

Digital Data Curation Task Force

Report of the Task Force Strategy Discussion Day

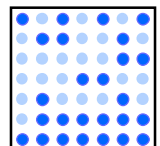
Tuesday, 26th November 2002

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

Digital Archiving Consultancy
2 Wayside Court, Arlington Road
TWICKENHAM
Middlesex
TW1 2BQ
UK



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 26th 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.

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 scene¹.

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 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

¹ 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).

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 Allden 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 Allden 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 Alden 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

Task Force Members

✓ Attendee at 26th 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 Allden | 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

Agenda - Digital Data Curation Strategy

Task Force Discussion Day – Chairman: Professor Tony Hey

Tuesday, 26th November 2002

Centre Point, London WC1

9:30 hrs.	Coffee	
10.00	Opening session:	
	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	e-Curation strategy: establishing the objective of the strategy:	<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)	Formulating a strategy to achieve this objective	<i>Round- table discussion</i>
	What are the fundamental questions? Global/umbrella strategy, plural strategies? Time frame? What do we have at starting point?	
	Pre-lunch summary, presentation of afternoon: break-out sessions after lunch look at factors.	
12.30	Lunch	
13.30	Break-out sessions: three groups:	
	<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	Reporting back from the break out sessions	Group representatives
14.50	Tea break	
15.00	Synthesis: Review of objective; identification of elements of strategy.	<i>Round table discussions</i>
	Closing remarks	<i>Tony Hey / Neil Beagrie</i>
16:30-17.00	End	

Appendix 3

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 Alden | <ul style="list-style-type: none"> • Proofs of concept – determining whether curation is worth it? • Proof of concept as a starting point to develop a strategy. • Test models of re-use models of data curation – citation would come into that. |
| Rolf Apweiler | <ul style="list-style-type: none"> • 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. |
| Sean Barker | <ul style="list-style-type: none"> • Inter-operability. • For instance, defining standards to make metadata inter-operable between repositories, allowing inter-disciplinary work. |
| Neil Beagrie | <ul style="list-style-type: none"> • Link data curation with the publishing process. • Proof of concept is needed. • Encourage thinking in the research councils on curation as an issue. • More research library initiatives are needed to guide us. • Stimulate thinking and existing good practice. |
| Peter Buneman | <ul style="list-style-type: none"> • Curation of databases/evolving datasets. • Developing models and tools for annotation and provenance. • Database archiving. <p>(Contribution received after the meeting)</p> |
| Mikael Dahlin | <ul style="list-style-type: none"> • Standards are needed for how to model processes. Broader models of processes are needed. • Storage – file formats need to be standardised. • Find out how to validate archival processes and information authenticity. • Define the borderlines between different processes – such as libraries, archives. |
| Peter Dukes | <ul style="list-style-type: none"> • 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

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 other 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

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

Alison Allden:

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).