KEEPING RESEARCH DATA SAFE 2

Neil Beagrie, Brian Lavoie and Matthew Woollard

with contributions by the Universities of Cambridge, Oxford, and Southampton, the Archaeology Data Service University of York, and University of London Computer Centre.

KRDS2 DATA SURVEY – SELECTION CRITERIA

Review Draft - 31 July 2009

Prepared by:

Charles Beagrie Limited

www.beagrie.com

A study funded by

JISC

Expressions of Interest/Comments can be sent to info@beagrie.com

THE DATA SURVEY

INTRODUCTION

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

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

The objectives of KRDS2 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 information on preservation costs for research data;
- undertake a survey of likely sources of information that may be appropriate for the aims of this study;
- analyse identified research data collections and associated preservation cost information to determine their validity for the purposes of this study;
- liaise and negotiate with research data collection owners and cost information providers to establish the terms on which information may be used;

- analyse the cost components and variables associated with the long-term management of the identified research data collections and to compare and contrast them with the model proposed in the “Keeping Research Data Safe Report”;
- make recommendations of suitability for the further analysis and exploitation of specific sources of information.

We have used our desk research and input from the project partners to prepare selection criteria for identifying appropriate sources of information to feed into our data survey. We are preparing a survey proforma to identify key research data collections with information on preservation costs and issues.

We have incorporated within our project, partners with known large-scale collections and existing historic cost information which can be utilised for the study subject to agreed terms and conditions. We explored structured sampling of these large collections and review of associated cost information 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).

Finally, we are making an open invitation via email lists and the project blog for others to contact us and contribute to the data survey if they had research datasets and associated cost information that they believe may be of interest to the study. This will be supplemented by targeted personal approaches to some services which we believe may have collections of interest.

Definition of Research Data

For the purposes of the KRDS2 study research data is defined as collections of structured digital data from any disciplines or sources which can be used by academic researchers to undertake their research or provides an evidential record of their research. Research data may be created in a number of different contexts: for reasons entirely unrelated to academic research; for academic research or as a by product of (academic) research. It includes a great variety and heterogeneity of data and its accompanying metadata and documentation to make it usable and understood, or the digital representations and records for physical

research data. In essence any type of research data already held in data repositories would be in scope. Examples could include: complex data used in climate modelling, aerodynamics, molecular modelling, bioinformatics; video and image archives used in archaeology, art history, anthropology and performance works; digital images/investigatory data of primary physical sources in the humanities; quantitative and qualitative data used in the social sciences; or electronic data and indices for fossils or skin tissue samples.

Our Selection Criteria

The following criteria define what material is required for this study – "Keeping Research Data Safe 2 (KRDS2)" which aims to build on and extend the original "Keeping Research Data Safe (KRDS1)" study report (Beagrie, Chruszcz, Lavoie 2008).

Each criterion is shown here in three sections: First, a simple single line criterion; second, a short paragraph with some guidance and detail for the potential respondents and participants; and third any internal information or guidance for project staff which may prove useful to the project in due course.

Criteria are divided into "essential" and "desirable" and further sub-divided into those which are repository specific and those which are needed at the level of the project as a whole.

Essential Criteria

For the Repository

1. The Repository must have cost information for research data relevant to at least one of the activities in the KRDS2 activity model or a KRDS2 key cost variable.

Guidance: the key cost variables and the activity model are both available in the Keeping Research Data Safe report and updates will be published in the data survey. Repositories should also explain where necessary where their internal procedures differ from the published activity model.

Cost information need not cover all aspects of the preservation life-cycle and may be at different levels of granularity across it. However, the more complete the coverage (i.e., all activities and all relevant cost variables) and the more detailed the coverage (i.e., below service level) the better. Cost information, if applicable should separate out costs of dealing with digital and non-digital research data. Reporting should also

include where other activities / cost variables are missing through unavailability or irrelevancy to the repository or are incurred elsewhere by other organisations/individuals.

Internal Project Note: Some material may have to be excluded from more detailed analysis if coverage is poor. The more complete the coverage the higher the priority for further analysis.

