e-Science Curation Report

Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision

prepared for:

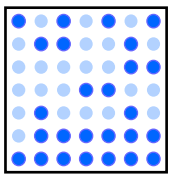
The JISC Committee for the Support of Research (JCSR)

by
Philip Lord and Alison Macdonald

The Digital Archiving Consultancy Limited

2 Wayside Court, Arlington Road, Twickenham, TW1 2BQ

e-mail: info@d-archiving.com



2003

Contents

…./cont'd.

Appendices (volume 2)

Preface

This report was funded by the 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 in the UK, particularly in the context of the UK Research Grid. Membership of the JCSR is drawn from across the research community including representatives from the Research Councils.

In recent years the JISC in partnership with other bodies has undertaken significant initiatives towards the long-term care of data in the UK academic sector. In 1998 a report for the JISC and the National Preservation Office was carried out by Professor Denise Lievesley and Simon Jones[1] into the digital preservation needs of universities and research funders. It recommended to funding bodies:

- The development of national guidelines covering the key areas of concern.
- The development of standards to allow kitemarking of centres where research data can be managed and preserved over the long term.
- The development of a national policy on research data and dissemination of information about this national policy.

In 1998 the JISC CEI (Committee on Electronic Information) Interim Preservation Strategy 1998-2001 was adopted[2]. This was superseded in 2002 by "A Continuing Access and Digital Preservation Strategy for the JISC 2002-2005"[3]. The new strategy set out the role of JISC in partnership with others in this field, outlined objectives and an implementation plan. The implementation of this strategy has initiated a number of studies to quantify requirements and implementation issues.

It is in this context that the JISC Committee for the Support of Research commissioned this study to establish the current provision and future requirements for curation of primary research data generated within e-Science in the UK.

Acknowledgements

Our particular thanks go to Neil Beagrie who has overseen the study and given extensive input, and to our steering group of David Boyd, Fred Hopper and Alan Rector for all their help and guidance. A large number of other extremely busy people have given us a lot of their time and expertise. These are not just those listed in our interviews in appendix 1; we would also like to express our thanks to Krys Bartoszewska in Professor Tony Hey's office,

---

[1] Lievesley, D. & Jones, S., 1998

[2] JISC CEI, 1998

[3] Beagrie, N., 2002

Alistair Knowles in the National e-Science Centre, Karen Mee in Carole Goble's office for their help, advice and sharing of knowledge.

Abbreviations

The following abbreviations are used frequently in the report and its appendices:

AHDS	Arts and Humanities Data Service
AHRB	Arts and Humanities Research Board
BBSRC	Biotechnology and Biological Sciences Research Council
CAMiLEON	Creative Archiving at Michigan & Leeds: Emulating the Old on the New (a JISC/NSF jointly funded project)
CCLRC	Council for the Central Laboratory of the Research Councils
CEDARS	CURL Exemplars In Digital Archives
DAC	The Digital Archiving Consultancy Limited
DCC	Digital Curation Centre
DPC	Digital Preservation Coalition
DTI	Department of Trade and Industry
EPSRC	Engineering and Physical Sciences Research Council
ESRC	Economic and Social Research Council
HE/FE	Higher Education/Further Education
HEFCs	Higher Education Funding Councils
HEI	Higher Education Institution
ISO	International Standards Organisation
JISC	The Joint Information Systems Committee
JSCR	The JISC Committee for the Support of Research
MRC	Medical Research Council
NERC	Natural Environment Research Council
NSF	National Science Foundation (US body)
OAIS	Open Archival Information System
OST	Office of Science and Technology
PPARC	Particle Physics and Astronomy Research Council
PRO	Public Records Office (now the National Archives)
RDN	Resource Discovery Network
RSLG	Research Support Libraries Group
UK DA	UK Data Archive, Essex University

Other abbreviations are introduced in the text.

Executive summary

Science is being transformed by accelerating change in information technology, with huge increases in computing power and network bandwidth, accompanied by an explosion in data volumes and information.

The term e-Science – or more inclusively "e-Research" - has been used recently to describe the research culture and opportunities enabled by these developments, and the collaborations of people and of shared resources that are needed to resolve new research challenges, whether in the sciences, social sciences or humanities.. e-Science enables a new order of collaborative, more inter-disciplinary research, based on shared research expertise, instruments and computing resources, and crucially increasing access to collections of primary research data and information - the knowledge base of research. The term e-Science is applied to these techniques when applied to the sciences. In this report we use the term e-Science.

There are challenges, however: these same technology changes put the very data they create and use at risk, and raise serious and complex issues of strategy and policy regarding its creation, management, and long-term care – its curation – for which top-level responsibility urgently needs to be adopted to protect and further UK research.

Our study examined the current provision and future needs of curation of primary research data in the UK, particularly within the e-Science context. At a strategic level we found:

1. Confirmation that the data revolution presents significant challenges and opportunities. However, our surveys show that the UK is not fully prepared to capitalize on the opportunities and urgently needs to address this.

2. There is a lack of a government-level, overall strategy for data stewardship and data infrastructure to which research administrators can refer, still less to support researchers in their evolving roles and duties with regard to data curation.

3. Existing data centres are usually supported by sponsors whose primary funding focus is research projects.

4. The current short-term funding models for the provision of curation are antithetical to its long-term nature and needs.

5. There will be an exponential increase in data volumes from e-Science over the next decade as planned new scientific instruments and experiments come on stream. However, to benefit fully from this major investment, further action is needed to support the curation of the data that will be generated.

6. Not all primary research data needs to be retained or has long-term value. Its potential value for generating new research will vary, and the level of investment in the curation of datasets therefore needs to be identified and graduated accordingly.

At a policy level we found:

7. Provision of curation is patchy, and more advanced in some disciplines than others; the basic life sciences, and "big" collaborative sciences such as particle physics and astronomy are examples where provision is most advanced.

8. Where retention and curation of primary research data is a requirement set by funding bodies, the majority of researchers surveyed stated this requirement was not funded. Where guidance is provided, researchers frequently felt that it was out of date or inadequate.

9. Awareness of the issues - particularly data longevity difficulties - is generally low among researchers. Consequently the good practice needed to assure data longevity is rare, putting valuable resources at risk.

10. For curation to be effective the researcher needs to be engaged in the curation of his or her own data, working in partnership with curators. But few incentives or procedures are in place to ensure that this engagement is achieved.

11. Whilst practice and experience in curation is increasing rapidly, areas of curation are still in a research and proof-of-concept phase. Much research and practical, exploratory activity is being undertaken in the UK, and its quality is world-class.

12. The data revolution raises many issues of trust which must be addressed before data-based research can flourish – issues of security, confidentiality, ownership, assured provenance, authenticity, and data and metadata quality.

13. There is little interaction and sharing in curation experiences between science-based industry and the academic sector. Within the next decade the curation of digital content and data is likely to be critical to science- and engineering-based industries and to knowledge-based economic activity.

Based on these findings we set out our major, strategic recommendations in the list overleaf. Our report details further specific recommendations within the body of the report, where we also outline proposals for the organisational structuring of curation provision and provide a table showing which recommendations address the findings summarised above.

These recommendations cross organisational boundaries and span organisational levels.

	Strategic recommendations
A1	Strategic-level advocacy for data curation is needed and should be assigned to a respected and influential champion so that strategic objectives are clearly articulated, to set the UK's curation agenda over the medium term, and to enhance the UK's standing, contribution and opportunities in this area.
A2	A curation task force made up of curation experts, practising researchers and research administrators should be established to inform and guide this agenda. This task force should work closely with and inform the work of the new UK Digital Curation Centre.
A3	The mismatch of short-term funding against the long-term needs for data retention needs to be addressed by providing new specific, long-term funding stream(s) for data centres and curation, thus ensuring that there is a strategic approach to data stewardship which addresses holding information indefinitely, makes it widely available and encourages cross-disciplinary usage, including linking to other digital information.
A4	Funding bodies should consider supporting research-led exemplars of curation to demonstrate and promote the benefits of curation for new research.
A5	Our findings endorse the need for the Digital Curation Centre and we recommend its establishment as part of a national provision for data curation in the UK.
A6	Criteria need to be established to determine what data we should keep, why and what level of curation is appropriate, together with mechanisms to monitor, validate and to modify them with accumulating experience.
A7	A programme of activities, both national and international, should be initiated to promote incentives which will foster a scientific culture of engagement in data care.
A8	Educational materials, guidelines and policy documents for researchers need to be developed and publicised.
A9	There should be increased investment, knowledge transfer, and cross-sector partnerships with knowledge-based and science and engineering industries to capitalize on UK expertise in data curation. This should be led by the DTI.
A10	Investment should be strengthened in those areas of curation research which will enhance data re-use; in particular we recommend focusing on those areas of research needed to establish trust in curated information.

It is our view that, as the highest priority, responsibilities should be assigned for the strategic recommendations. Following feedback from JCSR, the following responsibilities for taking actions are recommended:

	Action to be taken by:
A1	e-Science Core Programme to follow up with the RCUK.
A2	JCSR should take responsibility for establishing a Curation Task Force which could inform the strategic implementation of the digital curation agenda.
A3	HEFCE and OST.
A4	The production of research-led exemplars of curation could be co-ordinated by the new Digital Curation Centre.
A5	The Digital Curation Centre is now being established, managed by JCSR.
A6	These recommendations are the responsibility of the Research Councils and should be included in the paper which will be presented to the RCUK at a future meeting.
A7	The Digital Preservation Coalition and the Digital Curation Centre.
A8	e-Science Core Programme and Research Councils.
A9	e-Science Core Programme to follow up with the DTI.
A10	As for A2 above.

