

# **e-Science Curation Report - Appendices**

## List of Appendices

- 1. Interviewees**
- 2. e-Science Curation Task Force meeting Report**
- 3. Questionnaires: Detailed findings**
- 4. Questionnaires and covering letters**
- 5. Invitation to Tender**
- 6. The Digital Archiving Consultancy Team**

## Appendix 1: Interviewees

### Methods and Interview Panel

Interviews were arranged with a representative range of individuals chosen on the basis of their acknowledged expertise in this area, or because of their key position or because of their ability to represent a community of interest. The objective was to spread our net as widely as possible within budget and time constraints; we spoke to the most senior of academics through to support staff; the range of expertise in curation sampled ranged from the expert to those unfamiliar with the subject; all the major science disciplines were represented as well as the arts, administration and law. The sample included the Research Councils, universities, data archiving services, JISC, and the commercial sector.

This coverage gives us confidence that all major threads were identified. Assembling the one-day curation strategy task force and organizing the one-day meeting to discuss these issues supplemented this aspect of our work, and the report<sup>1</sup> provides a comprehensive discussion of the issues task force members raised.

Over thirty five interviews were conducted. The majority were held in person, some were conducted by telephone or supplemented by e-mail exchanges. Many very busy people were very generous with their time, spending much more than the expected hour and a half with us, and we would like to thank them for their notable and enthusiastic engagement with the process. Very full schedules did prevent us from seeing a few people we would very much liked to have spoken to, but this was often because they were out of the country. As interviews were generally held at the interviewee's offices they generally involved some travel, and to increase efficiency and reduce costs we attempted where possible to schedule visits on a city-by-city basis. We had no refusals, but one or two people did prove impossible to contact. There were many other people we could have approached, but time and budget were a constraint.

Other business took Philip Lord and Alison Macdonald to San Diego, and this enabled them to enjoy a second meeting with Professor Reagan Moore of the San Diego Supercomputer Centre at the University of California in San Diego; we were also able meet Constantine Scheder and Greg Granello of Nirvana, the division of General Atomics Inc. which is introducing a commercial version of the Storage Resource Broker (SRB) developed by Professor Moore's team. Unfortunately they were unable to arrange a meeting with Dr Jeff Rothenberg of the Rand Organisation at Santa Monica, to hear his views on the subject (although Mr Lord has had previous meetings with him). Jeff Rothenberg is an expert in the field of digital preservation and an advocate of emulation techniques for continued accessibility; he was one of the earliest workers to signal that there are significant barriers to digital longevity.

Professor Mike Freeston, though based at the University of Santa Barbara, was seen on a visit of his visit to Southampton.

---

<sup>1</sup> Macdonald, A. and Lord, P.

Dr. Jeremy Frey, of the University of Southampton kindly invited us to attend a meeting of the management team of the Comb-e-Chem project, one of the e-Science Core Programme pilot projects. This provided valuable insights into the management of projects and the interests and concerns of project participants.

The people interviewed are listed in the table below. We do not list the many people with whom we have contact in this area but who were not formally interviewed as part of this study. Nor do we include here the progress meetings held with Neil Beagrie and Professor Tony Hey, nor those met at the curation strategy task force or MRC Archiving Horizons days.

**Table A1-1 Interviewees**

<b>Interviewee</b>	<b>Affiliation</b>
Dr. Sheila Anderson	AHDS
Dr. Hamish James	
Dr Richard Baldock	MRC Human Genetics Unit, Edinburgh University
Dr. Dave Berry	National e-Science Centre, Edinburgh
Dr. David Boyd	
Dr. Kerstin Kleese	CCLRC
Dr. John Gordon	
Prof. Peter Buneman	University of Edinburgh
Dr. Andrew Charlesworth	University of Bristol
Dr. Dick Clements	University of Bristol
Dr. Andrew Coward	Occam, Southampton Oceanography Centre
Dr Martin Dove	NIEeS, Cambridge University
Prof. Tony Doyle	University of Glasgow
Dr. Peter Dukes	MRC
Dr. Susan Duthie	Rowett Research Institute
Prof. Mark Freeston	University of Santa Barbara
Dr. Jeremy Frey	University of Southampton
Prof. Carole Goble	University of Manchester
Prof. Julia Goodfellow	BBSRC
Mr Greg Granello	Nirvana Storage (General Atomics)
Dr Constantine Scheder	
Prof. Tony Hey	JISC
Dr. Bryan Lawrence	British Atmospheric Data Centre
Dr. David Lowe	John Innes Centre
Dr. Bob Mann	Institute for Astronomy, Edinburgh University

Interviewee	Affiliation
Dr. Colin Miles	BBSRC
Dr. Christine Thomson	
Prof. Reagan Moore	San Diego Supercomputing Center
Dr Anthony Morgan*	Cardiff University
Dr. Geoff Oldham	Institute of Animal Health
Mr. Vince Osgood	EPSRC
Prof. Norman Paton	University of Manchester
Dr Tim Philips*	University of Bristol
Prof. Alan Rector	University of Manchester
Dr. Seamus Ross	University of Glasgow
Mr. Fred Hopper	
Mr. Mark Thorley	NERC
Mr. Rod Bowie	
Prof. Kevin Schurer	UKDA, University of Essex
Dr Michele Shoebridge*	University of Birmingham
Dr. Ian Stewart	University of Bristol
Prof. Janet Thornton	European Bioinformatics Institute
Members of finance and administrative departments	BITS, MIMAS, EDINA, Wellcome Institute

Most interviews were conducted with a single person; a handful were joint (see groupings in Table A1-1).

It would have been inappropriate with such a diverse group as this to adopt a formal interview approach and use a proforma; indeed many interesting themes emerged during the discussions which the free format allowed us follow up with profit. Some of these excursions were quite extended.

There were three exceptions to this method, where a letter was distributed to Vice Chancellors, and responses were received from individuals nominated by them. (These responses are indicated by a \* in the table above.) These three replies showed a large degree of agreement: they all believe specific funding for this should be made available; the institutions do not have data retention and curation policies; all favoured establishing a UK central services unit; and all would prefer to see responsibility for curation staying with the originating institution, but would be prepared to share facilities.

## Appendix 2: Digital Data Curation Task Force Report

---

### Digital Data Curation Task Force

#### **Report of the Task Force Strategy Discussion Day**

Tuesday, 26<sup>th</sup> November 2002

Centre Point, London WC1

**Prepared by:** Alison Macdonald and Philip Lord

The Digital Archiving Consultancy  
2 Wayside Court  
Arlington Road  
TWICKENHAM  
TW1 2BQ

January 2003

## Management summary

We are entering an era in which digital data resources are becoming a central pillar of scientific research. Data volumes are increasing exponentially, as too is the complexity of the data itself; this will be magnified by the spread of Grid infrastructure and technologies. Some of the data will have considerable scientific value in itself, some data may have value from perspectives as diverse as commercial use or historical research. Already a significant amount of scientific work is conducted on previously collected data (collection-based science).

The data generated in this deluge requires active management to meet basic needs of access and re-use: data needs to be retained so that it survives, so that it can be found and retrieved as appropriate, understood within and across disciplines, and re-use must be possible; this needs to happen efficiently, fairly and affordably in contexts we cannot today predict. But in addition, digital technology may offer opportunities to incorporate such data more valuably into the knowledge base and extend the reach and value of the data. Ambition in this area could be rewarded by substantial and enduring benefit and scientific advance.

In the light of this problem and opportunity, Professor Tony Hey, chairman of the JCSR (the Joint Information Systems Committee's Committee for the Support of Research) assembled a task force to work towards defining and structuring a strategy for the "curation" of primary research data in the UK. The task force membership represents a broad range of expertise in the area of digital curation drawn from academia, the Research Councils and private industry. This report summarizes the discussion at a meeting of the Task Force which took place on 26<sup>th</sup> November 2002.

The application of the term "curation" is new, and in several ways the meeting found itself grappling with questions of scope, with frequent overlap with questions relating to digital preservation. It did not reach a definition of the term.

There was almost unanimous agreement that there are generic, inter-disciplinary areas where provision of a curation service and research would be useful. Above all, however, the meeting identified a need to establish a rationale for curation and proof of concept, suggesting exemplar research projects, science-led, which would demonstrate or otherwise the benefits and value of re-use of primary research data. The meeting touched upon but did not specifically explore the question of stakeholders.

An area of major concern at the meeting was the cultural problem of getting researchers to submit data and to provide the necessary contextual information for the data to be meaningful and useful in the future.

This is a factual report of the day's proceedings, which over its course touched upon many aspects and issues relating to the re-use of research data. The report seeks to set a firm structure to serve as basis for comment from task force members, towards developing an approach towards a curation strategy for this area, and the identification of issues requiring deeper consideration and areas not covered.

## Contents

Discussion was wide ranging and is reported here under seven headings:

1. What is curation?
2. What are we keeping?
3. Costs, benefits - why keep primary research data?
4. Exemplar research projects,
5. “How” - incentives, the role of journals,
6. “How” - data and curation,
7. Curation aims and strategy.

Appendix 1: Task force members

Appendix 2: Agenda

Appendix 3: Round-the-table priorities and topics for research

Appendix 4: “What a strategy for research data curation should address” - pre-submitted answers

Appendix 5; “The three most important issues to be addressed in a curation strategy” – pre-submitted answers.

## Background

At a meeting in October 2002 Professor Tony Hey asked Philip Lord and Alison Macdonald of the Digital Archiving Consultancy ('DAC') to bring together a task force for a one-day brainstorming session about the future shape of curation of the UK's primary research data, in particular scientific research data. The task force was assembled from individuals with expertise in this area, suggested by Professor Hey, Neil Beagrie and Philip Lord. The aim was to span different disciplines, with university, Research Council and corporate sector representatives. The full list of the task force members (members present at the meeting and those unable to attend) is given in Appendix 1; the agenda is attached as Appendix 2.

The DAC had the support of Neil Beagrie and the JISC team in organizing the day, finalizing the agenda, and would like to thank them all for their help and advice.

Ahead of the meeting task force members were invited to send a short note with suggestions for the three most important issues to be addressed in a curation strategy; these were circulated at the meeting (anonymized) and are attached here in Appendices 4 and 5 (with attributions).

Tony Hey summarized the reason for the meeting: we will be creating very large amounts of data in research in the next decade, and we are going to have to save some of it. He would like to understand where to put enough money to make a significant impact in this area, to support scientists and to create a centre of expertise in the UK, with a world-leading reputation.

The meeting began with short presentations from Neil Beagrie, Philip Lord and Rolf Apweiler, to set the scene<sup>2</sup>.

Neil Beagrie is Programme Director for Digital Preservation for the Joint Information Systems Commission (JISC) and also Secretary of the Digital Preservation Coalition. Implementation of the 2002-2005 JISC strategy for continuing access and digital preservation began in November 2002. Its aim is to provide a mix of national, possibly also regional, and institutional services, co-ordinating and partnering with other bodies as well. A cornerstone is the development of a digital curation centre. This is not envisaged as a data centre but will seek to provide a set of central services, standards and tools.

The Digital Archiving Consultancy is carrying out a study for the JISC's Committee for the Support of Research on the curation of primary research data, in particular in the context of the Grid and the UK's e-Science programme, assessing current provision and future requirements.

Rolf Apweiler is head of the SWISS-PROT Sequence Database Group, part of EMBL, the European Molecular Biology Laboratory. The SWISS-PROT Protein Knowledgebase is a

---

<sup>2</sup> Please contact the Digital Archiving Consultancy if you would like a copy of Neil Beagrie's and/or Philip Lord's Powerpoint presentations and/or a summary of Rolf Apweiler's presentation (taken from the meeting transcript).

curated protein sequence database that provides a high level of annotation and high level of integration with other databases. It is curated by 50-70 curators, who check for errors by submitters and provide annotation. Recruitment of curators is difficult. SWISS-PROT is human labour-intensive; usage is high, with about 200,000 scientists accessing it, and one million records accessed each day. SWISS-PROT is about 100 gigabytes, but doubling in size and complexity are problems. Standardization is difficult. Another need is to convince journals that data must be deposited before papers are accepted.

## 1. What is curation?

What do we mean by curation? Tony Hey took up the term which had been used by Dr John Taylor, Director General of the Research Councils, to distinguish the actions involved in caring for digital data beyond its original use, from digital preservation. The concept's reach extends beyond libraries.

For Seamus Ross, "curation in the museum sense" covers three core concepts – conservation, preservation and access. David Holdsworth noted that access implies preserving data and making sure that the people to whom the data is relevant can find it - that access is possible and useful.

Alison Allden noted that the interpretation of the word "curation" implied in the discussion was of an active management of information, involving planning. She also made the point that re-use of data is a core issue. If data is to be re-used, then it needs special treatment.

For Rolf Apweiler, access does not form part of curation. Indeed, the activity described by Seamus Ross and David Holdsworth he regards as conservation; curation in his eyes is when people add value to data. Jeremy Frey, however, felt that curation as described by Rolf is research work in itself - managing, improving, enhancing data.

