AN**
Tel: **01722 338482** Fax: **0709 204 8179**

1. Background

Data has always been fundamental to many areas of research but it in recent years it has become central to more disciplines and inter-disciplinary projects and grown substantially in scale and complexity. There is increasing awareness of its strategic importance as a resource in addressing modern global challenges and the possibilities being unlocked by rapid technological advances and their application in research. However, there are several significant challenges facing the UK academic community relating to the long-term curation, storage, retrieval and discovery of research data. Recognising this, JISC has invested heavily in Higher Education repository and digital content infrastructure initiatives and developing support for digital repositories and preservation.

We believe identifying and developing longitudinal data on preservation costs and benefits associated long-lived data collections is critical in justifying and sustaining this work and for forwarding planning and effective resource allocation.

The “Keeping Research Data Safe 2” project aims to extend previous work on digital preservation costs for research data. It will identify long-lived datasets for the purpose of cost analysis and build on the work of the first “Keeping Research Data Safe” study completed in 2008.

The first Keeping Research Data Safe study funded by JISC made a major contribution to the study of preservation costs by developing a cost model and indentifying cost variables for preserving research data in UK universities. That work has had considerable impact and received international interest. Over 3,400 copies of the report were downloaded from the JISC website during 2008 alone making it JISC’s most popular publication in 2008.

However it was completed over a very constrained timescale of four months so there was little opportunity to follow up other major issues or data sources it identified. It noted that digital preservation costs are notoriously difficult to address in part because of the absence of good case studies and longitudinal data for digital preservation costs or cost variables. The study had identified potentially valuable data both within the 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. Recommendation 9 in the final report of the study therefore stated: “JISC should consider further detailed study of longitudinal data for digital preservation costs and cost variables to extend the work of this study”.

Recommendation 8 in the final report also noted “Additional work [is required] 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”. This would be critical to development of any costing tools based on the study’s cost framework and for implementation by institutions. It needs to be pragmatic so that collection of cost data is feasible and scalable for institutions and the benefits of collecting cost data are not significantly diminished by associated new effort required for its capture.

Finally Recommendation 10 in Keeping Research Data Safe recommends that JISC and/or other funders should consider funding further work on quantifying the benefits of research data preservation alongside costs. Costs and cost benefit analysis are closely linked and we believe where possible consideration of quantifiable benefits alongside costs could be a desirable aspect of the proposed study.

In addition to work on Keeping Research Data Safe, JISC has funded work on preservation costs for digital publications (LIFE) lead by the BL and UCL. Although covering different materials and approaches, this work provided useful input to Keeping Research Data Safe (and vice versa as Life 2 progressed). Similarly the draft DCC Curation Lifecycle was included in the approach for the study and mutual feedback was established as the cost model and the final Curation lifecycle evolved. Another major input to the study was the NASA Cost Estimation Tool (CET), which has also recently gone through further development since the publication of the study report.

2. Aims and Objectives

The Keeping Research Data Safe 2 project commenced on 31 March 2009 and will complete in December 2009. The project aims to identify and analyse sources of long-lived data and develop longitudinal data on associated preservation costs and benefits.

The objectives of this study are to:

- understand current requirements for the gathering of evidential material that will increase understanding of the long-term costs [and where possible the cost benefits] of research data preservation;
- review international literature for relevant initiatives;
- establish suitable criteria for identifying appropriate sources of data;
- undertake a large-scale survey of likely sources of data that may be appropriate for the aims of this study;
- analyse identified sources of data and associated information to determine their validity for the purposes of this study;
- liaise and negotiate with data owners and information providers to establish the terms on which information may be used;
- analyse the cost components and variables associated with the long-term management of the identified data and to compare and contrast them with the model proposed in the “Keeping Research Data Safe Report”;
- make recommendations of suitability for the further analysis and exploitation of specific sources of information.

3. Overall Approach

To achieve these objectives we will utilise the Keeping Research Data Safe cost framework as a tool for organising and scoping our work. We will undertake a combination of desk research; a data survey; analytical work with national and disciplinary digital archives that have existing historic cost data for preservation of digital research data collections; and interaction with digital archives in research universities who have little or no historic cost data but a strong interest in identifying criteria and metrics for capturing cost data going forward and in quantifying benefits.