1 Introduction

This report examines the current provision of, and future requirements for, the curation of digital primary research data generated in academic and scientific research in the United Kingdom, with a focus on data enabled by "e-Science" – that is, data-intensive, computing-intensive, collaborative, dispersed.

Increasingly digital data is a vehicle for new discoveries. We can re-use it to extract additional value or simply avoid duplicating existing work. Digital technologies enable sophisticated collaboration and sharing within and between disciplines (where some of the most fruitful work lies). Proper retention of digital data is essential to demonstrate validity, and for respect of legal and ethical values. Digital data is already part of the history of science.

Observation and data underpin, validate and feed research activity. We now have unprecedented ability to observe and to record, from the sub-atomic particles to far galaxies. Models and simulations yield predictions and discoveries in every area of science. Computing power and telecommunications advances increase the volumes and reach of the digital data generated.

The current absolute volume and the rate at which digital data are increasing are astonishing. An estimate[4] in 2000 put the **annual** production of information to be 1.5 exabytes ($1.5x\ 2^{60}$ or $1.5x10^{18}$ bytes), of which only 0.003% was in printed form. Research is a formidable contributor to these figures; examples are legion, the most commonly quoted perhaps the petabytes (2^{50} or 10^{15} bytes) to be generated annually by the Large Hadron Collider when it comes on stream (and this is the consolidated value after raw data capture at rates up to 1 Gigabyte per second). A few other, more mundane examples from international science based in the UK emphasise the point:

- The EMBL Nucleotide Sequence at the EBI in Cambridge has tripled in size over eleven months; the database is of the order of terabytes (2^{40} or 10^{12} bytes)
- The European Centre for Medium Range Weather Forecasting (ECMWF) in Reading has about 330 terabytes, increasing at some 0.5 terabytes per day
- UK social science data holdings exceed one terabyte, more than doubling since 1995.[5]

We are moving into an era when exabytes of data and petaflops of computing power will not be thought extraordinary. There is a deluge[6] of data, in particular coming out of and used in primary research. While absolute volumes of data are increasing exponentially, the number

[4] Layman, P. and Varian, H.R., 2000; see also update at http://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm

[5] From Hey, T. and Trefethen, A, 2003

[6] Hey, T. and Trefethen, A, ibid.

of digital "objects" is increasing, and the objects themselves are heterogeneous and increasing in complexity.

The ways of performing research are changing in response to these developments: data-based research is becoming more important; there are drives for more data sharing, collaboration and cross-disciplinary work, much of it international. The development of Grid technologies[7] supports and drives the trend; in the UK the e-Science initiative[8] coordinates responses to it.

These challenges raise questions of curation of data for the Research Councils and higher education sector, which this report has been asked to address. The UK government invests some £7 billion annually in scientific research and development across all government departments; the UK's e-Science budget for 2001-2006 will total £213 million. Excellent digital curation is an opportunity to convert a proportion of these expenditures into capital in the form of an efficient, rich knowledge base, which itself supports and generates new science. Our report is presented in the belief that it can contribute to this agenda.

[7] Foster, I. and Kesselman, C. 1999

[8] See: http://www.research-councils.ac.uk/escience/

2 Methodology and work performed

The objectives of the data-gathering stage of this study were to audit current provision for the curation of primary research data in higher education and e-Science and to identify future curation requirements. To do this, we:

a) Carried out desk-top research examining literature referred to in the JCSR call for proposals and further sources identified by the authors. A list of references is provided in section 8.4 of this report (not all are cited in the text).

b) Administered a series of questionnaires addressed to researchers creating data, data centres, libraries and providers of information, IT service providers, and to policy makers. The detailed findings from the questionnaires are set out in Appendix 3. The questionnaires themselves are reproduced in appendix 4.

c) Conducted face-to-face interviews with experts and researchers. We consulted some 40 individuals formally and at length. Appendix 1 provides a list of the interviewees. We also spoke to many others informally.

During the study Professor Tony Hey, Chairman of the JCSR, asked the Digital Archiving Consultancy (DAC) to assemble a task force for a meeting to brainstorm the question of strategy for digital data curation. This forum met in late November 2002 and provided further valuable insights into current provision, future vision and requirements. A report[9] of the meeting was distributed in January 2003, and is included here in appendix 2.

The authors also attended the Medical Research Council's Archiving Horizons day, organized by Dr Peter Dukes and his team, exploring data sharing and preservation issues, with a particular focus on the epidemiology community.

In conducting the study we looked beyond the scientific environment, referring to relevant experience in the arts and humanities, administration and the private sector. (The term "science" or "scientific" is often used in this report in a wide sense, as in knowledge (or the German "Wissenschaft"), rather than just the physical and natural sciences.)

Work on this study began in mid September 2002 and was overseen by a steering group comprising Neil Beagrie of the JISC, Dr. David Boyd of CCLRC, Dr. Fred Hopper of NERC, and Professor Alan Rector of the University of Manchester, who all gave expert and valuable guidance. Data was collected in the winter of 2002 to spring 2003. The Digital Archiving Consultancy team is set out in appendix 6, the terms of reference in appendix 5.

In presenting our findings in the following chapters we have treated these on an issue-by-issue basis, rather than presenting separately the results of each activity described above, to avoid repetition and excessive cross-referencing. All views expressed to us have been anonymised.

[9] Macdonald, A. and Lord, P., 2003.

2.1 Working definitions

This is a relatively new field, and terminologies are not yet stable. For this paper we used **working** definitions of three key activities: "curation", "archiving" and "preservation".

> **Curation**: The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials[10].

> **Archiving**: A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity[11].

> **Preservation**: An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.[12]

Thus, very broadly speaking, these are terms of increasing specificity in this context: preservation is an aspect of archiving, and archiving is an activity needed for curation. All three are concerned with managing change over time.

2.2 Conventions

We indicate recommendations at appropriate points in the text by placing these as follows:

> Recommendation A99: We recommend that . . .

Recommendations are numbered in two series: the "A" series are our main, strategic recommendations, and the "B" series are less urgent and of a more tactical, detailed nature. Numbering of the "A" series reflects ordering in the executive summary; the "B" series is numbered by position in the full text. All are listed in section 8.3.

Quotations and similar materials are shown thus:

> Text in this style indicates a quote or highlight.

---

[10] Further discussion of the term curation is provided in chapter 5.

[11] The term archiving has widely varying professional use. The definition used here is closest to that used by traditional archivists. However, computer scientists often use the term to refer to professionally managed storage without the selection, authenticity, and preservation tasks included here.

[12] Elaborated by Hedstrom, M., 1998, and quoted in Cedars, 2002a and 2002b, as "the planning, resource allocation, and application of preservation methods and technologies necessary to ensure digital information of continuing value remains accessible and useable".

3 RATIONALE

3.1 Why archive and curate primary research data?

In the research context of our report, major reasons to keep primary research data include:

- Re-use of data for new research, including collection-based research to generate new science.
- Retention of unique observational data which is impossible to re-create.
- More data is available for research projects.
- Compliance with legal requirements.
- Ability to validate research results.
- Use of data in teaching.
- For the public good.

Some of these benefits may arise indirectly, rather than as the result of deliberate policy[13]. Indirect benefits include the provision of primary research data to commercial entities, and use in commercial products.

The following quotation[14], talking here about human genome data, provides one illustration of the breadth of potential for data re-use:

> "… these data [….] will be used for both diagnosis and prognosis. They will be analyzed for contextual effects on the frequencies of mutation. They will be used to study the origin and dispersal of human populations. They will be used to study the relationship between protein structure and function. They will be used by molecular biologists, forensic scientists, epidemiologists, demographers, public health personnel, population biologists, genome mappers, medical students, insurance companies, providers and servers of diagnostic tools, drug designers and the general public. There will be other purposes and users, as yet not imagined."

The curation of data will help maximize the potential of data, facilitating research, increasing its quality, extending the knowledge base through annotation, links and visibility.

However, without perception of benefit, digital curation could stay grounded - yet benefit can only be demonstrated by actually "doing" digital curation over a sustained period of time.

---

[13] The document prefacing the Proposal for a Directive of the European Parliament and of the Council on the re-use and commercial exploitation of public-sector documents (European Commission, 2002, Brussels.) states, "The new information society technologies have led to unprecedented possibilities to combine data taken from different sources and create added-value products". The report quotes an attempt at quantification in Pira International's 2000 report on "Commercial exploitation of Europe's public sector information", estimating the economic value of public-sector information in the European Union at around €68 billion.

[14] Michael Ashburner, writing when head of the European Bioinformatics Institute, and with acknowledgement to Jim Ostell.

Recommendation A4:	Funding bodies should consider supporting research-led exemplars of curation to demonstrate and promote the benefits of curation for new research.

It would be possible to track and provide some measure of the benefits of data retention and curation noted above (those marked ❖), and we discuss measurement in chapter 8 below.