Peter Buneman has subsequently noted that it is important to address the issue of curation of databases. The crucial observation is that databases, unlike documents, evolve: they change to reflect/represent the changing state of scientific knowledge. Much scientific "publication" now happens through a process of augmenting or modifying existing databases. As one example of the issues involved, most curated databases consist to a greater or lesser extent of data copied from other curated databases. Very few systems do a good job of telling you where they get their data from, and hence one has little guarantee of the quality of information. This is going to be a major challenge.

## 2. What are we keeping?

For Mikhael Dahlin, appraisal is part of curation. Appraisal is the selecting of data; selection also requires capture of the context of why the data was created, in what environments – without which the data itself will not be meaningful or useful in time. To date, we do not yet have the capability to capture all the contextual information or structure of data automatically. This data is most easily collected by its originators (see compliance issue below, section 5).

Selection is a difficult issue. For instance, who selects? While experts may know enough to decide whether to keep primary data from their own perspective, for scientific or

regulatory reasons, we know that we cannot predict the ways in which that data may be used in the future, whether for scientific, historical or other purposes, nor can we predict the tools which technology may produce for re-using old data. An examination of legacy data might reveal clues as to what to keep and what to discard.

One criterion for retention is reproducibility. Some data is readily identified as non-reproducible (typically, observational data), though Jeremy Frey also warned that we sometimes think data can be reproduced only to discover later that it cannot. Data which would be very expensive to reproduce could also be included in this category.

History is another criterion for keeping primary data, though it was noted both that historians are practised at working on fragments which have survived by chance, and that the needs of historians might not be sufficient to justify the cost of keeping data.

Alison Allden pointed out that in some cases what we should be keeping is not so much the data itself but the process by which it was reached. Data may be more valuable with the process which generated the data, or it may be just the process which is of value.

Philip Lord pointed out that data will have to undergo re-appraisal over its life.

David Holdsworth suggested that all data could be kept as a matter of course, overcoming the selection problem, because the cost of storage quickly becomes trivial - the cost of storing it arises primarily at the time of collection. This view was not universally shared. However, it was interesting to note that under Sweden's Freedom of Information law, which dates back to 1766, all publicly owned records are retained, in the interest of the nation.

### **3. Cost, benefit - why keep primary research data?**

At some stage, it is inevitable that cost will have to be justified. Sean Barker believed that it was important that in a curation strategy you have a consistent way of justifying the data kept, preferably in your own terms.

Alison Allden noted that the ESRC has made a strategic decision to collect and preserve data; this now represents quite a large overhead, both in terms of capital and recurrent costs, and is a long-term commitment. Looking forward, Seamus Ross remarked that data curators may find themselves having to decide, as traditional librarians before them, that they can stop holding data; Mark Thorley noted that librarians can get rid of a journal series, confident that there is always another library holding, but the same may not be true of data.

Tony Hey was struck in Rolf Apweiler's talk by the large numbers of people working on curation in EMBL and Swiss-Prot/TrEMBL (50-100 curators). This seems a huge cost – is it sustainable? In Rolf Apweiler's view this investment is extremely cost-effective, as it means that it is done centrally, so that everybody else does not need to do it on their own, which would be very much more expensive. Similarly in the corporate sector, at AstraZeneca for example, while some curation is necessarily dispersed, as much as possible is done centrally, allowing the organization to spend fewer resources and at the same time achieve higher quality.

Alison Alden noted that the level of re-use of data held in the AHDS and ESRC archives has been disappointingly low. This might be because for want of active encouragement of use. On the other hand, many examples were mentioned during the day of the benefits of re-use of data: Seamus Ross gave the example of the Hubble Telescope, where more data has been published in the last three years from original data re-used from the Hubble telescope than from new experiments.

At the macro-economic level, an open government policy on data sharing is reflected in increased financial revenues for government in the form of tax receipts – for the USA the income is significant, creating approx. €750 billion per annum, as compared with the €68 billion or so in the EU, which does not operate the same open policy, focusing on direct cost recovery. Commercial models may provide pointers to cost recovery through revenue generation from exploitation of data.

Peter Dukes made the point that the question of data value raised the ownership issue: ownership tends to impose rights and restrictions, as opposed to custodianship, which is about ensuring that quality is maintained over the data over a period of time. Another question is whether and how distributed ownership and custodianship might affect the quality of care of information, and the need for continuity of funding.

#### **4. Research project exemplars to identify benefit**

Mark Thorley said that NERC spends about £5 million per annum on data management, but he is not sure what benefit NERC derives from this. He would very much like to see research which seeks to establish benefits and value of data re-use. Indeed, at different junctures throughout the day there were calls for exemplar projects which could seek to identify benefits of re-using data. These projects should be limited in scope; they should be science-led so that results of scientific interest can be demonstrated to the community. These projects would work towards answering the question of what we need to know to use other people's data (probably variable from field to field). As Alison Alden noted, they would also provide a useful way of testing the preservation questions that need to be asked.

Peter Dukes noted that the re-use of data has already spawned a new research community around meta-analysis; in the epidemiological domain its work can be evaluated on the basis of the science generated.

Mark Thorley and Liz Lyon were both interested in the light which inter-disciplinary research might shed: Mark Thorley believed that inter-disciplinary work drives good data management, as it is a requisite for efficient collaboration. Liz Lyon is looking forward to seeing work generating inter-disciplinary datasets, and the issues and benefits which arise from this.

#### **5. “How” questions – incentives, the role of journals**

A key benefit of the demonstration of the scientific value of good data curation would be to encourage compliance on the part of the data creators. This is one of the major problems facing curation. At the moment researchers have no incentive either to submit data or to add contextual information to their datasets; their goals are publication of research papers in journals and subsequent citation.

Indeed, in the academic world the situation is deteriorating as researchers are increasingly anxious to put their own data on their own web sites rather than submitting data to journals.

The value of keeping the data is not apparent to its creator, nor does that value usually revert to its creator. As Sean Barker noted, this also applies to the corporate sector.

The meeting considered whether rules and penalties might help. Mark Thorley said that NERC's conditions of employment require researchers to submit their data to their data centres, but this was not entirely successful. Jeremy Frey wondered whether funders might achieve greater compliance by making, say, the last 10% of funding conditional on preparation of data for long-term retention and actual submission of data. Research Councils, universities and other funders should apply pressure on researchers in this regard, and journal publishers might also be recruited to this effort.

David Holdsworth wondered whether a mechanism could be developed to enable recognition of citation of or access to datasets, in place of and in addition to citation of papers. This would require accurate links to be maintained between data and papers – a provenance issue - and might also raise the profile of the datasets. It also implies the need for good curation.

In the traditional research process, typically data is gathered, from which information or knowledge is extracted, accumulating knowledge. The curation process should tap into this process. This process of research to publication has influenced the data life cycle. Publication generally lies at the end of this process – but the process is currently under challenge. Can the research-to-publication process be enhanced to help preserve value in data? Alison Allden made the point that one of the key strands of the strategy is to be able to tie the data to the information to the knowledge, and it needs to be seen to be of value. Once people are used to publishing on-line, the fact that a publication ties back to data will become more relevant. David Holdsworth referred here to the CEDARS architecture, which has a two-level follow-through; this raises an issue for the curator, if the life cycle includes removal of data, there may be “dangling” pointers without anything at the other end.

The meeting also mentioned the possibility of an on-line journal of data which is being publicly discussed, which might be maintained by an academic publisher.

The meeting agreed that a major need is a change in culture, so that curation, preservation actions become “what we want to do”. While training young scientists will help, this will not feed through in the near term. Awareness campaigns can also help, but the fundamental need is for incentives to submit data.

## **6. How - data and curation**

A curation centre should disseminate advice on good ways to store information in the first place, inform about preferred standards, data formats, and co-ordinate the development of standards where they are needed.

There are several areas where lessons might be learnt from the corporate sector. To a certain extent companies face the same problem of co-operation with data submission and

metadata provision, and it might be useful to examine companies' approach to this problem. In companies such as BAE Systems, of course, data generation takes place in defined contexts, and data goes into data management systems where a certain amount of context is pre-existing or pre-defined. Sean Barker and Seamus Ross both mentioned the ISO set of standards known as STEP as a useful model to study. There should be a life cycle approach to information, and also possibly to infra-structures within which curation takes place.

The systems and curation requirements may be quite different according to type of data. One distinction is between closed and open datasets (and variations in between). The European Bioinformatics Institute datasets are highly dynamic, multi-dimensional, with continuous accrual. With a closed dataset you get an end-of-study snapshot. Interestingly, the volume and complexity of the EBI data are increasing, and so too are the questions that people ask of the data, so the skills within the database need to be greater, ditto the resources needed to answer the questions.

The stage at which data is captured is another issue. For example, capturing data after calibration could add in a layer of risk for future reading of the data. Risk analysis is essential, therefore, before the data capture stage: preservation strategies must include risk analysis.

David Holdsworth pointed to the need for research into immutable, unique, persistent named entities so that data can be located reliably.

Different data have different types of value. Some data only gain their value after being corrected and cleaned, other data has a immediate value on creation; we need to be careful that the sharing strategies make due and adequate allowance for data originators to have fair use of their data before the data's release. Old data may also include inaccuracies, or be overtaken by more powerful tools and technologies - nevertheless, we cannot predict how old data might be used or why it might be needed, and inaccuracy or poor resolution, for example, should not necessarily be grounds for disposal of data.

We should look at curation retrospectively as well as prospectively. In particular there is a need to capture tacit knowledge before it becomes inaccessible. In newer ways of working, collecting this information will be built in prospectively to study design, it will be part of the culture according to which people work. Another concern raised by Peter Dukes is the problem of locating old datasets, and he wondered about the creation of a single portal through which searches for datasets might be directed.

## **7. Curation strategy, aims**

The general view was that there are common areas and principles, shared across disciplines, relevant to curation. Liz Lyon cited information discovery and access to data as examples of common functions, and the distinction between observational data (which cannot be recreated) and experimental data. Rolf Apweiler's was the dissenting voice, taking the view that there cannot be a single strategy - bioinformatics for instance is highly complex, domain-specific - you need to know your own domain and develop strategy accordingly.

It was also generally recognized that different disciplines have different problems. As aired by Philip Lord at the start of the meeting, there might be an umbrella strategy, with disciplinary pillars, in Peter Dukes' phrase: the overall strategy should provide coherence and prevent multiple re-invention of wheels. It will be important for each discipline to "put its house in order", agree its own vocabulary.

Data curation requires resources for as long as the data is managed, but curation strategy needs review, to take account of changing technology and also changing scientific context.

The activity of curation includes research into curation. In addition to exemplar research projects, there were suggestions for generic research. One of these from Alison Allden was modelling; she noted that there are probably some existing models on which work could draw, in the archival domain. Philip Lord pointed to the crying need for an agreed vocabulary for this domain, while Mikhael Dahlin stressed the need for clarification of where boundaries lie, between archiving and libraries, for example.

Alison Allden suggested that one of the first questions to ask is whether we have enough proof of concept before making significant investments in curation.

At the end of the day Tony Hey asked for each person's top issue(s), and these are given in Appendix 3, followed by suggestions for research projects made during the day.

## Appendix 1 (To Curation Task Force Report)

### Task Force Members

✓ Attendee at 26<sup>th</sup> November 2002 meeting

A = break-out session A, B = break-out session B, C= break-out session C\*

✓ A Tony Hey (Host) JISC Committee for the Support of Research, Director, e-Science Core Programme

✓ B Neil Beagrie JISC, Digital Preservation Coalition

✓ A Alison Alden University of Warwick (now at University of Bristol)

✓ B Rolf Apweiler European Bioinformatics Institute

✓ C Sean Barker BAE Systems

Peter Buneman University of Edinburgh

Andrew Charlesworth Bristol University

✓ B Mikael Dahlin AstraZeneca

Lorraine Estelle JISC

Mike Freeston University of Santa Barbara / University of Southampton

✓ B Peter Dukes Medical Research Council

✓ A Jeremy Frey University of Southampton

✓ A David Holdsworth University of Leeds

✓ B Philip Lord The Digital Archiving Consultancy

- |   |   |                  |  |
|---|---|------------------|--|
| ✓ | C | Liz Lyon         | UKOLN  |
| ✓ | C | Alison Macdonald | The Digital Archiving Consultancy  |
| ✓ | A | Bruce Pilsworth  | The Digital Archiving Consultancy  |
| ✓ | A | Seamus Ross      | University of Glasgow, Director of Humanities Computing and Information Management & ERPANET |
|   |   | David Ryan       | Public Records Office  |
| ✓ | C | Mark Thorley     | Natural Environment Research Council   |

\* Session A: Technical issues

Session B: Custodianship issues and implications

Session C: External factors affecting strategy.