Desk Research

We are already familiar with most of the international literature for relevant initiatives from work on Keeping Research Safe and more recently the cost/benefit work and literature review for the UKRDS Feasibility Study, and participation in the LIFE3 review workshop. We will update and review our existing research library from these projects to include recent work on LIFE2, the latest phases of development for the NASA Cost Estimation Tool, and other relevant initiatives. In addition to literature review, desk research will involve contacting existing relevant projects to obtain and share emerging reports, data and methodologies to feed into our data survey and analytical work. For example we have contacted NASA who have agreed to share their latest work. We have also approached MRC and agreed in principle collaboration with the new MRC Data Support Service. We will use our desk research to prepare draft criteria for identifying appropriate sources of data and potential models for terms and conditions of access to feed into our Data Survey and terms for future use of cost datasets created.

Data Survey

We recognise that the level of funding available will restrict the extent to which a data survey can be conducted. We therefore propose to use our existing knowledge to target resources appropriately to achieve the necessary scale. The UKRDS feasibility study has shown that one of the challenges will be that most universities have no previous activity costing data for preservation: the best longitudinal cost data for preservation often has come from national services that need to have charging policies or have long-term funding and reporting requirements. We have sought to address this in our methodology by distinguishing between those with existing cost data and those who have a prospective opportunity and interest in capturing and building up future longitudinal cost data. We hope by including both groups in the study draft criteria and data cost variables emerging from one group can be reviewed and if necessary tailored for application by the other.

The Data Survey will develop criteria and a survey proforma to identify key data collections and issues which will be piloted with two of our partners and then rolled out to other sites. We envisage completing the survey proformas via correspondence and site visits. We have incorporated within our project partners known large-scale collections in services with existing historic cost data which can be utilised for the study subject to agreed terms and conditions. We intend to explore structured sampling of these large collections and review of associated cost data as a major component of the data survey. These collections from project partners include:

The Archaeology Data Service at the University of York (staff: Catherine Hardman, Prof Julian Richards)

The UK Data Archive at the University of Essex (staff: Matthew Woollard)

The University of London Computer Centre (staff: Kevin Ashley)

In addition we will draw on collaboration with NASA to access data from relevant costing projects they are funding. Finally we will make an open invitation via email lists such as the JISCmail digital preservation list for others to contact us if they have research datasets and associated cost data that they believe would be of interest to the study. This will be supplemented by targeted personal approaches to some services such as the NERC Data Centres and STFC which we believe may have collections of interest.

For the second group, those who have a prospective opportunity and interest in capturing and building up future longitudinal cost data (and often some partial data on costs), we will work in the data survey with university project partners at:

University of Cambridge (DSpace@Cambridge repository: staff Elin Stangeland, Grant Young, Patricia Killiard)

University of Oxford (Scoping Digital Repository Services for Research Data Management project: staff Luis Martinez Uribe, Mike Fraser)

University of Southampton (Dept of Chemistry: staff Simon Coles, Prof Jeremy Frey)

Our university partners at Cambridge and Southampton were involved in Keeping Research Data Safe and we will be able to leverage their previous work and knowledge from the project. Oxford University were partners in the UKRDS feasibility study and have major research data preservation interests and plans for developing an institutional data service. The researcher survey conducted by UKRDS also has some potentially valuable input to the data survey.

Analytical Work and Case studies

Following completion of the desk research and data survey we will select the most promising collections and costs data for further analysis using the Keeping Research Data Safe cost framework as a tool for organising and scoping our work. We will consider a broad range of repositories so that differences in research data “collection levels” and consequent variations for example in metadata creation or access requirements are reflected in our cost data and methodology. The analytical work will combine face to face meetings and visits to project partners, data analysis and modeling, development of case studies and then review of draft findings with project partners. The UKDA case study may also target end-users of the ESDS and History Data Service for evidence of value-provided.

We will systematically analyse the cost components and variables associated with the long-term management of the identified data and compare and contrast them with the model proposed in the Keeping Research Data Safe Report. We will explore how the variables can be further quantified, and what additional data and data collection mechanisms are needed to support them. We recognise one of the most challenging areas will be identifying cost data and metrics for these variables which are both significant and available or easily captured by institutions. We will work with our partners to review all proposed metrics to ensure emerging findings and recommendations for future tool development and testing and validating cost models are realistic and achievable.