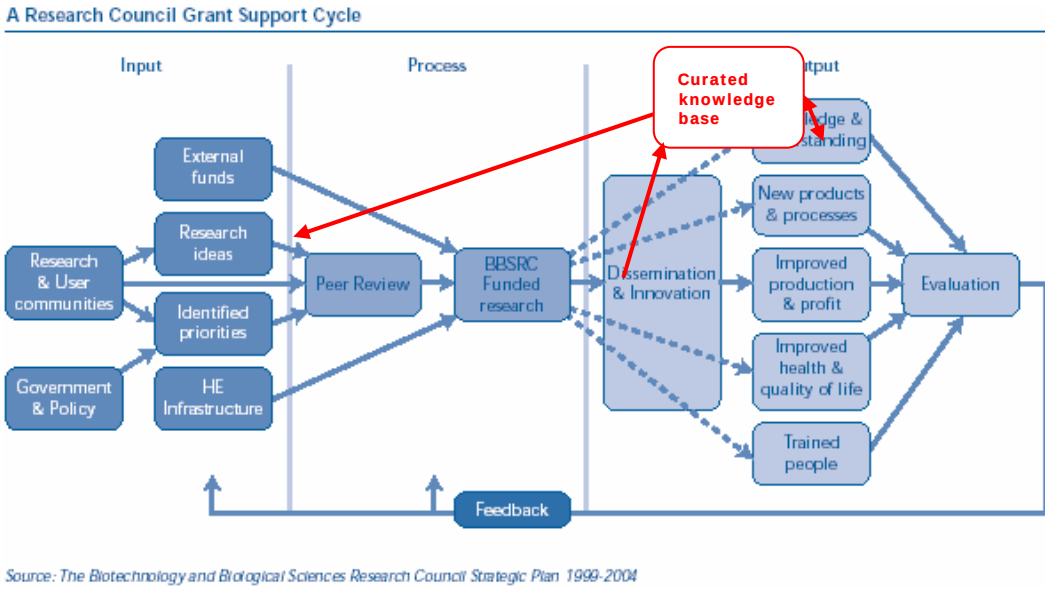
Recommendation B1:	Measures to quantify direct and indirect benefits from curation should be developed.

Without some quantification as to the size and type of resources entailed, and their timings and spans, it will not be possible for policy makers and funders to make informed resource allocation decisions.

Data is an output. A basic reason to keep the data is that it has been produced in the first place, often at considerable expense[15]-, generating and collecting data costs billions, whether pounds, dollars, or euros. Some data collected is unique and non-reproducible, which suggests we have a responsibility to keep the data, which in many cases is already recognized in mandates to bodies such as the NERC, with some holdings stretching back over centuries.

The diagram (Figure 1) summarizes a current Research Council support cycle, setting out the inputs, process and outputs as seen in 1999. Although these structures are now (December 2003) under review and may change, it shows data is now a significant element in the "knowledge and understanding" output, and should be recognized as such.

Figure 1: Inputs and outputs in the Research Council grant cycle



Source: The Biotechnology and Biological Sciences Research Council Strategic Plan 1999-2004

[15] Professor Cameron, European Bioinformatics Institute, quoted in European Commission Working Paper: Workshop report on managing IPR [intellectual property rights] in a knowledge-based economy – bioinformatics and the influence of public policy. Rapporteur Stephen Crespi, November 2001.

Our questionnaire findings show that scientists believe that a high proportion of primary research data generated can be re-used:

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

	Yes	No
Primary data	79%	21%
Summary / derived data	90%	10%
Published data	94%	6%

A subset of these respondents gave some indicators as to the nature of this value, indicating that the primary value lies predominantly in future scientific worth and scientific validation, with commercial value and historical value also mentioned[16].

3.2 Curation serves government objectives

Overall, "the Government sees science, engineering and technology as critical to our success as an economic power and to improving the quality of our lives, whether it is our health or the environment", in the words of Lord Sainsbury of Turville, Minister at the head of the Office of Science and Technology. Referring to the biotechnology sector in particular, the government sees one of its roles in this context as "creating the right environment for the sector to grow and prosper - to give those – the scientists, the entrepreneurs and the investors – who create the wealth the best environment in which to work." Curated databases are a critically important part of the sector's resource and investment – and it is surely positive for UK industry that several international databases are based here in the UK.

In our literature review we looked at strategy and policy documents of the DTI, OST, Research Councils, Higher Education Funding Councils, and more. Digital curation supports almost all objectives and priorities set. In several cases, data storage, data collection, and curation are specifically named as priorities (by the BBSRC, MRC and NERC).

Digital curation supports all four of the research objectives listed in the DTI's current Science Budget Strategic Objectives, in particular RO4:

[16] See Appendix 2

Table 2: DTI science budget and curation relevance

Relevant	Science Budget 2003-04 to 2005-06:
✓	**RO1.** To continue to improve the excellence, relevance and impact of the knowledge created from Research Council-funded programmes.
✓	**RO2.** To increase research capability and international competitiveness of the UK in new strategic areas.
✓	**RO3.** To increase the dynamism and flexibility of Research Council programmes to respond to changing requirements and opportunities, and to support effectively multi-disciplinary research, new researchers and higher risk research proposals.
✓	**RO4.** To maintain access for scientists working in the UK to the necessary major facilities, databases and supporting laboratory infrastructure that will enable them to deliver world-class research.

Digital curation also supports the Knowledge Transfer Objectives in this Science Budget. It is relevant to two of the four funding and policy commitments announced in the Treasury's "Investing in Innovation" document:

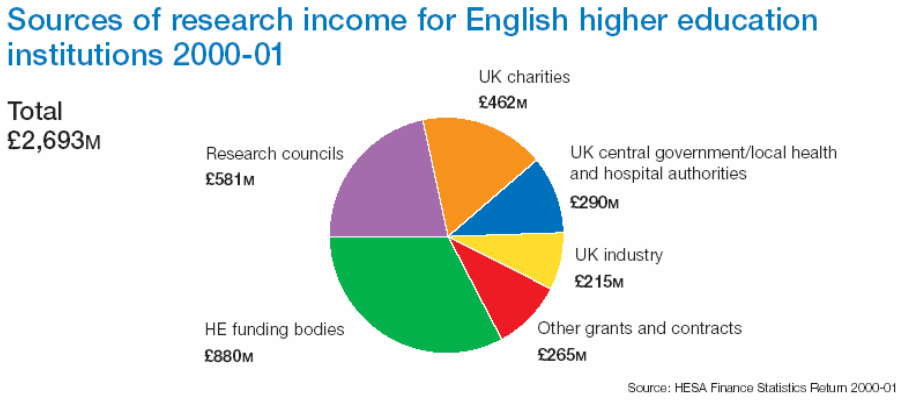
- "Measures to put UK university research on a long-term sustainable footing,
- Increased funding to further improve the exploitation of the knowledge and technologies generated by research in the science and engineering base."

4 Current situation and provision

> "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."
> *Questionnaire response (professorial)*

In general, primary research data is generated within three different time and funding frames: as (a) short-term projects (up to five years; the mode is three years), (b) longer-term large-scale projects, or (c) open-ended data collection. Types (b) and (c) are generally collaborative and have an international component. The publicly funded scientific research generating this data is conducted in the UK in higher education institutes and research or other specialist institutes. Most come under the education departments (the HEIs via the HEFCE) and the DTI - the seven Research Councils report to the Office of Science of Technology; they cover broad scientific streams and provide funding and strategic direction for their research establishments. The Research Councils also fund research in universities through grants and programmes. Charities (in particular the Wellcome Trust) fund a significant percentage of research as well (see Figure 2 below[17]).

Figure 2.



The range of funding stakeholders also includes trans-national organizations.

Facilities, structures and provision for the collection and archiving of digital research data vary between and within Research Councils and institutions. While there is contact and consultation between digital archive providers and research council officers at managerial level, there has been very little high-level formal collaboration between Research Councils

[17] Higher Education Funding Council for England Annual Report 2001-2002

with regard to primary research data, with the exception of the initiatives taken within the e-Science Core Programme. A recurring theme during interviews was the lack of overall policy to direct the issue of digital curation in the UK. This lack is reflected in feedback from the ground in comments from less senior researchers and from the questionnaire surveys: people do not have a clear view of what they need to do to archive their data and are seeking direction.

Recommendation A1:	Strategic-level advocacy for data curation is needed and should be assigned to a respected and influential champion so that strategic objectives are clearly articulated, and to set the UK's curation agenda over the medium term, and to enhance the UK's standing, contribution and opportunities in this area.
Recommendation A2:	A curation task force made up of curation experts, practising researchers and research administrators should be established to inform and guide this agenda. This task force should work closely with and inform the work of the new UK Digital Curation Centre.

Current provision for the care of and access to the primary data generated in research varies considerably, with some Research Councils providing data repositories, where in a few cases data submission is a requirement of funding, for example, ESRC and NERC. JISC itself funds national services or co-funds data archives (for example, MIMAS, EDINA, co-funds AHDS, and the proposed DCC.); in other domains the UK contributes to international repositories (CERN, for example). In other areas cupboards, drawers, and departmental servers were among the more common repositories we encountered during our study.

Our findings confirm that a considerable proportion of primary research data, some important, is at risk, going to waste or decaying for want of resources in the form of time, awareness, and tools.