## Appendix 2 (To Curation Task Force Report)

### Agenda - Digital Data Curation Strategy

**Task Force Discussion Day – Chairman: Professor Tony Hey**

Tuesday, 26<sup>th</sup> November 2002

Centre Point, London WC1

9:30 hrs.	<b>Coffee</b>	
10.00	<b>Opening session:</b>	
	Introduction	<i>Tony Hey</i>
	Presentation on JISC digital curation & preservation work	<i>Neil Beagrie</i>
	JISC's e-Science curation study – summary	<i>Philip Lord</i>
	Curation for the SWISS-PROT project	<i>Rolf Apweiler</i>
10:30	<b>e-Curation strategy: establishing the objective of the strategy:</b>	<i>Round- table discussion</i>
	Discussion initiation, presentation of summaries from task force members; discuss and agree definition of objective	
(15-minute coffee break c. 11.15)	<b>Formulating a strategy to achieve this objective</b> What are the fundamental questions? Global/umbrella strategy, plural strategies? Time frame? What do we have at starting point?	<i>Round- table discussion</i>
	Pre-lunch summary, presentation of afternoon: break-out sessions after lunch look at factors.	
12.30	Lunch	
13.30	<b>Break-out sessions: three groups:</b>	
	<i>Subject to review by the group, suggested topics for more detailed examination in break-out groups are:</i>	
	a) Technical issues	
	b) Custodianship issues and implications	
	c) External factors affecting strategy	
14:30	<b>Reporting back from the break out sessions</b>	Group representatives
14.50	<b>Tea break</b>	
15.00	<b>Synthesis:</b> Review of objective; identification of elements of strategy.	<i>Round table discussions</i>

Closing remarks

*Tony Hey / Neil  
Beagrie*

16:30-17.00 End

## Appendix 3 (To Curation Task Force Report)

### Round-the-table priorities and topics for research

The meeting ended with a round-the-table view each person's view on the one (or two) top priorities when formulating a strategy. The following were the replies:

- |               |  |
|---------------|--|
| Alison Allden | <ul style="list-style-type: none"> <li>• Proofs of concept – determining whether curation is worth it?</li> <li>• Proof of concept as a starting point to develop a strategy.</li> <li>• Test models of re-use models of data curation – citation would come into that.</li> </ul>   |
| Rolf Apweiler | <ul style="list-style-type: none"> <li>• Come up with controlled vocabularies / standards / ontologies. These cannot necessarily be shared across domains. One also needs to consider copyright problems, as some of the information may be from copyright sources.</li> </ul>   |
| Sean Barker   | <ul style="list-style-type: none"> <li>• Inter-operability.</li> <li>• For instance, defining standards to make metadata inter-operable between repositories, allowing inter-disciplinary work.</li> </ul>   |
| Neil Beagrie  | <ul style="list-style-type: none"> <li>• Link data curation with the publishing process.</li> <li>• Proof of concept is needed.</li> <li>• Encourage thinking in the research councils on curation as an issue.</li> <li>• More research library initiatives are needed to guide us.</li> <li>• Stimulate thinking and existing good practice.</li> </ul>            |
| Peter Buneman | <ul style="list-style-type: none"> <li>• Curation of databases/evolving datasets.</li> <li>• Developing models and tools for annotation and provenance.</li> <li>• Database archiving.</li> </ul> <p>(Contribution received after the meeting)</p>   |
| Mikael Dahlin | <ul style="list-style-type: none"> <li>• Standards are needed for how to model processes. Broader models of processes are needed.</li> <li>• Storage – file formats need to be standardised.</li> <li>• Find out how to validate archival processes and information authenticity.</li> <li>• Define the borderlines between different processes – such as</li> </ul> |

libraries, archives.

- Peter Dukes
- Data discovery examples – a quick win?
  - Generic tools and standards.
  - Controlled standards and vocabularies – best not developed in silos.
- Jeremy Frey
- Exemplar projects.
  - Planning, per discipline, from conception through to the end of the data lifecycle.
  - Leading to funding.
- Philip Lord
- We need an ontology for curation! – the day's discussion demonstrates this.
- Liz Lyon
- Try to identify some generic principles which work across domains.
  - Harnessing Grid technologies to produce more cost-effective solutions for data curation.
- Alison Macdonald
- Pressures on institutions – there will be change.
  - Risk analyses are needed.
  - Tools for cost control.
- Bruce Pilsworth
- Intelligent mining of data.
  - Annotation of large datasets to assist future users and the curation process.
- Seamus Ross
- Formal mechanisms for describing functions and behaviour of software so that you can measure performance.
  - Self-contextualising digital entities.
  - Automation of as much as possible of digital preservation activity.

These are not short-term wins.

- Mark Thorley
- Support investment in doing new science by looking at how we can re-use collections.
  - Publishing datasets as a means of placing curation into the research processes.

**N.B.** David Holdsworth had had to leave the meeting before this point.

### **Some suggested areas for research:**

#### **General research areas:**

- Exemplar research projects (see section 4 of report)
- What issues are involved in distributed custodianship and ownership? How will they affect the quality of care of information?
- Investigate quality control as a role within curation - how will peer review affect this, be affected by this? What tools are there?
- What lessons can we learn from the corporate sector about motivation, data management systems, training?

#### **Technical research areas:**

- Research to document systems functionality and behaviour.
- Find methods to reduce the labour involved in curation, including the application of autonomic systems.
- What is an acceptable loss of data? (Data compression, for instance, leads to data loss, and known acceptable data loss levels will play a role in compression decisions)
- Emulation and abstraction, and dependence of emulation on abstraction methods.
- What is the importance of location of technology, and what will be the effect of distribution of data?
- Research into the impact of use of the Grid on storage and distribution of data.
- Anomaly detection in datasets.
- Research into immutable, unique, persistent named entities (such as DOIs), so that data can be located reliably.

## Appendix 4 (To Curation Task Force Report)

### Answers to question (sent and received before the meeting): What a strategy for research data curation should address

- Alison Allden The strategy has to track two aspects – the first in simple terms is the relationship between data, information and knowledge as represented by the process of research, the second is the management of the lifecycle of the data in identifying the key aspects of curation from creation through to preservation and finally destruction. Therefore a data curation strategy has to address creation, maintenance, reuse, and finally the obsolescence of data. At the same time the data curation strategy has to support the dissemination of the research findings and be inter-digitised with the changing conventions of research publication.
- Sean Barker Strategy must first define requirements, cost/benefits (probably in terms of "real options") and risks. Second, it needs to identify existing experience/practice, particular in industry. Third, it needs to identify research issues that focus on highest risks. (There is probably little mileage in looking at cost reduction of storage technologies). Additionally, practice guidelines would probably be useful (e.g. strategies for stopping people keeping every last byte, just in case).
- Jeremy Frey The relationship between electronic data and the physical materials (when they exist) electronic data curation does not exist in a vacuum. With ideas of publication@source more and more of detailed experimental data (not currently normally available to others) could be kept and made available but if it is the responsibility of the research groups/labs/universities thus leads to a very devolved/distributed system which may be hard to control.
- The strategy should address the who, how and where of data curation in the UK.
- David Holdsworth There is a need to know what we have got, be sure that we really have it and can read it, and that we can understand its intellectual content. Nonetheless, we should be wary of setting such restrictive standards for data submission (ingest) that research groups lose interest in submitting their data. I have a strong preference for retaining original byte-streams, but the bulk of high- energy physics data may preclude that. However, it has been my experience that what seemed a lot of data a few years ago is now a trivial amount. My physics background means that I appreciate the data hosepipes of high-energy physics. I suspect that telemetry now permits space scientists to generate similar amounts.

- Philip Lord            Providing a model framework – or frameworks - within which the individuals and funding and organisations can plan, fund, and implement long-term curation easily and cost effectively.
- Liz Lyon                To define the key issues, policies and best practices associated with the cost-effective identification, description, storage and preservation of data, metadata and supporting infrastructure required to ensure the sustainable development of (e-)sciences.
- Mark Thorley            Why the long term management of research data is important; what needs to be done to manage research data effectively, when it needs to be done by and what will be the benefits to the research community; what are the responsibilities of organisations and individuals within the ‘data management chain’ - from the scientists collecting the data to the research funders. What data are to be covered by the strategy? Publicly funded research data are assumed, but are there over sources of data that should be included, and if so, what are the implications of this? For example, the management and exploitation of IPR. Strategy should also include guidance to the research community of what would be expected of them if they were to adopt the strategy. Strategy should not focus on specific technologies, however, it should give some indication of the overall technological framework within which activities should be carried out.

## Appendix 5 (To Curation Task Force Report)

### Responses to the question received before the meeting: “The three most important issues to be addressed in a curation strategy”

#### Alison Alden:

1. The elucidation of why data curation is required and what will be lost if it is not achieved and the recognition this is a demanding and developing research responsibility at individual, research group, institutional, national and international levels.
2. Awareness of curation requirements at start of any programme so that they can be planned and resourced, rather than emerge as an afterthought - including the relationship of data to the dissemination of research outcomes.
3. Series of curation models that can be specified for adoption across the range of research data and activities.
4. (Might add price tag as unspoken fourth most important issue)

#### Sean Barker:

Three topics of most importance (where risk is greatest):

1. Capturing and preserving the meaning of information and of the knowledge embedded in the interpretation of standards
2. Defining a common means of defining the context of information and of indexing against that context
3. Physical preservation of information long term (current industrial requirement is 70+ years)

#### Jeremy Frey:

1. Does the nature of the data (multimedia etc.) influence the nature of curation
2. How to assign a lifetime to data?
3. How to ensure the data can be used (e.g. legacy programs).
4. Business aspects: - How do patent and Health & Safety issues influence curation strategies (legal requirements etc.).
5. Where should the data curation take place?

#### David Holdsworth:

1. Media independence is vital - abstract data format (as per CEDARS) is the way to go.
2. Keep format conversion to a minimum (ideally zero)
3. Do not discard material solely on account of meta-data imperfections
4. Introduce/use a global naming scheme (c.f. CRID, DOI)

5. Have multiple archive stores (a la CEDARS)

**Philip Lord:**

1. Establishing criteria for selecting information for retention, and the purpose of retention.
2. Answering the question of funding the long-term curation of information in the light of its value.
3. Finding solutions to the problem of continued data accessibility in a period of rapid technology change, given that digital information is tightly bound to specific software, and that in turn to specific hardware architectures.

**Liz Lyon:**

1. Criteria for retention
2. Minimum standards for best practice
3. Cost-effective operational model.

**Mark Thorley:**

1. Justification – why should resources be invested, what will be the benefit?
2. Key activities – what needs to be done and when?
3. Reward schemes – why should researchers spend time doing data management – what will be the rewards to them for spending time on what is often currently seen as a non-productive activity (does not count towards career progression or is not seen as valuable by peers).

## Appendix 3: Questionnaires: Detailed Findings

### Methodology

Four sets of questionnaires were administered. The main questionnaire was directed at a broad population of researchers creating data, aimed at establishing the level of curation as practised now and exploring their attitudes to future needs. This is reported in detail in section A below. We asked this group if they would be willing to go on to answer a further set of questions which explored the issues in more depth. These responses are reported in section B. Time constraints limited the number followed up before editing these pages; due to the small numbers in this sample less analysis is presented, but the results show some interesting indicators and comments. Two groups of “service providers” were sampled. The academic librarians are reported in section C, and the views of data centre managers were explored and are reported in section D. Again the responses to these were low, and no detailed analysis was undertaken.

All questionnaires were administered by e-mail, except in the second stage of the data generator questionnaires, when one person opted for a telephone interview.

The four questionnaires and the covering letters are provided at the end of this appendix

### **A: Data Generators – Stage 1**

Amongst the data creators, our sampling aimed to cover as broadly as possible:

- the fields of scientific research
- the range of seniority (from junior post-graduate to eminent professor)
- types of institute (university, specialist institute, other bodies; amongst the universities we aimed to cover Russell Group to new universities)
- England, Wales, Scotland and Northern Ireland.

Names were sourced from Research Council funding lists, and university and institute web pages. All names and e-mail addresses were available in the public domain.

A questionnaire was developed and finalised after feedback was obtained from the steering group. See below for a copy and the covering letter.

A total of 275 questionnaires were sent to this sample using e-mail, to which a total of 48 replies were received (17.5%). Reminders were sent by email to 10%, but generated less than 10% return, evidently provided with less enthusiasm. While we regarded this as a disappointing return rate it is higher than the earlier comparable survey by Lievesley and Jones (13.0%, on a similar-sized population).

In one institution (Institute of Animal Health) replies were coordinated by one individual and a collective view was submitted; this is counted as one reply in the analysis; one individual (at a different site) escaped his net and replied on his own behalf. In another institution one respondent submitted two replies for different projects; these were quite different and are counted as two separate responses in the analysis.

Analysis of the replies against the whole sample showed no bias in response rates from different disciplines (though it was difficult to assign these accurately – and categories were drawn broadly on the basis of the host institutions' type). Small numbers precluded any detailed analysis. We also examined responses by presumed level of seniority, based on the title of the individual. This did not reveal any interesting patterns – Drs were about as likely to respond as plain Mr, Mrs, or Miss, though the percentage of professorial replies was gratifying (over 21%).

Results for each question are presented in turn. Quotations from replies are in *italic*.