We also believe there would be value in considering how quantifiable benefits can be identified as part of this work on costs. Important work has been done by projects such as eSPIDA on identifying intangible benefits of digital preservation. This new study on costs may also help to explore the associated tangible cost benefits from digital preservation. This is a major area of interest to UKDA and will be developed as part of a case study focusing on social science and historical datasets as part of their contribution to the project. We believe another cost/benefit case study on data preservation can be developed from longitudinal cost data held at the Department of Chemistry in Southampton and their experience of data creation costs and data loss as profiled in Keeping Research Data Safe. These cost benefit studies could be a valuable addition to the ITTs required work on longitudinal data costs.

Review of emerging draft reports
We will agree an advisory group to undertake reviews of drafts of the report with JISC. We would include our project partners and other nominated services in consultation with JISC (e.g. perhaps the DCC and LIFE project). In addition we propose to use Sheila Anderson and Mark Thorley as formal expert peer-reviewers (and scrutiny by the JISC Executive) as proposed in the ITT. Sheila is Director of the Centre for e-Research at KCL and was a project partner in Keeping Research Data Safe. Mark Thorley is Data Co-ordinator at NERC and has worked across the Research Councils on data issues.

4. Project Outputs

The main project output will be a final report. An interim report will be submitted to JISC on 1st July 2009. We will also submit a project completion report for internal JISC consumption with the final report on Friday 11 December 2009.

We propose that the final report should be written to be read by a wide audience including non-technical senior staff in universities, research centres, and research funding bodies. It will be concise with an executive summary for easy assimilation of key points and provide a full account of the project and assessment of its outcomes.

We anticipate it will be approximately 40 pages in length plus appendices and will consist of the following components:

- Title Page, Preface and Contents (3 pages);
- Executive Summary (2 pages);
- Introduction (2 pages);
- Study Methodology (2 pages);
- The Data Survey (10 pages);
- Preservation cost components and variables and metrics for them(15 pages)
- Conclusions, implications for the Keeping Research Data Safe Cost Model, and recommendations for future work (6 pages)
- Appendices: Cost benefit case studies; description of datasets analysed; any financial data that can be placed in the public domain with agreement of partners; outline description and

terms and conditions for access to any restricted financial data; supplementary information on cost components and variables.

5. Project Outcomes

We believe identifying and developing longitudinal data on preservation costs and benefits associated long-lived data collections is critical in justifying and sustaining this work and for forwarding planning and effective resource allocation. Through this new study we hope to build on the achievements of previous work and to provide a larger body of material and evidence against which existing and future data preservation cost modeling exercises can be tested and validated.

6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
UK Research Funding Bodies	Research benefits, efficiency and value for money	Medium
UK Higher Education Institutions	Institutional budgets and strategic planning, research costs and overheads,	High
UK Research Data Centres and Services	Institutional budgets and strategic planning, research benefits, efficiency and value for money	High
Project Partners	Digital preservation costs and benefits	High
International and Overseas Organisations	Comparative digital preservation costs and benefits	Medium
Researchers and Research Support Staff	Research data preservation costs and benefits; grant overheads; project planning	High

7. Risk Assessment

A summary of the key risks and mitigating actions is provided below. In accordance with good practice, a risk register and issues log will be maintained throughout the project.

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
The number of partners and the co-ordination required to successfully manage the project.	3	3	9	We have allocated significant time from Neil Beagrie to the project to lead and co-ordinate input from the partners. He has worked with all the partners in previous projects and has extensive project management experience to deliver the study.

The scale of work required could exceed budget allocated by JISC.	3	4	12	The perceived benefits of the study have allowed us to substantially increase resources available for the work through the institutional contributions of the partners.
No useable cost data associated with long-lived datasets will be located in the data survey.	2	5	10	The partners include some of the most significant data repositories and research and data intensive universities in the UK. Appropriate long-lived datasets and potentially promising costs data has been identified in them in previous studies. They will provide a reliable and accessible core to the data survey.
Staffing and relevant skills and experience	4	3	12	There is a wealth of relevant experience and knowledge for this field of study in the assembled team and partners. There will be cross-project resilience & backup from the other team members and company associates should this be required due to any unforeseen events.

8. Standards

Name of standard or specification	Version	Notes
International Organization for Standardization [ISO], (2003), ISO 14721:2003 *Space data and information transfer systems - Open archival information system - Reference model.*	ISO 14721:2003	Used as comparative reference model for Archive Functions

9. Technical Development

Not applicable to this study.