4.1 Repositories, existing curation activities

University-based centres

There are five university-based data archiving centres in the UK, none of which specifically focus on scientific information (in the natural sciences sense, though the UK Data Archive (UKDA) takes in population data sets from the MRC, which includes confidential medical data). These are the UK Data Archive based at the University of Essex, and the Arts and Humanities Data Service (AHDS) which has a central management unit at Kings College London and five distributed discipline-specific satellites. The University of London Computer Centre (ULCC) has a unit with archives data, but this is mainly public-sector data (such as for the National Archives, formerly the Public Record Office).

The ESRC's Economic and Social Data Service (ESDS) is a new national data service that came into operation in January 2003. Based at the UKDA, ESDS has been established as a distributed service, based on collaboration between UKDA, the University of Essex' Institute

for Social and Economic Research, MIMAS (see below), and the Cathie Marsh Centre for Census and Survey Research at the University of Manchester. It aims to promote and encourage data usage in teaching and research.

The UKDA and AHDS are examples of digital archives which manage **collections** of digital resources, taking preservation actions on individual resources within the collections as and when desirable or required. Most data centres we encountered during this study fell into the other, **unitary** category, however, where they provide a gateway into one resource (though these may have their own internal partitions). At these data centres preservation is usually limited to ongoing maintenance of a database which is upgraded as technologies are upgraded, and digital contents upgraded to cope with changing needs of presentation and dissemination. A third category are those **dynamic databases** which are continuously curated, such as SWISS-Prot.

MIMAS in Manchester University and EDINA in Edinburgh University are examples of JISC-funded centres which provide access to data and publications from third parties. EDINA staff has been looking into best practice for identifying the locus of archival responsibility for the various resources it hosts, although it does not hold archival responsibility for these. MIMAS at Manchester hosts a range of data sets and other resources – socio-economic (some in collaboration with the UK DA), scientific, spatial datasets. Another, non-archive service at Manchester, CSAR (Computer Services for Academic Research) operates an interesting charging system, whereby Research Councils make allocations to users in "generic service tokens" which can be exchanged for service provision, but these can also be traded within a trading pool as users' requirements change.

Science and technology research councils

The **BBSRC's** Bioscience Information Technology Services (BITS), located at the BBSRC Rothamsted site, is the common service provider of IT services for BBSRC-supported institutes. This includes a repository service and a central purchasing facility. In addition to the technical IT services, BITS has a small scientific applications team So that customers know exactly what BITS is costing, BITS splits its charges into service level agreement charges and standard charges.

Within units with BBSRC sponsorship, the availability of resources for curation of data, including genetic databases, is mixed. One scientist noted that there are databases of high value to the research community which need curation (in the form of annotation by expert scientists inter alia), but this has to be done by staff, already stretched, in their own time. The longer such curation work is left, the less useful the databases become.

On the science side, the Council for the Central Laboratory of the Research Councils (CCLRC) – one of Europe's largest multi-disciplinary research support organizations – operates several large-scale scientific facilities for the UK research and industrial community. This includes extensive data-centre provision in its role as a service provider to the academic community, at the Rutherford Appleton Laboratory and the Daresbury Laboratory. It is developing a curation capacity within this infrastructure. Its data centres hold many terabytes of data, also providing archive and back-up, for its own researchers and on behalf of researchers funded by other Research Councils, such as PPARC and EPSRC (which do not

have their own repository facilities). The advent of new instruments such as the Diamond synchrotron and the Large Hadron Collider project will require CCLRC to cater for tens of petabytes of data in the next three to four years, for which it is planning.

The **EPSRC** acts as a funding body and has no data centres of its own.

The **MRC** has local data centres within units, servicing local needs. These include several genetics units, such as the Human Genetics Unit in Edinburgh. This is the home of the Mouse Atlas, a curated functional genome database. This came on line in December 2002, so is a recent curated database arrival. It is currently funded by a five-year MRC programme, continuing to fund the database whose original development the MRC funded. Curation is carried out by a team of four, three scientists and one IT support person. Customers are the biological research community, expanding to the pharmaceutical sector; the database has been online since December 2002, and by February 2003 it was receiving some 380,000 "hits" per month (of which about 20% are believed to be from robots).

NERC has seven designated data centres, and a very large number of other data sets under its responsibility. It has a substantial budget for its data centres of £5 million per annum; it recovers some £2 million per annum through the provision of data and services outside the academic community. This budget exceeds that for many comparable organisations. As noted earlier, it also has responsibilities as a centre of legal deposit (for example, borehole cores obtained in the UK). Some NERC data centres' provision is supplied by third-party suppliers, such as high-performance computing centres. Data and data policy implementation is delegated to the seven designated data centres (the Antarctic Environmental Data Centre, British Atmospheric Data Centre, British Oceanographic Data Centre, National Geosciences Data Centre, National Water Archive, Environmental Information Centre, NERC Earth Observation Data Centre).

PPARC's involvement in "big" international projects has use of international facilities such as CERN. In terms of volume, particle physics already produce data sets of several hundred terabytes per annum. The experiments overtaking these data sets will come from the Large Hadron Collider ('LHC') in 2006, after which data sets of several petabytes per annum are expected from 2007. Astronomical data sets are also growing, with the advent of new telescopes taking volumes up to several hundred terabytes per year. Collaboration, national and international, is a major feature of both these areas.

Major e-Science applications being developed by PPARC-funded work are data grids and computational grids. The Astrogrid project, for example, is aimed at building a data grid for UK astronomy, forming the UK contribution to a global "Virtual Observatory". The long-term vision is one of a framework which allows data centres to provide competing and cooperating data services. Astrogrid will enable a consortium of UK data centres and software providers, pooling resources, storage and computing facilities. This was the one area we surveyed where actual storage represented a recognized major cost – unsurprisingly, given the volumes. (In general institutions and universities use computer clustering as opposed to "big iron" – very big, expensive computer storage boxes – to process and hold data.) It has been estimated that "data archival and processing costs at a "tier-1" LHC centre will be some £30 million over the next decade in the UK. It was noted that, while individual media costs fall, encouraging a continued shift away from the traditional bulk storage media (magnetic

tape, for example), this entails direct access solutions, which "will tend to slow the fall in the real cost of data storage"[18].

We found that e-Science project repositories are at an early stage, as the actual repository phases are placed at the latter end of projects, with the earlier phases concentrating on metadata and metadata schemas in particular, primarily in collaboration with CCLRC's Daresbury team. This was the case for example for the e-Science project based at the National Institute for Environmental e-Science in Cambridge. This e-Science project focuses on simulation data, including the archiving of that data. They have yet to reach the stage at which they need to consider where and how they would house the repository; they would like to see "a sort of "corporate" model, a national centre which would help them set up repositories".

Policies and practices

NERC has instituted policies to encourage sharing and curation, backed-up by a requirement for NERC funded researchers to offer datasets to an appropriate NERC data centre. The MRC has drafted a policy on data sharing and preservation; this will require grant applicants to set out their plan for the sharing and preservation of the data they will generate in their project. The ESRC has a policy of requiring grant-holders to offer data for deposit with the UKDA; the UKDA evaluates requests and accepts some 50% of them; if rejected there is no further obligation on the grant holder. The UKDA has an accessions committee which forms an opinion on criteria which include technical fitness for take-up (in which the cost of accession is also a factor), making a judgement as to future value and use. The AHDS applies similar criteria for potential acquisitions to its collections. (with copyright complications an additional criterion).

Table 3 summarises some of the practices and policies of two university-based data archiving organisations to whom we spoke:

[18] Professor Ian Halliday, paper supplied to authors.

Table 3: Practices and policies of AHDS and UKDA

Issue	AHDS	UKDA
Length of time data is kept.	Indefinitely	Indefinitely
Volumes	Terabytes A particularly large resource expected soon, with regular additions.	Terabytes
Dynamic datasets catered for?	Yes, a few. They are "problematic".	No. But would be intellectually interesting to explore this.
Metadata collection.	Determined by subject specialists in collaboration with depositors.	Determined by subject specialists in collaboration with depositors.
Data copies kept.	At least three. The original is always kept. A "migration copy" is made and to which preservation actions are performed. A dissemination version is used for delivery (not preserved).	At least four. One copy is stored off-site.
Software archived with data?	No. "A can of worms. Something for the Curation Centre?"	No.
Courseware?	No. But do hold some other learning resources.	No. "Likely to be a nightmare"
Consultancy services?	Yes	Yes
Advertise the services	Yes, in the designated community.	Yes, in the designated community.
Charges for access	No. Free at point of access.	Charges are made to commercial users, otherwise free for academic use.

4.2 Funding and costs - the repercussions

Table 4 summarizes the AHDS' and UKDA's funding position in 2002.

Table 4: Archive funding positions

	Annual budget	Number of staff	Current funding window	Funding sources
AHDS	£1.0 million	25	To 2005	AHRB (50%) JISC (50%)
UKDA	£1.5 million	50	To 2005	ESRC, JISC, University of Essex

Our analysis of costs in various repositories showed consistently that staff costs represent between 69% and 82% of the "total" – though we found that in many cases the "total" did not include all costs, as some infrastructure costs were absent, not visible because of cross-subsidies. Telecommunications represent a considerable portion of cost, and the largest item in BITS' standard charges.

Data centres noted that static funding over the period of their funding cycle (excepting inflation increases) is a problem, since the volume of work is increasing with continued uptake and more materials already held requiring preservation actions. It was suggested by more than one interviewee that the problem was really one of the academic funding structure overall, being based on the short term and inherently poorly guaranteed continuity, and the question could only be resolved at the DTI level. A project leader of a curated database would like to see a mechanism whereby funding for such databases can be sought on the basis that they provide a community service, rather than research. We discuss this at greater length in the following chapters.