A number of respondents offered some general comments on the questions:

- *“Archiving important for primary data - not ours.”*
- *“Different funders place different levels of commitment re archiving. While ESRC tends to require archiving all or nearly all data from projects our other main funders, European Union and Food Standards Agency, leave this to contractors. There clearly is a problem. While most data are kept in the form of summary information in publications more or less permanently, original questionnaires can be kept for only a limited time due to space constraints. Same for original tape recordings of qualitative work. Although raw quantitative data are easy to store electronically I have data files from projects from years ago which are on disks I no longer have a drive for on computers I no longer have access to or are no longer made or the software/operating system changes would make it extremely difficult to access any more. There are also problems that the nature of research work means a lot of short-term researchers over the years and a difficulty for a principal investigator to always keep definitive copies of all data plus backups. Also as PIs move around and collaborate with many people in other organisations it is pretty difficult to go back more than a few years with confidence that data will be adequately archived.”*
- *“I have answered your questions with respect to my scientific research efforts. In my other role as the head of the Institute's IT services I am very interested in providing services to staff to meet the very needs you have identified as leading to this questionnaire. We are implementing online Intranet-based resources to help with the management of data and have raised this area as one for further development over the next few years.”*
- *“Microarray will probably feature quite heavily in this survey. All users of this technology will need similar problems solving and the solutions should fit everyone. BBSRC should lead the development of these solutions and not leave it to individual institutes to develop different systems which cannot communicate easily. With microarray the whole is very much greater than the parts.”*
- *“These issues are especially pertinent re preservation, support, maintenance, domain name ownership of the programme website, after the cessation of ESRC programme funding.”*

The following section presents the findings from the replies, with some commentary.

**Question 1:** Which of your digital data will be of value or use after the end of your project(s)?

	Yes	No
a) Primary data	38 (79%)	10 (21%)
b) Summary / derived data	43 (90%)	5 (10%)
c) Published data	45 (94%)	3 (6%)

The majority of respondents therefore saw continuing value in their data - whether primary derived or published - after project end.

One might expect that the three categories of data might consistently be in ascending order primary to published. Cross-tabulations do indeed show this pattern, but two responses indicated primary data being of value and derived not and one where primary data was of value and the derived and published data was not. Then again another two saw value in derived data but not in the published. There was no indication of why this might be.

**Question 2:** Do the terms of your funding require data to be archived/preserved?

Yes	No	Don't know	In some cases
22 (46%)	15 (31%)	7 (14%)	4 (8%)

This question was examined in relation to question 1. In 38 cases respondents had indicated to question 1 that their primary data was of value and 20 of these (53% of this group) said their terms of funding required data retention. Only 2 of the 10 saying they did not see value in their primary data (Question 1) said were under a contractual obligation to keep data.

**Question 3: Who will look after project data after project end?**

This question elicited a wide variety of responses from 42 respondents, which were categorised as follows:

Response	No.	%
Self	11	23%
Don't know	9	19%
PI, project leader or supervisor	7	15%
Their institution	6	12%
"Colleagues"	3	6%
No one	2	4%
ESRC Archive (UKDA?)	2	4%
Successor	1	2%
"per regulations"	1	2%
No answer	6	12%

Clearly there is very little formal provision, or awareness of provision, for data retention judging by these answers.

**Question 4: Has financial provision been made for keeping project data after project end?**

Yes	No	Don't know	No answer
10 (21%)	27 (56%)	10 (21%)	1 (2%)

Thus, though these respondents in general see continuing value in their data, only 21% report definitely that financial provision has been made to look after it after project closure.

Looking in more detail reveals that where answers to question 2 show that data should be kept by contract (22 instances), only 8 (36%) of these say they have funding to do it, and 10 (45%) say they do not. (Three said they did not know, and one did not answer.). These are only very small numbers, but if they represent in any way the true situation there is a somewhat disturbing mismatch between directives and resources to achieve it. This was confirmed during some interviews.

**Question 5: Who owns the data you are generating during the project? Please specify:**

These answers were categorised as follows:

<b>Ownership</b>	<b>No.</b>	<b>%</b>
Their institution/employer	23	48%
Institution plus funding organisations	9	19%
Don't know	6	12%
Funding organisation	5	10%
Self	4	8%
"Public domain"	1	2%

There appears to be fairly good recognition of ownership in this population.

**Question 6: Have you been provided with any policies or guidelines regarding:**

	<b>Yes</b>	<b>No</b>	<b>Don't know</b>
a) Data preservation	21 (44%)	23 (48%)	4 (8%)
b) Records management	16 (33%)	26 (54%)	6 (12%)
c) Good data management	21 (44%)	24 (50%)	3 (6%)

Somewhat less than half these respondents had received any guidance on good data, records management and preservation. Cross tabulating these answers shows a high degree of correlation between them: provision of one is accompanied by provision of the other and vice versa; we might reasonably assume that generally they are all referring to the same documentation or training in the case of the positive replies. Two responses from the same department were corroborative.

Provision of guidance materials shows some interesting points in relation to some of the other questions. Question 2 explored the contractual obligations to keep data and comparing those replies with provision of guidance on preservation shows that where there is an obligation to keep data (22 cases) some 32% (7) of these report having received no guidance on preservation and only 64% (14) had. There was one don't know. Again one needs to interpret this cautiously in view of the low numbers.

Of the 21 respondents who reported receiving guidance, they quoted that their sources of advice as follows:

Source	No.
Their institution	12
IT department and advisors	7
Research Council	7
Funding organisation(s)	3
UKAS	2
"Training"	1
QA department	1

Some respondents quoted more than one source of information. One comment:

*"we found that the document was really a set of motherhood statements and that it did not help us to solve the practical and financial problems of long-term data archiving."*

#### Question 7:

(a) Does your project data contain confidential information?

Yes	No
27 (56%)	21 (44%)

(b) Will that data remain confidential after the project?

Yes	No
20 (42%)	28 (58%)

These data show no anomalies. Cross tabulations show 5 respondents indicating that after project end their data ceased to remain confidential. It is not clear what the reason is for this declassification.

(c) Are there any other conditions which you feel should be imposed on, or capabilities enabled, for the data after project end?

Yes	No
12 (25%)	36 (75%)

If "Yes", please specify:

Some of the remarks made were as follows:

- *“This largely applies to qualitative data in my experience, which requires context to be available in order to fully appreciate its meaning and where revelation of this context could compromise the interviewees' anonymity.”*
- *“Third-party data mining should acknowledge.”*
- *“Storage in searchable database. Some projects confidential, others not.”*
- *“Probably Clinical Governance rules.”*
- *“Patenting may be appropriate.”*
- *“Only access for academic purposes.”*
- *“How the data were derived and for what purpose.”*
- *“Data may need to remain confidential after project for patent issues. Conditions should be imposed for public release of data.”*
- *“Appropriate recognition given to originators. Data may need to be “visible” but not “captureable” without permission.”*
- *“7a: Yes but: Some projects funded by industry will be confidential. Data from other projects will be confidential until the IP position has been determined. 7b: Some data for projects will remain commercially in confidence. However, the basic principal is to publish all data from publicly funded projects and from industry funded projects where possible. 7c: Data should be actively assessed by a Project Team and archived by the Project Manager.”*

**Question 8: Will future users need any of the following to use the data?**

	Yes	No
a) Special software	20 (42%)	28 (58%)
b) Special hardware/instrumentation	7 (15%)	41 (85%)
c) Explanatory documentation	36 (75%)	12 (25%)

A disturbingly high number of respondent’s report using special software whose longevity may be questionable.

**Other (please specify):**

Some of the notes made are as follows:

- *“Yes to all of the above for certain types of data.”*
- *“We use European Data Format EDF for patient data which is standard.”*
- *“The special software is not essential but it is highly desirable.”*
- *“Specific yes, Special no.”*
- *“Some data yes, most no.”*
- *“Software: probably, but unlikely to be of much use dependent on time since data produced. H/w/instrumentation: Probably but unlikely to be available again dependent on time since data produced. Other: a mechanism to restore archived data and re-archive in the new format to maintain its usefulness.”*
- *“Re special software: some aspects of the processed data would be far easier to access with specialised software, but most data is in standard formats.”*
- *“My belief is that the material archived should be self explanatory and accessible.”*
- *“Explanatory documentation, because nearly all qualitative data needs to be placed in context.”*
- *“End users will not require these at their own sites, but it will be necessary either to maintain the database server and database management software at the site where the data are held, or to arrange export of the data in a "universal" format when the project comes to an end. Much of the value of the data would thereby be lost because it is currently maintained in an object-oriented database with sophisticated linkages.”*

### Question 9:

**(a) Could you please confirm the start and end dates of your current research project(s) (mm/yyyy):**

The aim of this question was twofold: (1) is much data being collected which may be subject to preservation problems during project life due to the length of the project? and (2) to gauge the turn-over of new, completed datasets from projects. Thirty one respondents provided figures. The distribution was as follows (note many respondents quoted for the multiple projects they were engaged in and so these data refer to number of projects, not to respondents):

<b>Period</b>	<b>No.</b>	<b>%</b>
2 years	7	(17%)
3 years	18	(45%)
4 years	3	(5%)
5 years	4	(10%)
> 5 years	5	(12%)
Indefinite duration	4	(10%)

Only a minority of projects span time periods over which preservation issues may arise during the project itself. The data “turnover” rate would seem to be centred on three years.

**(b) Are you collaborating with other institutions on your project(s)? :**

Yes	No	Sometimes	No answer
38 (79%)	7 (15%)	2 (4%)	1 (2%)

Overwhelmingly respondents were working collaboratively, but the questionnaire could not probe the details of this collaboration (whether intra-institution, across institutions and industry, or internationally). We also noted that several respondents who had received funding for different projects from different Research Councils and funders.

### Question 10:

**(a) What are the main types of specialist commercial or open source software you are using, if any (e.g. Maple, TurboChrome, etc)? :**

This elicited a surprising variety of software systems – from this sample of 29 replies some 49 systems were mentioned, most of them just once. (This counts a reply such as “Microsoft Office” as one system.) Of those mentioned more than once, MatLab was mentioned 7 times and Fortran (sic) and MS-Office both 4 times; a further nine were mentioned twice.

**(b) Are you using software you have written for your project?**

Yes	No
22 (46%)	26 (54%)

This recalls question 8a, whether special software is in use. It is not clear how this self-generated software is to be preserved if need be, and a nearly 50% response indicates that the potential problem may be of considerable size.

**(c) What are the principal data formats for the data you are generating (e.g. XML, ASCII, TIFF, etc)? :**

As might be expected from the answers to question 10a there were many different responses to this question – 25 different formats were mentioned by 34 people. 6 said they did not know, 4 merely said “many” and 4 entries were blank. Again, respondents gave multiple answers.

There was rather more agreement on formats than for actual software; formats mentioned more than once were:

<b>Format</b>	<b>No.</b>
ASCII	20
TIFF	11
Excel (.xls)	9
Access (.mdb)	3
HTML	3
FITs	2
XML	2

**Question 11:**

(a) Roughly how much data are you generating? :

Is this per month or per year? :

The replies from 27 respondents were aggregated to produce the following logarithmic distribution on an annual basis:

<b>Volume per annum</b>	<b>No.</b>	
Terabytes	2	(an astronomer and oceanographer)
Gigabytes	15	
Megabytes	8	
Kilobytes	2	

Three people said they did not know, ten replies were blank or could not yield a useful interpretation. One respondent said “Lots” and one respondent reported not generating any data at all.

(b) Is your data static [ ] or dynamic [ ]?

Static	Dynamic	Both	No answer
21 (44%)	15 (31%)	2 (4%)	10 (21%)

(c) Where and on what is your data kept? :

Where data is kept was another question eliciting an uncomfortably varied response. Significantly all responses bar three replied in terms of storage media; the exceptions were three responses of “Archive” (unspecified) and one of NAS (Network Attached Storage). Four respondents referred to paper: “on hard copy” (3) and laboratory notebooks/logs (1). Two respondents said “database” – it was not clear what this implied. The following table lists the frequencies of various replies (again, multiple answers were received from some individuals):

Medium	No.
Local hard disk	22
Server / Network	18
CD R/RW	13
Media (various, excluding CD-R/RW) <sup>3</sup>	8
“Back-ups”	8
“Archive”	3
Hard copy	4
Database	2
NAS	1

There was one don’t know, and two did not answer.

These replies indicate a lack of awareness in practice of the problems of keeping data over the longer term.

---

<sup>3</sup> Media mentioned were DLT, DAT and Exabyte tapes; “tapes”; DVD-R; floppy discs (twice); ZIP drives; MO disks (Magneto-Optical). Note: the reply “Zip” format and could have referred to the “archive” compression format.

**Question 12: Is there any data which cannot be recreated if lost? :**

<b>Yes</b>	<b>No</b>	<b>Maybe</b>	<b>No answer</b>
27 (56%)	13 (27%)	1 (2%)	7 (14%)

**Additional question: Would you be willing to participate in a second stage of the survey?**

26 (54%) of the respondents at this stage elected to answer a further set of questions – see below.

## B: Data Generators – Stage 2

Of the 26 people from the data generator stage 1 survey who said they were willing to answer further questions, most electing to do so by e-mail. At time of writing 15 of these had been conducted, with replies received from 10. We present the results below, using the format of the second questionnaire to present the numbers. As above, quotations from replies are in italic.