10. Intellectual Property Rights

Under the terms of the project contract, all rights in the final report will be given to HEFCE on behalf of JISC.

The lead contractor will sign confidentiality agreements with the project partners who provide costs and other data for the project study. We anticipate that costs data identified in the study survey will be either be in the public domain (with the agreement of the data owner); only made available for this study; or potentially be available to any future third parties on request and subject to appropriate signed agreements. All such requests would need to be made direct to the data owner. Appropriate data status and contact details will be provided in the final report to facilitate any enquiries by future researchers.

11. Project Partners

The project is being undertaken by a consortium consisting of 8 partners as follows:
Charles Beagrie Ltd (lead contractor and project management)
OCLC Research, (co-applicant and contributing Brian Lavoie's expertise on economics of digital preservation to the study)
UK Data Archive (co-applicant and case study site, contributing Matthew Woollard to the study)
University of Cambridge (university partner)
University of Southampton (university partner and contributing case study)
Archaeology Data Service (Data Service partner)
University of Oxford (university partner)
University of London Computer Centre (Data Service partner)
All the partners bring considerable relevant expertise, knowledge and resources to the project and have significant data collections and interests in preservation costs.

Key personnel in the study are as follows:

Neil Beagrie is lead consultant and team leader for this study. He is an experienced project manager and a leading expert on digital preservation and curation with an international reputation across the archive, library, science and research sectors in the long-term management of digital assets. 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. Neil chaired the UK Office for Science and Innovation Digital Preservation and Curation sub-group and wrote its report. He is also a past chair of the Medical Research Council Data Sharing and Preservation Technical Experts Panel. In addition at AHDS he developed the application of the lifecycle approach to digital preservation which was subsequently published as a JISC/NPO preservation study. This approach has become an essential methodology for costing digital preservation utilised by Hendley, and more recently by the JISC-funded LIFE project.

Julia Chruszcz is a senior consultant for this study and has over 15 years senior managerial experience in UK HE academic computing service. She initially trained in industry as a Systems Analyst, and then moved to academic computing services in 1980. By 1990 Julia had moved into service management at the University of Manchester, becoming the founder Director of MIMAS and also worked with the Research Councils, principally EPSRC and NERC in establishing a national HPC service, CSAR. In 1995, as Head of National Services, Manchester Computing (MC) her role included a seat on the MC Board of Management. She subsequently assumed additional responsibilities as Head of Department for Academic Computing Services in January 2001. In October 2004, with the establishment of the new University of Manchester, Julia was appointed Deputy Director, MC. 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 is our adviser on cost and economic modelling for the study. He has a first degree and doctorate in economics and joined OCLC in 1996. He is currently a 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.

Matthew Woollard is contributing to cost and benefits analysis within the study. Matthew is Associate Director and Head of Digital Preservation and Systems at the UK Data Archive (UKDA). He is in charge of the development, implementation and maintenance of the UKDAs digital preservation policy and its framework. He has strategic and line-management responsibilities for the Digital Preservation and Systems section of the UKDA and primary responsibility for the development of vision and policy in the area of Digital Preservation more generally. He is responsible for the preservation of some 6,000 data collections. He represents both the UKDA and ESDS in national and international forums where such issues are debated and developed. He also retains oversight over the History Data Service and is a member of the editorial boards for the International Journal of Digital Curation (IJDC) and the International Journal of Humanities Computing.

12. Project Management

The study and project team will be managed by Neil Beagrie. The team is geographically dispersed so the project will utilise telephone and video conferencing, email and secure online document filestores and calendars provided by the Charles Beagrie for most of the project work supplemented by face-to-face meetings. Project team conference calls will be held fortnightly. A mid-project review meeting will be held approximately half-way into the project at a date to be agreed with the JISC Programme Manager. The project plan and milestones will provide the overall framework for monitoring the project.

Appendix – Detailed Work Packages

We propose to undertake this study in seven work packages as follows:

WP 1 - Project initiation;
WP 2 – Desk research and data survey;
WP 3 – Analysis and case studies;
WP 4 - Review by advisory panel, peer reviewers and JISC;
WP 5 - Report production and delivery of the draft and final reports;
WP6 - Dissemination;
WP 7 - Project management.