The cost profile of any long-term data store is for a large peak at the time of accession, when decisions are made, data prepared, possibly reformatted, metadata added, indexing provided, and data loaded; further peaks are seen when data is subject to reappraisal and/or preservation actions; as indicated above, a further peak is found when data is deleted. (The profile is different for repositories where curation is continuous.)

Both the UKDA and AHDS reported that deleting information is more expensive than simply keeping it, since at the current time the cost of storage is low and deletion (changes to indices, metadata and databases) would incur systems and staff overheads. Until this balance changes, disposition will probably take the form of passive cessation of further curation or preservation work. Both organizations operate accession policies in which cost reduction is a factor.

For some services, their level of funding meant that they were unable to encourage further use of their holdings by a wider user community – a curation point we develop in the following chapter. The question was raised whether data centres should not apply a disclaimer when providing data, as units do not have enough staff at the moment to do validation of the quality of the data they receive. Another practitioner noted an unexpected problem: they have to handle blame for problems inherent in data sets which properly should be addressed to the data originators. The two last points highlight the fact that curation begins at data generation stage.

Recommendation B2: Clear terms of reference regarding the limits of data validation are needed for repositories.

University-based research projects see a high turn-over of staff, including during the life of projects. This is the nature of research – staff move on, post-graduates in particular. For data preservation and curation, however, this is an additional problem and risk, as these people take knowledge about data (content, context, technical) with them. Some areas are less vulnerable than others, where good practices are driven by habits or by the nature of the work they are doing (epidemiologists, astronomers for example). The loss of tacit knowledge was an area of particular concern for several interviewees, including the most senior. Nowhere did we meet any systems or procedures which addressed this problem.

Recommendation B3: Methods to capture tacit knowledge need to be researched and then introduced, particularly for staff moving off projects.

The short term affects provision for primary research data in other ways. Firstly, materially, in terms of continuity of provision. Another frequent manifestation was change of location of services, which was felt by users as disruptive and creating extra work during transitional periods.

Funding for data repositories is signed off in three- or five-year cycles. This is the case for example for the UK Data Archive, the NERC data centres (including centres of required data deposit), European Bioinformatics Institute databases (which receive half a million "hits" a week), or the Mouse Atlas database. As the managers of these facilities themselves all confirm, they are happy to be subject to regular performance or other review. Feedback from interviewees and questionnaire respondents, however, noted that the short-term nature of funding equals an uncertain future, which also affects staffing.

Recommendation A3: The mismatch of short-term funding against the long-term needs for data retention needs to be addressed by providing new specific, long-term funding stream(s) for data centres and curation, thus ensuring that there is a strategic approach to data stewardship which addresses holding information indefinitely, makes it widely available and encourages cross-disciplinary usage, including linking to other digital information.

4.3 Hybrid repositories

Many of the data centres are also repositories of physical samples – seed banks, geological cores, tissue samples, etc. While management of links between these holdings, paper archives and digital records was not regarded as a problem - and indeed, examples we saw were impressive - several interviewees stressed the importance of maintaining the links.

4.4 Digital Curation Centre

Our consultation with scientists and researchers firmly endorsed the need for generic support in the area of digital curation.

In early 2003 the go-ahead was given to establish during 2003 a Digital Curation Centre (DCC) to:

- Establish a vibrant research programme into the wider issues of data curation;
- Become an international centre for developing tools and techniques for long term, secure data curation;
- Develop a reliable, sustained repository of generic tools, software, and documentation, to support curation, preservation and use of digital resources;
- Develop testbeds and certification for systems, tools, and curation services;
- Pilot development of services for recording and monitoring file formats and preservation planning tools utilising these services;
- Establish a close relationship with JISC and Research Council funded services and repositories and the relevant HE community;
- Provide advisory services on curation 'best practice' and to be pro-active in raising awareness of curation issues;

For the first three years this will be jointly funded by JISC and from the e-Science programme. A business plan for continued financing of the centre beyond the initial three years is to be developed by the DCC. It is hoped that the DCC will provide common support services for data curation to the whole sector. Awards for the operation of the DCC will be the subject of a tender process during 2003.

Recommendation A5: Our findings endorse the need for the Digital Curation Centre and we recommend its establishment as part of a national provision for data curation in the UK.

4.5 Libraries

A central premise of the Follett report[19] a decade ago (1993) was the shift in emphasis in academic libraries, away from the idea that their essential role was provision of physical holdings and towards the concept that a library provides access to information, not necessarily within its own collections. There is a move towards access rather than holdings. One interesting example is the Edinburgh Engineering Virtual Library which "has demonstrated

[19] Follett, Sir Brian, 1993

the need for and value of structured information gateways to networked information"[20]; another in the commercial sector is GlaxoSmithKline's Research & Development organization, which replaced its physical libraries with virtual libraries.

Funded programmes followed the Follett report, with experiments for online journals, cataloguing programmes which sought to penetrate and describe scattered and under-used subject and special collections of national importance, all contributing to better disclosure of resources.[21] We note several parallels between the Follett libraries report and the digital curation situation. One example is the difficulty in distinguishing costs, another is diminution of local use of collections: when academic focus moves on and away from the topic area, the materials are more likely to languish. Local priorities change, but the importance of the materials generated there in the past may actually grow. For digital data, this represents risk.

There is an increasing trend for universities to bring IT and libraries under one umbrella. While this would perhaps tend universities to see digital curation as the responsibility of the merged, information services entity, responses did not indicate that the situation is or would be clear cut.

We heard conflicting opinions regarding the role of libraries (and the university libraries in particular) in curating primary research data, both views stated with emphasis. One view is that they are the natural heirs for data, and that their role should adapt to acquire necessary technical skills to enable them to become custodians of data; they were already skilled practitioners at organising information and disseminating it. The opposing view was that they are stuck in a mindset which does not make them able to adopt this role in the near future and they currently lack some of the requisite skills, such as IT, and that they may also lack the necessary imagination to change.

In practice the organisational focus of curation may be a local implementation depending on local circumstances and skills. We would argue that the role of researchers and discipline specialists remains at the heart of curation, but many different professional skills including those of librarians and IT providers will also need to be brought to bear on the curation process.

Journals

In many cases, primary research data relates to secondary and tertiary materials, often published papers. At the moment, these publications are generally one of the major sources of reference to the primary data. It is essential that links between primary and secondary and tertiary materials exist and are **persistent**. There is research work being done in this area, such as that by Professor Reagan Moore in San Diego and the work on unique digital object identifiers; this may be an area for the Digital Curation Centre, the JISC Information

[20] Peter Fox, Cambridge University Librarian, in his introduction to the new Betty and Gordon Moore Library in Cambridge, 2001.

[21] An interesting example is the Cairns (Co-operative Academic Retrieval Network for Scotland) project (cf Appendix 3). This is just one example of the many initiatives relevant to the digital curation field as examined.

Environment, and the Research Grid. It should be the curator's responsibility to ensure that these links are in place and are persistent, and that his collection has the tools and systems to support link persistence. At present, there are no consistent, standard or formalized links between publication and underlying datasets. For instance, the work that is being done on URIs[22] and DOIs[23] (uniform resource indicators, digital object identifiers) is relevant here.

This also has implications for the persistence of materials referencing the primary data, such as journals, and implying some responsibility in this respect on the part of the journals. Joint work will be required between the various parties.

Publishers are realizing the commercial benefits of "deep content databases". Elsevier, for example, has said that it intends to maintain electronically archived copies and has entered into an agreement with the Netherlands National Library[24]. However, commercial dictates do not always align with academic research priorities, with publishers applying different criteria in their management of archives.

4.6 Digital repository initiatives in the university sector

There are some initiatives emanating from larger universities in the USA which are examples of institutional libraries finding a role in digital curation; significant among these are the DSpace[25] project at the Massachusetts Institute of Technology (with Hewlett Packard), the CalTech Library System Digital Collections[26], and Lockss[27] (Lots of Copies Keeps Stuff Safe) at Stanford.

The DSpace system has adopted an open-source policy regarding the underlying software, and we have heard that over 2,000 copies of this have been downloaded. It is being implemented in the UK at Cambridge University, and is being widely promoted to higher education institutes, with questionnaire responses from other UK libraries and IT departments indicating interest. The Cambridge MIT Institute is involved in the DSpace research and development programme.

The DSpace system is simple, allowing basic metadata information to be collected. It supports standard format files. The aims for DSpace are ambitious but are not fully realised yet for much primary research data. However, in time, it may help achieve the deeper objectives of digital curation, by enhancing archiving, access and linking to other related categories of material such as e-theses and e-prints.

---

[22] See http://www.w3.org/Addressing/

[23] See http://www.doi.org/

[24] See: http://www.kb.nl/kb/pr/pers/pers2002/elsevier-en.html

[25] See http://www.dspace.org/

[26] See http://library.caltech.edu/digital/

[27] See http://lockss.stanford.edu/

4.7 Guidelines

We probed to see how far researchers were aware of guidelines to help them with the practical issues of coping with data retention and good data management. The results are summarised below in Table 5:

Table 5: Survey – Provision of guidance

Has guidance been provided on:	Yes	No	Don't know
Data preservation	44%	48%	8%
Records management	33%	54%	12%
Good data management	44%	50%	6%

Somewhat under 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.