**Question 1: Can you characterize the value of your data after the end of your project?**

Further scientific value: [9]  
 Potential commercial value: [4]  
 Evidential value, to  
 confirm conclusions drawn: [9]  
 Historical value: [2]  
 Other (please specify):

- *“Taxonomic convention requires that 'type' specimens of newly described species are curated in perpetuity. My work will inevitably result in 'virtual' type specimens of this sort.”*
- *“Additional data for patent development, licensing etc.”*
- *“Note - all these are possible but not necessarily for all data”*

**Question 2: Do you think new science can be built on primary data you have generated?**

Yes [9] No [1]

**Question 3: Do you believe that continuing funding to keep the data you are generating would be:**

Justified? Yes [10] No [0]  
 Would prove a good investment? Yes [8] No [2]

**Who/ what body do you think should take care of the data after project end?**

- *“Central archives - many already in place (e.g. for space missions and ground-based telescopes)”*
- *“A traditional museum, preferably the museum in Oxford, should certainly hold responsibility for the data - where it is actually stored and who actually 'curates' it at any one time would seem unimportant, as long as a long-term museum is maintaining (a) interest in it, and (b) a portal to it.”*
- *“The Institute's Data Manager.”*
- *“University Library, sponsor companies.”*

- *“The Institute. Also need to keep tissue (plasma) samples - good investment for both data & tissue. Her experience is that organisations do NOT like funding this kind of thing - they don't want an endless drain of cash. [answers given by phone - see paper form].”*
- *“Probably SCRI or its commercial arm, MRS Ltd, who own the IP.”*
- *“Principal Investigator.”*
- *“Note - for Qs 2 and 3 this will be dependent on the specific data. Re Q3(2): [who take care] This will change with time. The generator has a responsibility to make sure the data is stored appropriately but the Institution has an obligation to provide the facilities. The onus probably shifts from the generator to the Institution over time.”*

**Question 4: Effective preservation of data depends on providing good-quality description (context, technical, indexing). Much of this is best provided by the data originator at the time it is created.**

**For your data do you:**

Feel you have enough time and Funding to do this?	Yes [3]	No [7]
Think training about this would be valuable?	Yes [7]	No [3]

**Question 5: would you be willing to see your data kept in:**

A repository managed by your institution's

a) Library:	Yes [5]	No [5]
b) Data centre/IT group:	Yes [9]	No [1]
c) Other:		

- *“I don't think it matters too much where it is physically held, as long as it is available through a museum.”*
- *“!Library does not have the right skills. Other: BioSc & Stats Group in Scotland. There will be reluctance from scientists unless linked with publication. They will use only if there is a measure of control, e.g. link to publication.”*

A national repository:	Yes [7]	No [3]
An international repository:	Yes [6]	No [4]
Should this be a general facility? or discipline-specific?	[6]	[3]

**Note:** One person responded to both options; one person did not answer.

**Question 6: If your project data contains confidential information, what is the nature of this confidentiality:**

Data on individuals(e.g. medical):	Yes [4]	No [6]
Commercial secrets:	Yes [3]	No [7]
Patentable information:	Yes [2]	No [8]
Information, disclosure of which might compromise this project or future projects for you:	Yes [7]	No [3]

The following comments were received to this question:

- *[re Compromise] “Potentially in a few cases.”*
- *re Q6(4) (compromise – telephone interview): “possibly”: Cited cases where the press had got hold of data before peer review/publication.*
- *“Note - All of these will apply to some but not necessarily all data generated by the Institute.”*

**Question 7 a): Do you mine or re-use your own old primary data?**

Yes [3] No [3]

**b): Have other people used these data?**

Yes [3] No [6] Don't Know [1]

**Question 8: Have you ever experienced difficulty locating or using other people's data:**

Yes [5] No [5]

If "Yes", was this caused by:

a) Inadequate indexing or descriptive information?

Yes [4] No [1]

b) Lack of access to software to read it?

Yes [4] No [1]

c) Other:

- *“Confidentiality.”*
- *“Lack of standards; no SPSS licence and too expensive.”*
- *“Poor storage conditions (hard copies.)”*

**Question 9:** Do you anticipate using or conducting "collection-based" science in the future (where discovery is made by investigating existing data rather than generating significant new primary data)?

Yes [8] No [2]

**Question 10:** Do you apply standard vocabularies in your work:

Yes [1] No [9]

What are they:

- *"Not sure what you mean here - but the data I generate contains no written language other than file titles, so I am fairly sure the answer is no."*
- *"Ecological and entomological terms."*

**Question 11:** Do you feel there are sufficient tools and technologies available to ensure long-term accessibility to your specific data:

Yes [4] No [4] No answer [2]

**Question 12:** Are you aware of the emerging Research Grid technologies which allow vast access to shared and distributed computing resources and vastly increased storage?

Yes [5] No [5]

If "Yes": What benefits, if any, do you do you believe it will yield for you:

- *"Faster access + better natural language queries spanning multiple databases."*
- *"Probably few or none. The volume of data with which I deal is within the capabilities of web/internet technologies."*
- *"Very hard to tell at present."*
- *"Throughput. More data analysis per unit time."*
- *"Re q11: don't really know, lack of understanding of issues."*
- *"Easier access and exchange of information in different formats."*
- *"Not convinced that it will deliver much more than at present."*

**OTHER COMMENTS:**

- *“Insufficient storage at X for older, hard copy data. Tends to get forgotten or damaged by poor storage conditions. Electronic data is back up by IT.”*
- *“Re Qs 7 & 8 - don't know [Institute response], but q9 is going to be increasingly important.”*

## C: Librarians

We aimed to cover a range of university and institute librarians throughout the UK. Forty seven questionnaires were sent by e-mail, and nine replies were received (19.1%). The replies are summarised in the following with some commentary. (Note: Inconsistencies in numbering in this questionnaire are retained here – 1 d) and 7) (numbers omitted).

**Question 1: a) Has the proportion of electronic holdings (of all types) in your library increased in the last two years?**

Yes: [9] No: [0]

**b) If yes, has this resulted in an increase or decrease in:**

**- the number of user queries?**

Increase: [9] Decrease: [0]

**- the time spent on user queries?**

Increase: [9] Decrease: [0]

**If increases, do you think this is a transitional phase?**

Yes: [3] No: [6]

The following comments were made to this question:

- *“What sort of user queries? The number of user queries in general has increased but it isn't possible wholly to determine the cause - i.e. increases in overall business and in student numbers coinciding with increase in e-journals. Therefore it isn't easy to tell if this is transitional. Probable, but tentative, answers.”*
- *“Unable to measure properly. Anecdotally it feels as if traditional paper-based queries have been replaced by the same number of queries relating to e-materials.”*

**c) Does your library have its own IT staff?**

Yes: [8] No: [1]

**e) Is your library responsible for information services throughout your institution?**

Yes: [6] No: [3]

**Comments:**

The following remarks were made:

- *“If you mean are we converged, the answer is no. However, we would say that the Library is responsible for information services, while the Computing Service is responsible for information systems. As they once said: the Library is the 'I' and the Computing Service is the 'T' in 'IT', which though over simplified is more or less how we work.”*
- *“Library is part of an Information Services Dept.”*

- *“Information Services is an integrated service department responsible for computing, library and media services.”*
- *“There is a small number of small, specialist libraries which are staffed separately from the University Library.”*
- *“[We] work in partnership with University Computing Services.”*

**Question 2:** Are you considering, or do you already share specific electronic collections, journals with other institutions?

Yes: [6] No: [3]

If "Yes", have the participants come together for geographic reasons, for subject area reasons, or other?

- *“A number of electronic resources are shared by partners within the Institute.”*
- *“We are at the beginning of considering whether this would be a possibility, but as yet no further forward. It would be a) via our purchasing consortium which started as a regional consortium and grew; or b) for our joint medical school, and therefore subject and geographic reasons.”*
- *“Purchasing consortia.”*
- *“Yes to both, plus organisational reasons.”*
- *“Yes for both reasons.”*

**Question 3:** a) Who is responsible for managing digital security in your library?

The following were noted:

- *“College Network Supervisor looks after digital security for the college - we do not have anyone within Learning Resources who looks after this area.”*
- *“Various staff members.”*
- *“Our library management systems librarian maintains data security for the LMS. Our electronic service librarian manages our web site, through which access to the electronic journals to which we subscribe is provided, and she manages our Athens registration with Computing Service. Our serials librarian manages the licences for subscription databases and e-journals. But we own no digitised records (see 4 below).”*
- *“No-one.”*
- *“Systems and Security team.”*
- *“The Library Systems Team in liaison with the University Computing Service.”*

- *“Computer Centre.”*
- *“The Collections Management Team, together with University Computing Services.”*
- *“I.T. department.”*

**b) Does that role also cover maintaining integrity of the digital records?**

Yes: [2] No: [5] No answer: [2]

**c) Who is responsible for digital preservation in your library?**

Eight people responded:

- *“Ultimately, the Learning Resources Manager although we are not currently active in this area.”*
- *“Electronic Collections Librarian.”*
- *“No-one.”*
- *“Collection and Access Management Section, Information Services.”*
- *“No one. We do not, as yet, have a clearly articulated policy in this area.”*
- *“Systems team.”*
- *“Archive project managers, Library archivist, University Computing Services.”*
- *“The Keeper of Special Collections.”*

**Question 4: a) Do you have digital holdings other than electronic journals and other commercially available electronic materials?**

Yes: [7] No: [2] Don't know: [0]

**b) If "Yes", do they pose any particular accession or management problems?**

Three answers were given:

- *“Not at the moment.”*
- *“Yes - insufficient funds available.”*
- *“Re X - dedicated member of staff to manage collection. With other collections the Library has no control over the content & in some instances important documents have been removed with[out] our prior notification.”*

**c) Do you accept digital materials in any software format?**

Yes: [4] No: [3] No answer [2]

One of the “No” replies added “Not yet”.

**Question 5:** a) Does your library have any holdings or archives of primary research data?

Yes: [7] No: [2]

b) If so, are any of them in digital form?

Yes: [5] No: [2]

c) Do you receive special funding to support these holdings (whether paper or digital)?

Yes: [3] No: [6]

d) Are staff or researchers required to deposit digital copies of research papers and publications with your library?

Yes: [1] No: [8]

The “Yes” response added that it was a pilot scheme; one of the “No” responses added “not yet”.

e) Are submitters required to provide descriptive information about the digital materials they are lodging?

Yes: [2] No: [3] Not applicable: [4]

f) Are submitters provided with any guidelines for submission of digital materials?

Yes: [3] No: [2] Don't know: [4]

**Question 6:** a) Does your library specialize in any area(s)? If so, which?

Yes: [5] No: [3] Not applicable: [1]

Four responses cited collections in the humanities and the fifth merely said “many”.

b) Do any of your specialist collections have their own curator(s)?

Yes: [4] No: [5]

c) Has curation work relating to your specialist collections generated new research?

Yes: [5] No: [4]

d) Do any special collections now include digital items?

Yes: [5] No: [3]

**Note:** One response gave no answer to this question, and omitted questions 6 e) to 6 g) following.

e) Do these collections attract specific funding?

Yes: [4] No: [2] Not applicable: [2]

f) Do they generate, directly or indirectly, revenue for your institution?

Yes: [4] No: [2] Not applicable: [2]

g) Have you received any bequests of digital materials?

Yes: [3] No: [5]

[Note – 7 omitted from numbering]

**Question 8:** Digital technology offers the possibility to link annotation to records. Do you think this would be a valuable extension to the knowledge base (whether discipline-specific, interdisciplinary or wider)?

Yes: [8] No: [1]

Is it practical?

Yes: [6] No: [3]

One of the “No” replies added “not yet”.

Who/which body/bodies do you think would be most appropriate to provide/oversee these services?

Four respondents passed on this question.

- *“No easy answer as it depends who holds/owns the data, which is obviously distributed.”*
- *“JISC”*
- *“The Open Archives Initiative has been exploring solutions for citation impact analysis and reference linking in large-scale OAI open-access archives. This work may provide pointers to a solution.”*
- *“Services or standards? I'm assuming that this is best done as a function of libraries in terms of a service. That's certainly how we're experimenting in x with projects on collection level descriptions. We have found that no matter how good the metadata standards, they are only as good as their application (almost self-evident!). We then find benefit in having a central institutional unit ensuring consistency (aka the cataloguing department!). Standards setting is perhaps best done by JISC. But that's the easy part as I'm not sure if you mean that kind of metadata annotation or something closer to a threaded discussion, or simply some description by the researcher depositing the data.”*

- *“Information Science/Library professionals.”*

**Question 9:** Has digital technology resulted in the increased creation of islands of information and information resources within departments? If so, should catalogue information be coordinated with/by your library?

Yes: [4] No: [3]

Two respondents answered as follows:

- *“Ideally, in order [to] provide the widest access to an institutional resource; providing resource to do this would be more problematic.”*
- *“Islands of information and information resources have always tended to exist within organisations. I am unsure as to whether digital technology has increased this tendency. However, digital technology certainly does provide the potential to link these islands and the institutional 'library' would have a role in this. This issue raises the extent to which 'libraries' increasingly include management of information resources generated WITHIN their institutions in their remit - as well as acquiring and managing information resources created outside the organisation and then making them available (i.e. libraries' traditional remit).”*

Two of those answering “Yes” added:

- *“Absolutely and standards set and policed.”*
- *“Yes - in an ideal world.”*

**Question 10:** a) Have you been made aware of developments involving the Grid?

Yes: [8] No: [1]

b) If so, do you think the Grid will bring structural change, benefits or entail any particular difficulties for your department?