A start date of 31st March 2009 is assumed given existing commitments. Each work package, its staff allocation, milestones and deliverables is also described in more detail below:

WP 1 - Project initiation
Staff: allocation: Neil Beagrie 2 days, Matthew Woollard 1 day

Given the proposed project start date on or about 31 March, we would ask that JISC makes available a date between 2- 9 April for the project initiation meeting. The key outcomes of this meeting will be:

- Confirming project scope, timescales, structure and format of deliverables;
- Discussion and agreement of individuals on the proposed advisory panel.

The deliverable from this WP will be the project plan agreed by JISC and Charles Beagrie Ltd.

WP 2 – Desk research and Data Survey
Neil Beagrie 5 days, Julia Chruszcz 4 days, Brian Lavoie 6 days, Matthew Woollard 15 days, Research Assistant 10 days, other partners 20 days

This work package will consist of desk research and the data survey from mid-April to early July. Desk research and the data survey will be carried out by the consultancy team and our project partners as set out in our Methodology in section 3 above. The deliverables from this work package will be: draft criteria for identifying appropriate sources of data; potential models for terms and conditions of access to feed into our Data Survey; terms for future use of cost datasets created; the internal study files to generate sections of the draft final report and to inform development of the cost analysis and case studies.

WP 3 - Analysis and case studies
Neil Beagrie 6 days, Brian Lavoie 10 days, Matthew Woollard 15 days, Research Assistant 20 days, other partners 15 days

The work will be carried out between mid July and late September. Following completion of the desk research and data survey we will select the most promising collections and costs data for further analytical work. We will analyse the cost components and variables associated with the long-term management of the identified data collections and compare and contrast them with the model proposed in the Keeping Research Data Safe Report. We will assess the quality and completeness of cost information and how this should be factored into the study. In addition we will consider how benefits might be quantified as part of this work in two case studies at UKDA and University of Southampton. Deliverables from this work package will be the draft preservation cost components and variables and metrics for them and case studies sections of the draft final report.

WP4 - Review by advisory panel, peer reviewers and JISC
Neil Beagrie 1 day, Peer reviewers 6 days, Matthew Woollard 3 days, Other partners 5 days

We propose to seek review comments of our draft findings and sections of the final report from from mid October to mid November. Review comments will feed into the final version of the report.

WP5 - Report production
Neil Beagrie 10 days, Julia Chruszcz 1 day, Brian Lavoie 4 days, Matthew Woollard 4 days

Neil Beagrie will lead production and editing of the report by the study team. We will produce sections of the draft report from late September and peer review interim drafts (see WP5). We will then incorporate review comments and deliver the draft of the complete final report as a Word file by email on 23 November 2009. We would request JISC feedback by Friday 4 December. A final version of the report incorporating JISC feedback will be presented as a PDF file by email by Friday 11 December 2009.

WP6 – Dissemination
Neil Beagrie 8 days, Matthew Woollard 5 days, Brian Lavoie 2 days, Research assistant 2 days

At the commencement of the project we will establish a project webpage providing details of the project and links to other relevant work e.g. previous studies. This will be regularly updated as the study progresses. In addition we will make regular posting to the blog on the Charles Beagrie website (this has an established readership with over 200 subscribers to the news feed). Additional publicity may link to these sources from our project partners. During the course of the project we will participate in JISC programme meetings and other fora to publicise the project. We recognise the most important phase for dissemination and assuring take-up of the findings will be after completion of the project and beyond the funded phase of work. We have therefore committed 5 days as an institutional contribution from Charles Beagrie and 3 days from UKDA post completion to work with JISC on this key phase of dissemination.

WP7 - Project management
Neil Beagrie 5 days, Brian Lavoie 2 days, Matthew Woollard 2 days, Research assistant 1 day

The study and project team will be managed by Neil Beagrie. The team is geographically dispersed so the project will utilise telephone and video conferencing, email and secure online document filestores and calendars provided by the Charles Beagrie for most of the project work supplemented by face-to-face meetings. Project team conference calls will be held fortnightly. A mid-project review meeting will be held approximately half-way into the project at a date to be agreed with the JISC Programme Manager. An agreed project plan and milestones will provide the overall framework for monitoring the project. We will submit a project completion report for internal JISC consumption with the final report on Friday 11 December 2009.