Recommendation A8: Educational materials, guidelines and policy documents for researchers need to be developed and publicised.

Provision of guidance materials shows some interesting points in relation to some of the other questions. Comparing responses on contractual obligations to keep data (see Figure 5 in section 4.14) with those on provision of guidance on preservation shows that, where there is an obligation to keep data, some 32% of these report having received no guidance on preservation. One needs to interpret this cautiously in view of small numbers.

Of the respondents who reported receiving guidance, they quoted as sources of advice (in order of frequency) their institution, their (host) IT department or its support staff, and then the funding Research Council.

Where a research council does have guidelines for its staff, setting out policies and procedures to follow, these were not always seen as helpful.

> " …. 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."
> *Questionnaire response*

The Royal Statistical Society has issued guidelines for researchers on the preservation and sharing of statistical materials, working in collaboration with the UK DA[28]. This booklet contains a code of best practice, and provides useful lists of resources.

4.8 Standards

The use of standards is clearly important for data curation. These are numerous, and span a wide range of issues (e.g. metadata, archival systems, data interchange, records management, file formats, etc.) and are variously applicable to wide or narrow ranges of disciplines. There are a growing number of metadata standards for specific fields (an example is ISO 19115[29] for geospatial data). Notable for wide applicability is the Dublin Core metadata set, asserting its currency for resource discovery in the library and information management community, as confirmed during this study with the announcement of its status as an ISO standard[30]. The Dublin Core metadata set is basic; CCLRC's generalized metadata project in development for the scientific community involves substantially more extensive datasets. The Open Archival Information System (OAIS) reference model for digital archives[31] also acquired ISO status during preparation of this report. Questionnaire responses revealed a total lack of knowledge of OAIS and almost total lack of knowledge of metadata tools such as the Dublin Core.

There are useful guidelines and standards in the commercial sector, notably in the form of the engineering sector's STEP standard[32] and the US Food & Drug Administration's 21CFR Part 11 regulation, which is accompanied by guidance documents relating to the retention of digital records[33]. We also encountered a total lack of awareness of these during our study.

Recommendation A9:	There should be increased investment, knowledge transfer, and cross-sector partnerships with knowledge-based and science and engineering industries to capitalize on UK expertise in data curation. This should be led by the DTI.

4.9 Preservation

Since Professor Lievesley's 1988 report[34], the Digital Preservation Coalition[35] (DPC) has done considerable work towards achieving the first of the objectives she identified. Indeed,

---

[28] Royal Statistical Society, 2002. (with the UK Data Archive)

[29] ISO 19115, 2003, Geographic information – metadata.

[30] ISO 15836, 2003, The Dublin Core Metadata Element Set

[31] ISO DIS 14721, 2003 (OAIS: Reference Model for an Open Archival Information System)

[32] ISO 10303, 1994, and some 90 related standards for the computer-interpretable representation and exchange of product data

[33] Food and Drug Administration, 1997 and 2002, currently being revised and to be reissued in 2004.

[34] Lievesley, D., and Jones, S. ibid.

[35] See: http://www.dpconline.org/graphics/

more widely, activity in the digital preservation and records management fields has expanded massively, to the extent that there are now so many initiatives in the public sector that the territory is hard to navigate.

One highlight of current preservation work was celebrated in Leeds University in early December 2002 with the demonstration of the rescue of the 1986 Digital Domesday Disk through emulation[36]. The BBC Domesday project involved fairly complex digital objects for its time. There is an increasing body of practice in digital preservation, particularly for the more simple types of data. However, for the more complex data types there are few, if any, acknowledged examples.

The UK Data Archive at The University of Essex expressed the view that no-one has any real experience yet with preserving multi-media formats. Difficult formats are not addressed, nor is dynamic information.

Methods for long-term preservation of digital data aim to preserve content (information) and systems (applications) behaviours over time as successive hardware and software technologies to read and interpret bit-streams become obsolete. There are variants within each option, but they may be summarised as follows:

- **Migration**: This requires transforming data from one format to another successively as technologies change. This is a well understood process and occurs when systems are upgraded. Except for the simplest data structures and/or over short timescales, it is likely to result in information loss, and/or changes in systems' behaviours or computed results. It can be expensive and time-consuming to perform, and may rely on expert knowledge which may no longer be available when needed. Costs are recurring and errors are cumulative (and may not be detectable). It is sometimes referred to as conversion.

- **Emulation**: This entails keeping the original data and application software and creating programs as and when needed which emulate the behaviours of successive computer systems, thus enabling the original application and data to be processed - emulated - on contemporary architectures. This may prove more cost-effective than migration, and promises more faithful preservation of both content and behaviours. The work of Rothenberg[37] and the CAMiLEON[38] project provide examples of research in this area.

- **Formal descriptions**: The use of a Universal Virtual Computer (UVC) has been proposed by Raymond Lorie of IBM[39]. The behaviours of the original application are encoded at the originating time in a format which can be understood by the UVC in the future; the abstract UVC is designed so that a real, functioning instance of it will

[36] See: http://www.si.umich.edu/CAMILEON/domesday/domesday.html

[37] Rothenberg, J., 1995, 1999, 2000

[38] See: http://www.si.umich.edu/CAMILEON/

[39] Lorie, R., 2001, 2002

be easy to create in the future and which will be able to emulate the original application on contemporary architectures. This method is still in development.

Two other strategies have proponents, but have limited use in our context:

- **Digital archaeology**: Analogous to the recovery of physical artefacts, it involves recovery in the future on an as needed or exploratory basis. It transfers cost to the future at considerable risk of loss or future misinterpretation.
- **Computer museums**: This strategy proposes to archive whole systems, including hardware and systems software, so that they can be used in the future. Continuing costs, dwindling available expertise and physical decay of hardware will limit this approach. (However, it was, essentially, one of the suggestions from the USA's Food and Drug Administration original guidance to 21 CFR Part 11 – guidance which has since been withdrawn for review; at least one commercial company has proposed providing this as a service.)

It is likely no single approach (or some method yet to be discovered) will dominate; the outcome will depend on the material to be preserved, the degree of technical success achieved and on economic and organisational factors. Some common-sense rules are generally agreed – the original data bit streams should be preserved alongside any "preservation versions" which are also kept; the existence of high quality metadata and documentation from the original research greatly enhances its re-use and preservation; the use of non-proprietary, well-documented data format standards like ASCII/UNICODE and XML increases the chance of future recoverability. All methods assume application of good data management practices, and implementation of secure storage, and institutional or organisational continuity.

Our surveys revealed general ignorance of the digital obsolescence problem (and thus of the need to apply these techniques in the first place). On the coal face in university and institutions, awareness of the digital preservation as an issue was low, though not entirely non-existent. There were several instances where researchers and managers were confronting preservation or management issues. Few departments made use of potentially relevant guidelines such as those prepared by Cedars[40]. At one unit, scientists were daunted by the problem of migration of data to the extent that it was inhibiting their reliance on data archives. This unit was unaware of the preservation work cited here.

4.10 Heterogeneity and categories of data

Our questionnaires confirmed some of our fears regarding the ease with which mainstream research data can be preserved. Heterogeneity increases the preservation burden. As noted above, the use of non-proprietary, internationally recognised data format standards is likely to promote easier interoperability and preservation, not to mention lower costs in the future. Promoting these would be valuable, particularly given indications that use of them is patchy.

[40] Cedars, 2002a, b

When asked what would be required to use their data in the future (Table 6), a disturbingly high number of respondents report using special software whose longevity is questionable.

Table 6: Survey - Requirements to use data in the future

Future requirement to use data	Yes (%)	No (%)
Special software?	42	58
Special hardware/instrumentation?	15	85
Explanatory documentation?	75	25

46% of respondents had written their own software for their projects, which further increases the risk of future inaccessibility.

> "End users will not require these [software systems] 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."
> *Questionnaire response*

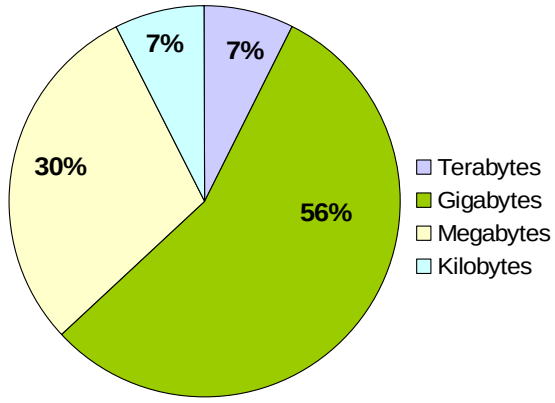
Asked what are the main types of commercial or open source software used there was a surprising variety of response – from a sample of 29 replies some 49 different systems were mentioned, most of them just once (this counts a reply such as "Microsoft Office" as one system). Of those mentioned more than once, MatLab was mentioned seven times and Fortran (sic) and MS-Office both four times.

As might be expected, there were many different responses to the question, what are the principal data formats for these data – 25 different formats were mentioned by 34 people. There was rather more agreement on formats than for software systems; formats mentioned more than once were as shown in Table 7.