Yes: [2] No: [3] Don't know: [4]

**Question 11:** a) Does your institution publish any academic journals?

Yes: [5] No: [4]

b) Do you contribute to your institution's policy with regard to journals?

Yes: [5] No: [2] No answer: [1]

The other respondent said they did not understand the question.

**Question 12:** what proportion of your costs does computer storage (hardware, media) represent?

One respondent said less than 5%, and two said 10% or less. The others were not sure or could not answer.

a) Are there any research programmes relating to digital information services or digital libraries which you would like to see?

Three respondents offered possibilities:

- *“I would be more concerned that the outcomes of existing research programmes are better known, and that the issue of how the long-term preservation of digital data is resourced receives more attention.”*
- *“YES a big push on OAI [= Open Archives Initiative]”*
- *“A national electronic library initiative that is adequately funded and is not 'just' another research programme.”*

b) Are you involved in any digital library initiatives or research programmes?

Yes: [6] No: [1] Not at present: [1]

There was one blank reply. Three people amplified their answers as follows:

- *“Yes, a project for institutional archiving, about to start. This project is under the JISC FAIR initiative: project SHERPA, a consortium project between CURL and White Rose University Libraries, the BL and MIMAS.”*
- *“JISC-funded FAIR project on e-prints.”*
- *“Yes we have a 16-person Centre for Digital Library Research up to its neck in this stuff.”*

Please add any further comments you may have:

- *“I am unclear as to how this study relates to JISC's other digitisation initiatives - for example, the digitisation strategy as recently published.”*
- *“This is a co-ordinated response on behalf of Subject Librarians and the Keeper of Special Collections.”*

## D: Data Centre Questionnaire

Fifteen questionnaires were sent to a range of heads of university, institute and departmental data centres, and five responses were received (33.3%).

The responses from these for are summarised below in-line using the text of the questionnaire itself in bold **text**. The five responders are denoted by the letters A to E where appropriate.

### Question 1: What services are you/is your department responsible for:

End-user support:	Yes [5]	No [0]
Systems management:	Yes [5]	No [0]
Network support:	Yes [5]	No [0]
IT security (e.g. firewalls):	Yes [5]	No [0]
Access controls/account management:	Yes [5]	No [0]
Programming assistance or services:	Yes [3]	No [2]
Other (Please specify):		

E: “*all centrally managed information systems services (eg. Telephony, data networking, e-learning platforms, web publication tools, electronic data storage & management, computationally intensive facilities, ....)*”

### Question 2: Approximately how much storage is on systems under your management?

On-line: See the table at the end of this section relating staffing, budgets and data volumes  
Off-line: idem

### Question 3: Do you think that new Grid technologies will affect your department's responsibilities?

Yes [3] No [2]

If "yes", how:

C: “*We already run two major White Rose Grid systems at x providing both systems and user support, file backup and management, and collaboration with University researchers. We also work with regional grid support personnel in developing the scope of the White Rose Grid.*”

E: “*We are already a National Level 1 and 2 facility, and have a 500+ node computational resource*”

**Question 4: Do you provide any of the following facilities to help users to preserve their data over the long term (>5 years):**

- |  |   |         |        |
|--|---|---------|--------|
|  | Long-term storage for media:                              | Yes [3] | No [1] |
| *  | Regular checks on media integrity:                        | Yes [1] | No [4] |
|  | Migration of data from old media to a new:                | Yes [3] | No [1] |
|  | Services to facilitate data format migrations:            | Yes [2] | No [3] |
| **   | Monitor and record hardware and software used:            | Yes [2] | No [2] |
|  | Provide advice about data preservation:                   | Yes [4] | No [1] |
| For systems to be de-commissioned do you - |   |         |        |
| **   | Have policies to save data off these systems:             | Yes [4] | No [1] |
|  | Retain specifications of these systems:                   | Yes [1] | No [4] |
|  | Provide other services for preservation (please specify): |         |        |

E: \*: [re long-term storage for media] "but probably <10 years"

\*\* "only central systems"

C: "The Computing Service at Leeds University has a long tradition of providing archiving facilities, enabling data from previous legacy systems to continue to be available. X is also involved in the CEDARS and CAMiLEON projects (as noted in the questions below)."

**Question 5: Do you have policies which govern the retention of data when de-commissioning systems?**

Yes [2] No [3]

C: "No general policy covering all data, but because of our heritage with archiving systems (see above), a lot of data is placed into the Archiver."

E: "[No] only rather ad hoc, sadly"

**Question 6: Which of these national and international initiatives on digital preservation are you aware of:**

The Data Preservation Coalition (DPC) in the UK?	Yes [1]	No [3]
Digital Library initiatives, such as D-Space?	Yes [4]	No [1]
Open Archival Information System (OAIS)?	Yes [3]	No [2]
CEDARS and CAMiLEON (at Leeds and Michigan)	Yes [1]	No [4]
Others (Please specify):	None	

**Question 7: Do you believe that data centres have a role in tackling the digital preservation problem?**

Yes [5] No [0]

If so, how?

A: "Integrity and preservation."

B: "Providing resilient storage."

C: *“By providing the underlying infrastructure for institutional data management including long-term storage using technologies such as SANs and/or HSM systems.”*

D: *“Informing on good practice and offering services to those that do not have the necessary facilities.”*

**Question 8: Do you provide guidance for users on the following:**

Good data management: Yes [4] No [1] Don't know [0]

Records management: Yes [0] No [4] Don't know [1]

Data preservation: Yes [2] No [3] Don't know [0]

**Question 9: Is data backed up from all machines regularly (including desk-top systems)?**

Yes [3] No [2]

C: *“University data management guidelines discourage holding data ONLY on desktop systems - secure, networked, backed-up server storage is advised.”*

E: *“only centrally managed (ISS) systems, plus those dpts who ‘contract’ with us to do so for them”*

**Question 10: Is your group responsible for backing up:**

\* Local desktop machines? Yes [2] No [3]

Local servers? Yes [4] No [1]

Central servers? Yes [4] No [1]

\* E: *for those we manage*

**What media are used?**

A: *DLT and LTO*

B: *DAT, DLT*

C: *LTO, DLT, DAT, Exabyte*

D: *Tape*

E: *AIT Tape, plus mirrored and “broken mirror” disks*

**Question 11: How much data is backed up each day?**

See the table at the end of this section relating staffing, budgets and data volumes.

**Question 12: For how long is backed-up data kept?**

- A: 1 year  
 B: 3 to 4 months  
 C: “Varies - mainly either a month (backup cycle tapes), a year (take out tapes) or indefinitely (take out tapes or archived).”  
 D: “Backups exist to the earliest use of centralized backup at the Institute. It is intended to continue to keep long term backups at yearly intervals for the foreseeable future.”  
 E: “3 months unless ‘archived’ upon request (pls note ‘archiving’ is NOT the same as backup).”

**Question 13: How would you describe your department:**

Central data centre/IT group? Yes [4] No [1]  
 Local data centre/IT group? Yes [0] No [2]  
 Other (please specify)? **None**

**Question 14: Is your group in the same department as the library (or information services) group?**

Yes [3] No [2]

- C: “Although, close collaboration with the Library and media Services under the umbrella title of ‘Academic Services’.”  
 E: “As part of a ‘federation’ of “Academic Services” where constituent units remain professionally managed separately but in a spirit of partnership.”

**Question 15: Number of staff (full-time and part-time):**

See the table at the end of this section relating staffing, budgets and data volumes.

**Question 16: What is the group's annual budget?:**

See the table at the end of this section relating staffing, budgets and data volumes.

**Question 17: Does this budget include costs for:**

Staff? Yes [4] No [1]  
 Media for backups? Yes [5] No [0]  
 Purchase of hardware? Yes [4] No [1]  
 Purchase of software licenses? Yes [5] No [0]

Table of comparative answers to questions 2, 11, 15 and 16.

Question	Responses					
	A	B	C	D	E	
2. Total system storage	On-line	about 1Tb	Est. 5Tb	6 Tb	0.5TB	12 Tb
	Off-line	- *	Est. > 5Tb	286 Tb	1Tb	30+ Tb
11: Back-up volume per day	Ca. 550Gb	100 Gb	Varies. (See note 1)	Varies. (See note 2)	~1Tb	
15: Number of staff	50	60	200	10	140 Full time 10 part-time	
16: Annual Budget	- *	£3.0m	£9.7m	~£0.2m	£8.1M	

Note 1. *“Depending on backup cycles (full vs incremental etc). Not known exactly owing to historic multiplicity of varying backup schemes. (Hopefully, to be consolidated with centralised backup system to be procured as part of SAN Project).”*

2. *“Varies from day to day: full backups of the order of 200GB occur a number of days per month and incrementals of the order of 20GB on other days.”*

\* No answer.

## Appendix 4: Questionnaires and covering letters

### Data generator stage 1 questionnaire and cover letter

Dear \_\_\_\_\_,

We are writing to you on behalf of Professor Tony Hey, Chairman of the JISC Committee for the Support of Research and Director of the UK's e-Science Core Programme, who has asked us to investigate provision for the long-term care and preservation ('curation') of the data being generated in research in the UK.

The increasing power of computing and its collaborative capabilities will have significant implications for the curation of research data: We need to be able to ensure that data can continue to be accessed and re-used over time, that we can validate our research, and that our research can contribute to dynamic knowledge bases and future research. The potential for loss and waste is immense, but so too are the opportunities.

We realize how heavy the demands on your time are, but given your important role and the importance of the issue we would be very grateful if you would complete the attached questionnaire and return it to pwl@d-archive.co.uk by Thursday, 12th December 2002. All replies will be treated in strict confidentiality. By way of thanks, and to reflect the theme, all returned questionnaires will go into a draw for a bottle of vintage champagne (or other, if the winner is not a champagne person), to reach the winner for the festive season.

The questionnaire is attached as a Word form and also below within this e-mail, whichever you prefer. If you use the Word form, please save it as "Form". For those involved in several research projects, you may prefer to reply with regard to one or two of your research projects.

If you have any other questions, please do not hesitate to contact Philip Lord (telephone 0208-607 9102) or Alison Macdonald (0208-744 9322), or e-mail us at pwl@d-archive.co.uk. We will be happy to answer any questions you have.

Your input is important and much appreciated.

Philip Lord

Digital Archiving Consultancy

2 Wayside Court

TWICKENHAM TW1 2BQ

Tel: 0208-607 9102

Fax: 07050-675 010

=====

#### DATA CURATION QUESTIONNAIRE

Please mark appropriate boxes with an x, thus [x]

#### THE LONG-TERM USE AND VALUE OF YOUR DIGITAL DATA

1. Which of your digital data will be of value or use after the end of your project(s)?

Primary/raw data:                    Yes [ ] No [ ]  
 Summary/derived data:                Yes [ ] No [ ]  
 Published data/publications:        Yes [ ] No [ ]

#### DATA RESPONSIBILITIES

2. Do the terms of your funding require data to be archived/preserved?  
 Yes [ ] No [ ] Don't know [ ] In some cases (specify):

3. Who will look after project data after project end? :

4. Has financial provision been made for keeping project data after project end?

Yes [ ] No [ ] Don't know [ ]

5. Who owns the data you are generating during the project? Please specify:

6. Have you been provided with any policies or guidelines regarding:  
 Data preservation:    Yes [ ] No [ ] Don't know [ ]  
 Records management:    Yes [ ] No [ ] Don't know [ ]  
 Good data management: Yes [ ] No [ ] Don't know [ ]  
 By whom? :

7. (a) Does your project data contain confidential information?

Yes [ ] No [ ]

(b) Will that data remain confidential after the project?

Yes [ ] No [ ]

(c) Are there any other conditions which you feel should be imposed on, or capabilities enabled, for the data after project end?

Yes [ ] No [ ] If "Yes", please specify:

#### READING THE DATA IN THE FUTURE

8. Will future users need any of the following to use the data?

Special software                        Yes [ ] No [ ]  
 Special hardware/instrumentation    Yes [ ] No [ ]  
 Explanatory documentation            Yes [ ] No [ ]  
 Other (please specify):

#### BACKGROUND AND TECHNICAL INFORMATION

9. (a) Could you please confirm the start and end dates of your current research project(s) (mm/yyyy):  
(b) Are you collaborating with other institutions on your project(s)? :
10. (a) What are the main types of specialist commercial or open source software you are using, if any (e.g. Maple, TurboChrome, etc)?:  
(b) Are you using software you have written for your project?  
Yes [ ] No [ ]  
(c) What are the principal data formats for the data you are generating (e.g. XML, ASCII, TIFF, etc)? :
11. (a) Roughly how much data are you generating? :  
Is this per month or per year? :  
(b) Is your data static [ ] or dynamic [ ]?  
(c) Where and on what is your data kept? :
12. Is there any data which cannot be recreated if lost? :

#### GENERAL INFORMATION

Please confirm your name (first name, surname):  
Department, institution:  
Project name(s) (Abbreviated titles are sufficient):  
Please add any further comments you may have:

#### NEXT STEPS

Would you be willing to participate in a second stage of the survey?