2. The cost information provided must be decomposable into a form that at least roughly approximates the KRDS2 key cost variables or activity model.

Guidance: see appendices. Reasonable adjustments must be able to be made where the mapping between a repository's activities / cost elements is not completely aligned.

Internal Project Note: Some material may have to be excluded if it is not easily decomposable.

3. The Repository must be able to provide a good sample for its costs information.

Guidance: the information available must be sufficient to capture variability in costs for the same type of collection/material and/or different types of collection/material e.g. it is not based on a single example.

For KRDS2 Project Overall

4. For the study to succeed it must have enough repository sources to cover a range of the KRDS activities/sub-activities and key cost variables.

Guidance: The study probably needs to be comprehensive at the highest Archive Phase activity levels e.g. Acquisition, Ingest, etc. Sub-activity information e.g. negotiate submission agreement may be less easy to obtain and need to be more selective. It will be important to identify cost elements which are not applicable, unknown/too difficult to obtain and why.

Internal Project Note: May not be able to be comprehensive on Pre-Archive Phase. Support Services and Common Services costs may be based on formulae.

5. For the study to have broad applicability there must be a range of repository types to illustrate the possible differences which may arise because of: repository mandates /types/ service levels; repository reliance on specific disciplinary types of data; repository reliance on specific types of file.

Guidance: The study has the potential to be less widely applicable if any one of these is not dealt with. Without representation from across these three (potentially) interlocking criteria the study will be of more limited value. The underlying heterogeneity which these three elements / contexts may provide could be used to interpret and explain significant differences or patterns in the cost information.

Internal Project Note: given limited resources, the study also has the potential to fail / be inconclusive if it seeks to analyse further in depth too extensive a number of repositories from the data survey.

DESIRABLE CRITERIA

For the Repository

6. The Repository should be able to provide time series for costs.

Guidance: Wherever possible, repositories should provide comparative cost information over time. For the purposes of this study, retrospective cost information is preferable to estimates of future costs. The latter will be considered if the underlying assumptions and models can also be made available and if they are limited to a 12 month time horizon. Note the project could be interested in past archiving cost projections for projects and what you have learnt from actual outcomes over time.

7. The Repository may be able to provide some illustrative material to assist in understanding potential cost/benefit relationships.

Guidance: The KRDS2 study will include two case studies to illustrate benefits as well as costs and further illustrations would be welcome. Note this is subsidiary aim

of the study: KRDS2 will be carried out at the same time as a RIN/JISC study on use/impact of data repositories and these studies should dovetail where possible.

Internal Project Note: This may be best captured in discussions rather than in a formal survey. See if we can illustrate cost variations as service and benefits change.

For KRDS2 Project Overall

8. Detailed analysis of cost information by the project should be from the UK repositories where possible or costs information should be able to be adjusted/qualified for international environments.

Guidance: All international sources of costs information may be of interest to the data survey. If there are lacunae within the UK based samples then cost information from elsewhere which can be adjusted to take account of any potential differences in international environments can be considered for further analysis.

Internal Project Note: Additional workload may be necessary to disentangle some of the potential issues arising from different legislative, legal and rights environments if more detailed analytical work is proposed by the project.

9.The Repository allows cost information to be made available for scrutiny by other researchers.

Guidance: this is not a mandatory requirement but may help others leverage work completed by the project. Terms of potential use/access and confidentiality can be set by the repository and open access is not assumed given some information may be commercial or sensitive.

Internal Project Note: A confidentiality agreement is available for use by the project. This sets tight access criteria as a default which can be relaxed in writing by the repository i.e. they set terms of access and dissemination for the project and to others.

10. To have information sufficient to provide results of interest and relevance to both national and institutional services in the UK.

Guidance: Currently national data services are those most likely to have information on preservation costs. Their service levels and target user communities may differ substantially from university services but may still help to inform costs data models for others. Universities services are included amongst KRDS2 project partners to ensure differences can be identified and appropriate models and metrics developed.

Internal Project Note: This is perhaps an obvious statement, but must remain in our minds.

REFERENCES

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