Table 7: Survey – Data formats used

Format	No.
ASCII	**20**
TIFF	**11**
Excel (.xls)	**9**
Access (.mdb)	**3**
HTML	**3**
FITs	**2**
XML	**2**

More reassuring at first glance was the finding that a minority of projects (22%) span time periods over which preservation issues may arise during the project itself, set at a somewhat low period of five years. Nevertheless, this is still a substantial proportion. These datasets pose different management and access problems which we discuss in chapter 5.



Our survey investigated the amounts of data now being generated per annum by individual researchers; the ranges of data volumes are shown in Figure 3. Among those who responded, gigabyte ranges were by far the most common. It was interesting that a number of people could not tell us how much data they were generating.

Figure 3:
Survey - Data volumes generated per annum

Risk can be managed better –it was noted there is often a vast quantity of poorly managed data in research environments which could represent liabilities (such as following from loss or inappropriate access). At one major university the estimated total volume of data at risk not backed up was put at one petabyte.

Our questionnaire replies showed a surprising percentage of researchers claiming their data contained **confidential** information, as shown in Table 8. We cannot be sure about the interpretation made of "confidential" by respondents – whether this meant confidential information about third parties or information they wished to be kept in confidence.

Table 8: Survey - Confidentiality of data

	Yes	No
Does your project data contain confidential information?	56%	44%
Will that data remain confidential after the project?	42%	58%

Some 25% of these respondents felt that some restrictions should be placed on future use of their data after project end. The following is a representative sample of the reasons or circumstances given to support this answer:

> "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 [source]."

> "Storage in searchable database. Some projects confidential, others not."

> "Probably Clinical Governance rules."

> "Patenting may be appropriate."

> "Only access for academic purposes."
>
> "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."
>
> "Yes but: Some projects funded by industry will be confidential. Data from other projects will be confidential until the IP position has been determined. […] 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. […] Data should be actively assessed by a Project Team and archived by the Project Manager."

One respondent in the field of clinical research is currently working on a confidentiality project to meet new EU guidelines relating to patient data.

Static/dynamic: Our sample of researchers showed that most data is static, though a sizeable minority (35%) were producing dynamic information (Figure 4).

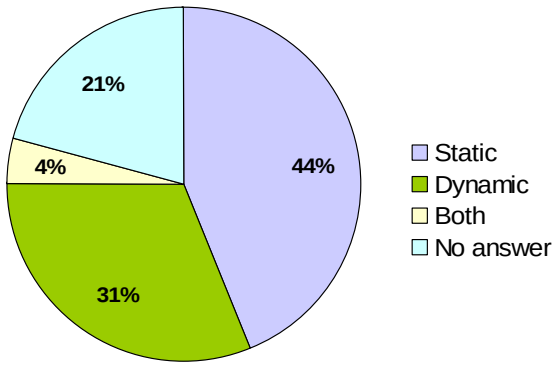


Figure 4: Survey - Data categorisation - static/dynamic

Medical and biology researchers drew attention to the need to distinguish between "**dirty" data** and clean, **validated data**. Again, in the medical field a distinction was made between epidemiological data and data from those working at the bench. This distinction has informed the formulation of the MRC's policies. In general terms "e-scientists" create larger volumes, the data is less "clean" and is often dynamic. Epidemiological data is usually the result of investment in getting it clean up-front, and the originators feel more ownership in it. These considerations will influence curation of the datasets.

Data nowadays is rarely verified when it is entered; we have all become amateur data entry typists and proof readers or we rely on automatic data collection devices. This point was made by senior computer scientists with many years' experience working with scientific data issues, noting that data validation was a general problem throughout science, except in certain areas (like high energy physics). Another drew attention to the policies of the Mayo Clinic in the USA: the clinic attributes its success and excellence in large part to a decision made in the early 1900s to base its work on the good management of records. Since that date that "they have only lost some 380 records".

Without exception, data centre managers we surveyed stated that while they meet demand requirements, they have insufficient time to validate the quality of the data submitted. Of course, this may not apply to all data managers.

Taken as a whole these findings do not present a reassuring picture for future preservation and re-use.

4.11 Metadata

At present, data is stored in many file systems and databases, with no common way of accessing or searching the records. There is a lot of activity globally to establish metadata frameworks and these will assist the future scientist to locate, retrieve and make use of data. This work is taking place at a general level, concerning itself with access issues and the re-use of common office document files (examples are the Dublin Core and from Cedars, both referred to above).

There is also work being undertaken to satisfy the specific needs of science and different specialities. CCLRC is developing a web-based portal with the aim of offering a single method of searching the CCLRC resources. Interaction with the metadata catalogues which they are developing is based on a metadata model for representing scientific data being developed by CCLRC. This important work is being conducted in collaboration with other European institutes which work with large volumes of scientific data. This feature itself is of considerable importance, given the international nature and scope of research work and the need for common standards and terminology.

One of CCLRC's key requirements for its model is generality. It aims to be a high-level generic model which can be adapted to scientific disciplines.

CCLRC's Daresbury Laboratory development work involves collaboration with several e-Science projects. One is the NERC DataGrid, which aims to allow users to discover and access data without having to know storage details, values or parameters:

> "[The NERC Datagrid aims] to build a grid which makes data discovery, delivery and use much easier than it is now, facilitating better use of the existing investment in the curation and maintenance of quality data archives. Further we intend to make the connection between data held in managed archives and data held by individual research groups seamless in such a way that the same tools can be used to compare and manipulate data from both sources. What will be completely new will be the ability to compare and contrast data from an extensive range of (US, European, UK, NERC) datasets from within one specific context."[41]

[41] NERC DATAGRID project abstract, January 2002

4.12 Ontologies, mark-up languages

> "The "semantic Web" is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is the idea of having data on the Web defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications. The Web can reach its full potential if it becomes a place where data can be shared and processed by automated tools as well as by people."

"Semantic inter-operability" is key to this, and "… the notion of the *Semantic Grid*, [applies] Semantic Web technologies in Grid computing developments, from the machinery of the Grid infrastructure (such as the Grid services of the Globus Toolkit) up to the Grid applications. It is important to note that the 'semantics' permeates the full vertical extent of the Grid and is not just a semantic (or knowledge) layer on top: it is semantics in, on and for the Grid."[42]

"Ontologies" fall into this context, their purpose being to support the sharing and re-use of data; they are a set of concepts - such as things, events, and relations - that are specified in some way (such as specific natural language) in order to create an agreed-upon vocabulary for exchanging information. Several interviewees referred to the need for ontologies, though not necessarily using these terms (taxonomies were also mentioned). This is an active area of research. A considerable amount of work is being undertaken into developing ontologies and tools to help them, notably the OIL[43] and DAML[44] initiatives. The UK is making a significant contribution to this work.

It was interesting that more than once researchers involved in e-Science projects remarked that they had moved more into work involving domain-specific mark-up languages, with work on chemical mark-up languages mentioned more than once. These are usually based on XML. We noted that there is a tendency in this area (whether in a commercial or academic context) for different strains of a mark-up language to develop or to be mooted, with the result that more than one version of a mark-up language can be in circulation for a while before one particular strain establishes its single currency.

4.13 Users - awareness

Our questionnaires revealed insights into aspects of use of facilities.

We asked researchers who they thought would look after their data after the end of their project(s). The answers showed that there is very little formal provision, or awareness of provision. The question produced a wide variety of responses, as shown in Table 9.

---

[42] Professor David De Roure, from www.semanticgrid.org

[43] See: http://www.ontoknowledge.org/oil/

[44] See: http://www.daml.org/

Table 9: Survey – who will look after data?

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

With regard to these services we heard suggestions that where the centre for deposit was outside the originator's own institution there was sometimes reluctance to pass on data.

In the universities, some interviewees felt that among PhD students attitudes in this respect were poor - "students do not care about the fate of their data after award of their PhD".

An interesting comment was made that digital science was seen as a threat by some communities: libraries, university computer centres, some researchers and some in the computer science community. If present to a significant degree it could significantly impede progress.

Recommendation B4: The responsibilities of the various parties to the curation process should be articulated and communicated to researchers and scientists and those responsible for curation.

As noted above in regard to metadata collection, curation and preservation as currently practised require input by the creator to prepare data for future curation; often their input is also needed later as the data undergoes curation and preservation processes. To get the most from e-Research also demands a culture of providing resource (including time) to make this possible, and of data sharing. Interviewees stressed the need to change the culture of the research environment in order to engage and motivate data creators in these activities. Some experience suggests that depositors, when interested in keeping data, are more interested in immediate access and usage, and not in the preservation issues, denying this is even an issue. The value of keeping the data is often not apparent to its creator, nor does that value usually revert to its creator. This also applies to the corporate sector.

In a few cases, funding bodies make submission of data a requirement, but this is not well policed. The general view of those we consulted, however, is that sticks are less effective than carrots - people must want to provide their primary research data and be given incentives to undertake the curation work which benefits the wider research community rather than the individual data creators themselves directly.

Recommendation A7:	A programme of activities, both national and international, should be initiated to promote incentives which will foster a scientific culture of engagement in data care.
Recommendation B5:	Where conditions in grants for research stipulate that data should be cared for after the project end, methods to track compliance to these conditions should be introduced.

4.14 Users - funding, incentives, cultural

In general researchers themselves see continuing value in their data, (though not quite to the same extent as they do for the final publication), as Table 10 from our survey shows, when asked which of their data would have continuing value.