Yes [ ] No [ ]

If yes, would you prefer this to be conducted by

telephone [ ] or e-mail [ ]?

Would you like to receive more background information to this questionnaire or about some of the issues investigated?

Yes [ ] No [ ]

#### DATA PROTECTION AND CONFIDENTIALITY

The Digital Archiving Consultancy ('DAC') is being registered as data controllers under the 1998 Data Protection Act. The data you have provided on this form will only be used by the DAC for research purposes as part of the e-Science Curation Study. Any findings published as a result of this research will be done in an anonymised format only. If you have any questions relating to the processing or confidentiality of your personal data, please contact Philip Lord of the DAC in the first instance.

If you would like to receive a copy of the vintage champagne prize draw rules, please e-mail the DAC at [support@d-archive.co.uk](mailto:support@d-archive.co.uk).

## Data generator stage 2 questionnaire and cover letter

Dear ,

We are now following up on the survey we sent last month as part of our investigations into provision for the long-term care, preservation and re-use ('curation') of digital research data in the UK. Results from the first round of questionnaires have been extremely useful, highlighting issues which require deeper exploration - thank you very much for your answers.

You indicated that you would be willing to take part in a second round of the survey, and below we attach a second short questionnaire designed to explore these issues a little further. At the end there is a space for any comments you would like to make. We would be particularly interested in what risks and/or opportunities you might see in the use, re-use of digital research data over time, what additional resources might be beneficial, or suggestions you might have for areas to research.

We have put background information on the study at [http://www.philiplord.com/e-Science\\_info/web%20info.htm](http://www.philiplord.com/e-Science_info/web%20info.htm).

Please do not hesitate to contact Philip Lord (telephone 0208-607 9102) or me (0208-744 9322), or e-mail us at [pwl@d-archive.co.uk](mailto:pwl@d-archive.co.uk) if you would like any further information or pointers.

Again, we realize how heavy the demands on your time are, but given the importance of the issue (which indirectly may boost research funding) we would be very grateful if you would complete the questionnaire below and return it to [pwl@d-archive.co.uk](mailto:pwl@d-archive.co.uk) as soon as possible. All replies will be treated in strict confidentiality.

Once again, thank you very much.

Alison Macdonald  
Digital Archiving Consultancy  
2 Wayside Court  
TWICKENHAM TW1 2BQ  
Tel: 020-8607 9102  
Fax: 07050-675 010

=====

### DATA CURATION QUESTIONNAIRE

Please mark appropriate boxes with an x, thus [x]

#### THE LONG-TERM USE AND VALUE OF YOUR DIGITAL DATA

1. Can you characterize the value of your data after the end of your project - please mark (X) all that apply.

Further scientific value: [ ]  
Potential commercial value: [ ]  
Evidential value, to confirm conclusions drawn: [ ]  
Historical value: [ ]  
Other (please specify):

2. Think new science can be built on  
primary data you have generated? Yes [ ] No [ ]

3. Do you believe that continuing funding to keep the  
data you are generating would be:

Justified? Yes [ ] No [ ]

Would prove a good investment? Yes [ ] No [ ]

#### DATA RESPONSIBILITIES

3. Who/ what body do you think should take care of the data after  
project end?

4. Effective preservation of data depends on providing good-quality  
description (context, technical, indexing). Much of this is  
best provided by the data originator at the time it is created.  
For your data do you:

Feel you have enough time and funding to  
do this? Yes [ ] No [ ]

Think training about this would be valuable? Yes [ ] No [ ]

5. Would you be willing to see your data kept in:

A repository managed by your institution's

a) Library: Yes [ ] No [ ]

b) Data centre / IT group: Yes [ ] No [ ]

c) Other:

A national repository: Yes [ ] No [ ]

An international repository: Yes [ ] No [ ]

Should this be a general facility? [ ]

or discipline-specific? [ ]

6. If your project data contains confidential information, what  
is the nature of this confidentiality:

Data on individuals(e.g. medical): Yes [ ] No [ ]

Commercial secrets: Yes [ ] No [ ]

Patentable information: Yes [ ] No [ ]

Information, disclosure of which might  
compromise this project or future projects  
for you: Yes [ ] No [ ]

#### READING THE DATA IN THE FUTURE

7. a) Do you mine or re-use your own old  
primary data? Yes [ ] No [ ]

b) Have other people used these data? Yes [ ] No [ ]

8. Have you ever experienced difficulty locating or using other  
people's data:

Yes [ ] No [ ]

If "Yes", was this caused by:

a) Inadequate indexing or descriptive information? Yes [ ] No [ ]

b) Lack of access to software to read it? Yes [ ] No [ ]

c) Other:

9. Do you anticipate using or conducting "collection-based" science in the future (where discovery is made by investigating existing data rather than generating significant new primary data)?

Yes [ ] No [ ]

#### TECHNICAL ASPECTS

10. Do you apply standard vocabularies in your work:

Yes [ ] No [ ]  
What are they:

11. Do you feel there are sufficient tools and technologies available to ensure long-term accessibility to your specific data:

Yes [ ] No [ ]

12. Are you aware of the emerging Research Grid technologies which allow vast access to shared and distributed computing resources and vastly increased storage?

Yes [ ] No [ ]

If "Yes": What benefits, if any, do you do you believe it will yield for you:

OTHER COMMENTS:

#### GENERAL INFORMATION

Please confirm your name (first name, surname):  
Department, institution:  
Please add any further comments you may have:

#### DATA PROTECTION AND CONFIDENTIALITY

The Digital Archiving Consultancy ('DAC') is being registered as data controllers under the 1998 Data Protection Act. The data you have provided on this form will only be used by the DAC for research purposes as part of the e-Science Curation Study. Any findings published as a result of this research will be done in an anonymised format only. If you have any questions relating to the processing or confidentiality of your personal data, please contact Philip Lord of the DAC in the first instance.

## Librarian Questionnaire and Cover Letter

Dear ,

Professor Tony Hey, chairman of the JISC Committee for the Support of Research and director of the UK's e-Science Core Programme, has asked us to investigate provision of, and make recommendations for the long-term 'curation' of the digital data being generated in primary research in the UK.

We need to be able to ensure that data can continue to be accessed and re-used over time, that we can validate our research, and that our research can contribute to dynamic knowledge bases and future research. It is therefore of key importance to seek the views and experience of the academic research library community.

We realize how heavy the demands on your time are, but given the importance of the issue we would be very grateful if you would complete the questionnaire below and return it to [pwl@d-archive.co.uk](mailto:pwl@d-archive.co.uk) as soon as possible. Most questions are Yes/No. All replies will be treated in strict confidentiality.

Further brief information on the study can be found at [http://www.philiplord.com/e-Science\\_info/web%20info2.htm](http://www.philiplord.com/e-Science_info/web%20info2.htm)

If you have any further questions, please do not hesitate to contact Philip Lord (telephone 0208-607 9102) or Alison Macdonald (0208-744 9322), or e-mail us at [pwl@d-archive.co.uk](mailto:pwl@d-archive.co.uk).

Thank you very much for your help. Your input is much appreciated.

Philip Lord  
Digital Archiving Consultancy  
2 Wayside Court  
TWICKENHAM TW1 2BQ

Tel: 0208-607 9102

Fax: 07050-675 010

=====

1. a) Has the proportion of electronic holdings (of all types) in your library increased in the last two years?

Yes [ ] No [ ]

- b) If yes, has this resulted in an increase or decrease in:

- the number of user queries? Increase [ ] Decrease [ ]  
- the time spent on user queries? Increase [ ] Decrease [ ]

If increases, do you think this is a transitional phase?

Yes [ ] No [ ]

- c) Does your library have its own IT staff?

Yes [ ] No [ ]

e) Is your library responsible for information services throughout your institution?

Yes [ ] No [ ]

Comments:

2. Are you considering, or do you already share specific electronic collections, journals with other institutions?

Yes [ ] No [ ]

If "Yes", have the participants come together for geographic reasons, for subject area reasons, or other?

Yes [ ] No [ ] Other:

3. a) Who is responsible for managing digital security in your library?

b) Does that role also cover maintaining integrity of the digital records?

c) Who is responsible for digital preservation in your library?

4. a) Do you have digital holdings other than electronic journals and other commercially available electronic materials?

Yes [ ] No [ ] Don't know [ ]

b) If "Yes", do they pose any particular accession or management problems?

c) Do you accept digital materials in any software format?

Yes [ ] No [ ] Don't know [ ]

5. a) Does your library have any holdings or archives of primary research data?

Yes [ ] No [ ] Don't know [ ]

b) If so, are any of them in digital form?

Yes [ ] No [ ] Don't know [ ]

c) Do you receive special funding to support these holdings (whether paper or digital)?

Yes [ ] No [ ]

d) Are staff or researchers required to deposit digital copies of research papers and publications with your library?

Yes [ ] No [ ] Don't know [ ]

e) Are submitters required to provide descriptive information about the digital materials they are lodging?

Yes [ ] No [ ] Don't know [ ]

f) Are submitters provided with any guidelines for submission of digital materials?

Yes [ ] No [ ] Don't know [ ]

6. a) Does your library specialize in any area(s)? If so, which?

b) Do any of your specialist collections have their own curator(s)?

Yes [ ] No [ ]

c) Has curation work relating to your specialist collections generated new research?

Yes [ ] No [ ]

d) Do any special collections now include digital items?

Yes [ ] No [ ]

e) Do these collections attract specific funding?

Yes [ ] No [ ]

f) Do they generate, directly or indirectly, revenue for your institution?

Yes [ ] No [ ]

g) Have you received any bequests of digital materials?

Yes [ ] No [ ]

8. Digital technology offers the possibility to link annotation to records. Do you think this would be a valuable extension to the knowledge base (whether discipline-specific, inter-disciplinary or wider)?

Yes [ ] No [ ]

Is it practical? Yes [ ] No [ ]

Who/which body/bodies do you think would be most appropriate to provide/oversee these services?

9. Has digital technology resulted in the increased creation of islands of information and information resources within departments? If so, should catalogue information be co-ordinated with/by your library?

10. a) Have you been made aware of developments involving the Grid?

Yes [ ] No [ ] Don't know [ ]

b) If so, do you think the Grid will bring structural change, benefits or entail any particular difficulties for your department?

Yes [ ] No [ ] Don't know [ ]

11. a) Does your institution publish any academic journals?

Yes [ ] No [ ]

b) Do you contribute to your institution's policy with regard to journals?

Yes [ ] No [ ] Don't know [ ]

12. What proportion of your costs does computer storage (hardware, media) represent?

a) Are there any research programmes relating to digital information services or digital libraries which you would like to see?

b) Are you involved in any digital library initiatives or research programmes?

#### GENERAL INFORMATION

Please confirm your name (first name & surname)

Department / institution:

Please add any further comments you may have:

#### Data Protection and Confidentiality

The Digital Archiving Consultancy ('DAC') is being registered as data controllers under the 1998 Data Protection Act. The data you have provided on this form will only be used by the DAC for research purposes as part of the e-Science Curation Study. Any findings published as a result of this research will be done in an anonymized format only. If you have any questions relating to the processing or confidentiality of your personal data, please contact in the first instance Philip Lord of the DAC.

If you are returning this questionnaire by fax, please send to:

07050-675 010

No cover page is needed.

## The Data Centre Questionnaire and Letter

Professor Tony Hey, chairman of the JISC Committee for the Support of Research and also director of the UK's e-Science Core Programme, has asked us to investigate provision of and to make recommendations for the long-term care, preservation and use ('curation') of the digital data being generated in research in the UK. This area may receive significant funding.

At the moment, there is a lot of "old" research data sitting on servers and PCs. The volumes that data centres have to manage are growing massively, as is the complexity of data and networks. A lot of this research data is just gathering dust, but a lot of it should be preserved and also kept in such a way that it can be accessed, re-used, enhanced, contributing to dynamic knowledge bases and future research.

We would therefore be very grateful if you could help us with information and, above all, your views, in the questionnaire below. Please feel free to put in fuller comments, or give us a ring. All answers are treated in strict confidence.

Could you please return the questionnaire to [pwl@d-archive.co.uk](mailto:pwl@d-archive.co.uk) as soon as possible.

Further information on the study can be found at [http://www.philiplord.com/e-Science\\_info/web%20info.htm](http://www.philiplord.com/e-Science_info/web%20info.htm). If you have any further questions, please do not hesitate to contact Philip Lord (telephone 0208-607 9102) or Alison Macdonald (0208-744 9322), or e-mail us at [pwl@d-archive.co.uk](mailto:pwl@d-archive.co.uk).

Your input is very important and much appreciated.

Philip Lord  
 Digital Archiving Consultancy  
 2 Wayside Court  
 TWICKENHAM TW1 2BQ  
 Tel: 0208-607 9102  
 Fax: 07050-675 010  
 =====

### DATA CURATION QUESTIONNAIRE

Please mark appropriate boxes with an x, thus [x]

#### RESPONSIBILITIES

1. What services are you/is your department responsible for:
  - . End-user support: Yes [ ] No [ ]
  - . Systems management: Yes [ ] No [ ]
  - . Network support: Yes [ ] No [ ]
  - . IT security (e.g. firewalls): Yes [ ] No [ ]
  - . Access controls/account management: Yes [ ] No [ ]
  - . Programming assistance or services: Yes [ ] No [ ]
  - . Other (Please specify):
  
2. Approximately how much storage is on systems under your management?

On-line:                      Off-line:

3. Do you think that new Grid technologies will affect your department's responsibilities?

Yes [ ] No [ ]

If "yes", how:

#### LONG-TERM PRESERVATION OF DATA

4. Do you provide any of the following facilities to help users to preserve their data over the long term (>5 years):

- . Long-term storage for media:                      Yes [ ] No [ ]
- . Regular checks on media integrity:                      Yes [ ] No [ ]
- . Migration of data from old media to a new:                      Yes [ ] No [ ]
- . Services to facilitate data format migrations: Yes [ ] No [ ]
- . Monitor and record hardware and software used: Yes [ ] No [ ]
- . Provide advice about data preservation:                      Yes [ ] No [ ]
- . For systems to be de-commissioned do you -
  - Have policies to save data off these systems: Yes [ ] No [ ]
  - Retain specifications of these systems:                      Yes [ ] No [ ]
- . Provide other services for preservation (please specify):

5. Do you have policies which govern the retention of data when de-commissioning systems? Yes [ ] No [ ]

6. Which of these national and international initiatives on digital preservation are you aware of:

The Data Preservation Coalition DPC) in the UK? Yes [ ] No [ ]  
 Digital Library initiatives, such as D-Space? Yes [ ] No [ ]  
 Open Archival Information System (OAIS)? Yes [ ] No [ ]  
 CEDARS and CAMiLEON (at Leeds and Michigan) Yes [ ] No [ ]  
 Others (Please specify):

7. Do you believe that data centres have a role in tackling the digital preservation problem?

Yes [ ] No [ ]

If so, how?

#### DATA MANAGEMENT

8. Do you provide guidance for users on the following:

Good data management: Yes [ ] No [ ] Don't know [ ]  
 Records management: Yes [ ] No [ ] Don't know [ ]  
 Data preservation: Yes [ ] No [ ] Don't know [ ]

9. Is data backed up from all machines regularly (including desk-top systems)?

Yes [ ] No [ ]

10. Is your group responsible for backing up:

Local desktop machines? Yes [ ] No [ ]

Local servers? Yes [ ] No [ ]  
 Central servers? Yes [ ] No [ ]

What media are used?

11. How much data is backed up each day?

12. For how long is backed-up data kept?

YOUR DEPARTMENT/ORGANISATION

13. How would you describe your department:

Central data centre/IT group? Yes [ ] No [ ]

Local data centre/IT group? Yes [ ] No [ ]

Other (please specify)?

14. Is your group in the same department as the library (or information services) group?

Yes [ ] No [ ]

15. Number of staff (full-time and part-time):

16. What is the group's annual budget?:

17. Does this budget include costs for:

Staff? Yes [ ] No [ ]

Media for backups? Yes [ ] No [ ]

Purchase of hardware? Yes [ ] No [ ]

Purchase of software licences? Yes [ ] No [ ]

GENERAL INFORMATION

Please confirm your name (first name, surname):

Department, institution:

Please add any further comments you may have:

DATA PROTECTION AND CONFIDENTIALITY

The Digital Archiving Consultancy ('DAC') is being registered as data controllers under the 1998 Data Protection Act. The data you have provided on this form will only be used by the DAC for research purposes as part of the e-Science Curation Study. Any findings published as a result of this research will be done in an anonymised format only. If you have any questions relating to the processing or confidentiality of your personal data, please contact Philip Lord of the DAC in the first instance.

SP-IT

## Appendix 5: JSCR Invitation to Tender

### **Joint Information Systems Committee**

### **Draft Invitation to tender for a Consultancy**

### **Data Curation for e-science in the UK: an audit to establish requirements for future curation and preservation**

This enquiry document invites proposals to report on the current provision and future requirements for data curation and long-term preservation of primary research data in e-science in the UK.

#### **1. Introduction**

##### **1.1 The e-science Core Programme**

The DTI and the Research Councils are committing £118M to a government-industry programme on e-Science. The reason for this investment is that GRID technology is seen as the natural successor to the world wide web and the UK wants to take a leading role in order to develop solutions for its scientists and developing opportunities for its industry.

The world wide web has revolutionised the way companies do business and fundamentally altered people's personal lives but it can no longer cope with the demands being placed on it by science. The world wide web allows very easy access to information, Grid allows that same easy access to computing power, data processing and communication of the results. The opportunities are immense, it will allow the efficient manipulation of vast amounts of information such as that contained in the human genome or the results from experiments in CERN's new Large Hadron Collider. It will also allow the ability to mine data again and again by comparing existing data sets collected for one purpose with new and previously unrelated information, so generating new knowledge.

There will be significant implications for the future curation of primary research data if we wish to ensure that such data can continue to be accessed and re-used over time. Digital information is now enabling new methods of research, dissemination and collaboration in areas ranging from environmental science to genomics. Requirements for data curation

varies between disciplines but persistence of this information is increasingly important: not only for validation of research but because it contributes to dynamic knowledge bases or future research. Already in the US, many are predicting that the major science driver of high-end computing will soon be data. Digital preservation is now being seen as a core requirement for the US Cyberinfrastructure and the Research Grid in the UK.

The e-Science core programme has been established to co-ordinate the research effort on e-science in the UK. Within the programme each research council is funding a number of pilot projects in their own application areas. There are 22 pilot projects in total.

## **1.2 JISC Committee for the Support of Research (JCSR)**

The JCSR is responsible for ensuring that the JISC provides appropriate infrastructure and services to support the needs of researchers, particularly in the context of the UK Research Grid. In addition to the necessary networking infrastructure, this includes ensuring that key issues such as authentication and data storage/retrieval are addressed. Membership of the JCSR is drawn from across the research community including representatives from the Research Councils. The responsibilities of the JCSR include:

- supporting the requirements of the research community by identifying relevant areas of work appropriate to the JISC;
- helping ensure that appropriate JISC activities remain relevant to the research community where appropriate and, if necessary, identify ways in which these activities can be modified or extended to increase their usefulness in supporting research. In particular this includes work on research related aspects of the information environment, network infrastructure and applications, middleware, technology watch and standards, scholarly communications and on-line resources etc;
- managing JISC involvement in e-Science and the Research Grid with the particular aims of:
  - improving the outreach of the Grid to the wider education environment;
  - seeking a common information environment approach through standards in areas such as authentication and security, the DNER and preservation and thus act as a bridge between the Grid world and the web/XML world;
- advising other JISC sub-committees as appropriate and provide a report to each JISC meeting;
- liaising with other organisations (eg Research Councils, OST, Wellcome Trust, NSF).

## **1.3 JISC Digital Preservation Focus**

The JISC Preservation Focus was established in June 2000. Its role is as follows:

- Developing a long-term retention strategy for digital materials of relevance to HE/FE institutions in the UK.
- Providing a UK focus for the development of practices, policies and strategies for the preservation of digital materials.
- Generating support and collaborative funding from and promoting inter-working with appropriate agencies worldwide.

The JISC is currently developing its Continuing Access and Digital Preservation Strategy for 2002-5. This will replace the JISC Interim Digital Preservation Strategy approved in November 1998, which set out JISC's objectives for the Digital Preservation Focus and related initiatives. The consultation draft of the new strategy includes proposals to create a Digital Preservation Development Centre and for working with research councils and others to support long-term curation and preservation of primary research data.

#### **1.4 Further Information**

**For further information about data curation in e-science see:**

*An Investigation into the Digital Preservation Needs of Universities and Research Funders* (Denise Lievesley and Simon Jones 1998) available from:

<http://www.ukoln.ac.uk/services/papers/bl/blri109/datrep.html>

*Digital Curation: digital archives, libraries, and e-science* 19 October 2001

This was an invitational seminar sponsored by the Digital Preservation Coalition and the British National Space Centre. The seminar aimed to share practical experience of digital curation in the digital library sector, archives, and e-sciences. Available from:

<http://www.dpconline.org/graphics/events/digitalarchives.html>

**For further information about the JISC Continuing Access and Digital Preservation Strategy for 2002-5 see:**

<http://www.jisc.ac.uk/dner/preservation/dpstrategy2002b.html>

**For further information about the e-science programme see:**

<http://www.escience-grid.org.uk/>

**For further information about the JISC Committee for the Support of Research see:**

<http://www.jisc.ac.uk/jcsr/index.html>

## **2. Aims & Objectives**

This consultancy is being funded by the JISC Committee for the Support of Research chaired by Prof Tony Hey (Director of the e-science Core Programme).

The consultancy will establish the current provision and future requirements for curation of primary research data being generated within e-science in the UK. This will include the e-science core programme but is anticipated to extend beyond this to other e-science research and primary research data. A consultancy report will provide a synthesis of findings and make recommendations for future action.

The consultancy will support JCSR's aims to manage JISC involvement in e-Science and the Research Grid, and to work in partnership to support the research community through activities such as its digital preservation programme.

## **3. Requirements**

To establish the current provision and future requirements for curation of primary research data within UK e-science the consultancy will:

- through desk-top research, synthesise:
  - existing reports to provide a context for the study;
  - existing practice, policy and guidance to provide an overview of the current provision for curation of primary research data in the UK;
  - requirements within the relevant research communities for future curation in the medium term (5-10 years) and long-term (10 years plus);

It is anticipated that sources for this will include:

- sources noted in 1.4 Further Information above;
  - relevant practice, policy and guidance from the e-science programme, JISC, Research Councils, Arts and Humanities Research Board, Data Centres and Services;
  - reports from the e-science sub-group of the Research Support Libraries Group on disciplinary needs (copies will be made available to the consultant).
- 
- through a combination of postal/telephone survey and interviews, audit:
    - Ownership and responsibility for long-term curation of primary research data;
    - Perceived future value and re-use;
    - Provision made for future curation;
    - Grant conditions for curation and re-use;
  
  - Relevant guidelines, standards, tools, and funding available for projects to prepare data for future curation;
  - Primary research data being produced by the e-science programme projects and identify what kind of material (eg closed or dynamic datasets) and how big (relatively) are they likely to be;
  - Other materials being produced or dependencies critical to their future curation and re-use (eg metadata, technical documentation);
  - Procedures and standards followed in the creation and validation of data and other materials;
- 
- report on:
    - Key findings and issues relating to current provision for curation of primary research data in the UK;
  
    - Future curation requirements for e-science in the UK;
  
    - Recommendations to JCSR;

- Proposed implementation plan and funding required for implementation of recommendations and options outlined in the report.

The deliverable for the consultancy is a structured report with an Executive Summary and appropriate appendices covering the above requirements.

#### **4. Scope**

The report will cover:

- e-science in the UK;
- Primary research data and any additional material (eg metadata or documentation) required for their effective curation and re-use;
- Current provision and future requirements for data curation.

#### **5. JISC management of the consultancy**

Neil Beagrie (Programme Director JISC Digital Preservation Focus email: [preservation@jisc.ac.uk](mailto:preservation@jisc.ac.uk)) will be responsible for managing the consultancy. A small steering group will be established to advise the consultancy and comment on draft documents produced.

#### **6. Proposal requirements and timescales**

Proposals should:

- Outline methodologies, timescales and deliverables. This must include a process for consultation and validation of draft documents.
- Itemise resources required to carry out the study, with costs (including VAT if applicable)
- Indicate skills and expertise of the applicant(s)

**Responses to this enquiry document marked "e-science curation consultancy" must contain four print copies and one electronic copy of the proposal and arrive not later than 5 p.m. on Friday 9<sup>th</sup> August 2002 with**

Rachel Merrett  
JISC Executive  
Northavon House  
Coldharbour Lane  
Bristol BS16 1QD

Tel: 0117 931 7124  
Fax: 0117 9317255

Email: r.merrett@hefce.ac.uk

.

**We expect to agree a contract for the consultancy by 1st September 2002.**

**A draft report will be required by Friday 15 November 2002 for consideration by JCSR on 25 November. A final version of the report will be required by 15 January 2003.**

## **7. Evaluation**

Proposals will be evaluated by referees. The following criteria will be used for evaluation purposes:

- Demonstration of understanding of the issues (including technical)
- Demonstration of ability to carry out the work required by the required deadline
- The proposer's experience, knowledge, and track record of carrying out similar work
- Sound methodology
- Appropriate use of resources

## **8. Publication**

A copy of the report will be published on the JISC website. Reports should be delivered in an appropriate electronic format, to be agreed with JISC. Proposers must be prepared to assign copyrights (including electronic) for the study outputs to JISC, or its nominee.

## Appendix 6: The Digital Archiving Consultancy Team

Lead consultant: Mr. Philip Lord  
Senior consultant: Miss Alison Macdonald  
Consultant: Mr. Jon Friend  
Consultant: Dr. Bruce Pilsworth  
Research Assistant: Miss Helen Tingle