Table 10: Survey – Value in data types

Data type	Yes	No
Primary data	79%	21%
Summary / derived data	90%	10%
Published data	94%	6%

As Figure 5 shows, some 46% of respondents are working under terms of funding which require their data to be archived. However, when asked whether financial provision had been made for archiving their data, only 21% answered with a definite "yes" – see Figure 6.

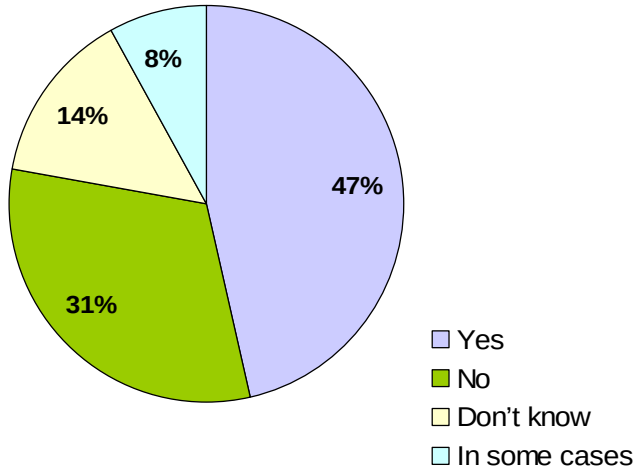


Figure 5:
Survey - Does funding stipulate archiving?

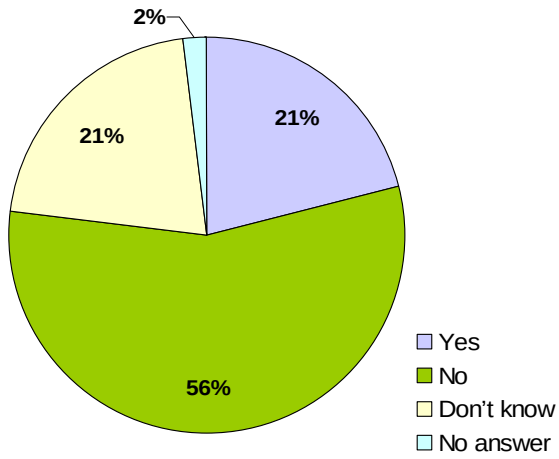


Figure 6:
Survey – Financial provision for archiving?

The implication of these figures is that though 47% these respondents are under an obligation to archive their data, only 21% report definitely that financial provision has been made to look after it after project closure. See Figure 7.

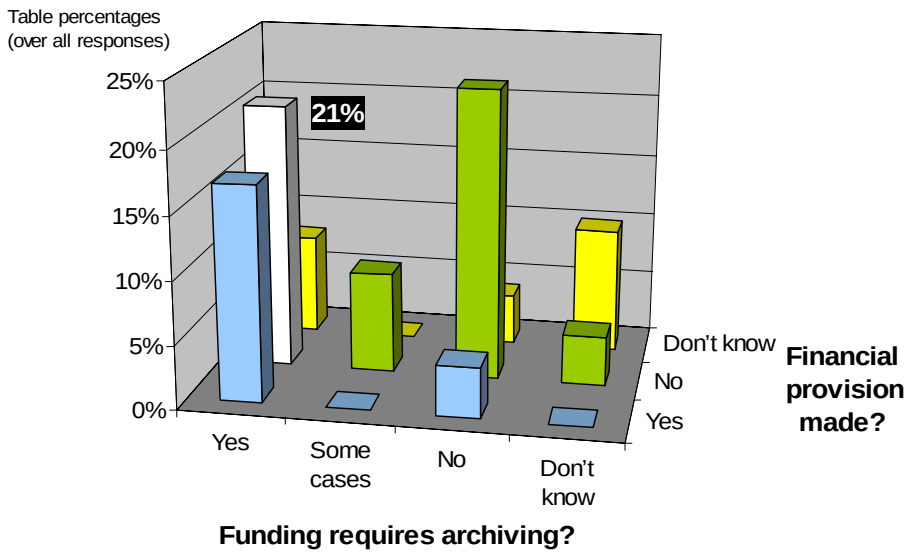


Figure 7: Survey - Archiving requirement vs funding

A more detailed look reveals that where data should be kept by the funding contract, 45% say they do not have financial provision to do this. These figures are based on small numbers, but if they represent in any way the true situation there is a somewhat disturbing mismatch between directives and resources to achieve them. This mismatch was confirmed by feedback from our interviews.

4.15 Commercial

It was noted that at the DTI there is a high degree of motivation in promoting the commercial interests of UK plc through the e-Science agenda - a requirement of DTI funding is that projects must obtain matching funding from the commercial sector. Various interviewees noted the enthusiastic engagement of commercial companies. IBM and Oracle in particular are engaged in Grid technologies, with many collaborative initiatives. At the end of January 2003 IBM made an announcement about developing commercial grid software with Globus (on OGSA, the open grid services architecture). Another multinational company active here is HP, which has been supporting the DSpace work at MIT, the electronic library development mentioned above. Sun is also involved in Grid technology and e-Science projects.

There was some criticism of companies for not providing sufficient support to the research community within the e-Science programme. Comments implied they would like to see more active engagement, and perhaps a bit more cooperation with the pilot projects. Smaller, specialist companies are engaged in this work, including those working on the Grid infrastructure.

A recent study[45] for the Digital Preservation Coalition confirmed a fairly low level of understanding of, and engagement in, digital preservation within the commercial sector at that time, and we would extrapolate this as being certainly true for the wider curation and archiving questions. As digital content continues to increase and customers become more aware of the longer-term curation issues, we might expect this position to change.

[45] Lord, 2002

5 Future requirements

In this chapter we discuss the definition and future provision of curation. First our view of the evolving nature of curation is presented; then consideration of the acquisition of data, curation activities once the data has been acquired, organisational questions, and some management proposals. Funding and intellectual property questions are discussed in chapters 6 and 7 respectively.

For our analysis and presentation of future requirements for digital curation for primary research data, we began with the following premise: That the objective of digital curation of primary research data is to keep data which is valuable, potentially valuable or which is required to be kept; and in such a way that it is accessible and usable by others (while observing relevant restrictions), that its value is maintained and, where possible, enhanced; and that this activity and service should be provided at affordable and justifiable cost.

5.1 The evolving curation picture

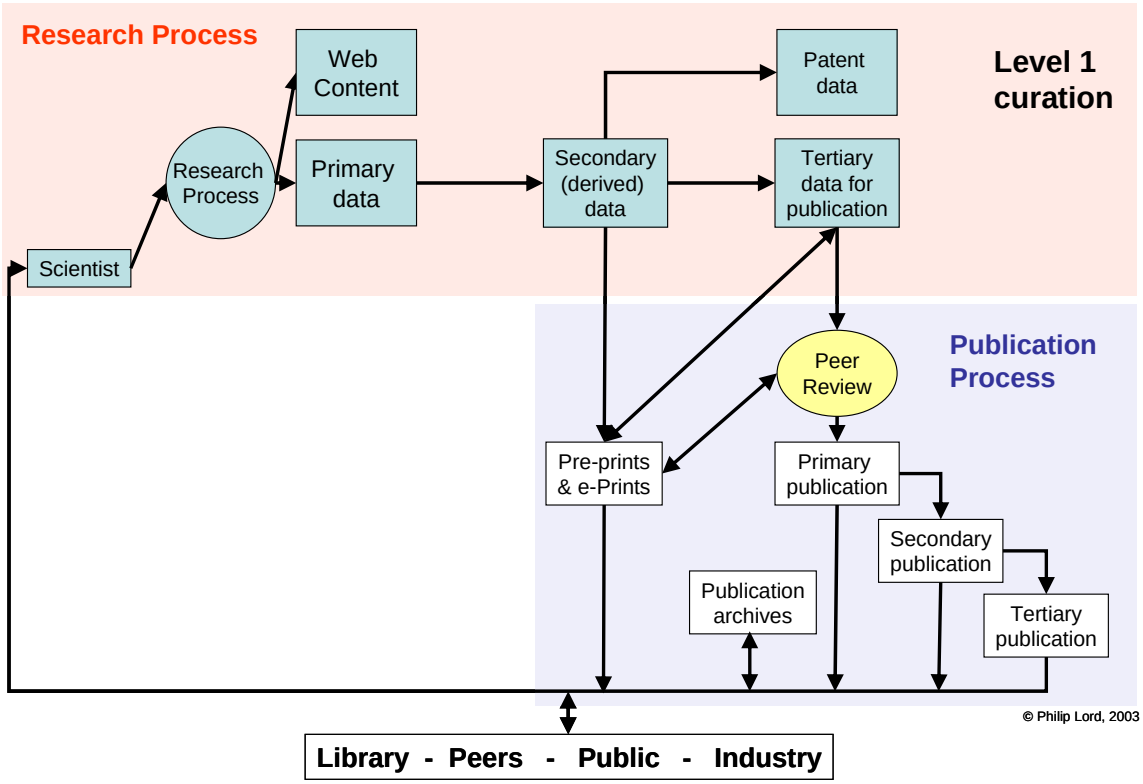
In the course of writing this paper we developed a model of the current research and publication process, and from that extended it by adding first a layer for data archiving processes, and then a layer for "full" data curation activity. The model shows the evolution of flows of information from the scientist working at the research process, and its cycling back via increasingly diverse channels to inform further research. The evolution is represented here by a series of diagrams. Any such model must be a simplification, but we believe it does draw out the main elements of processes as they will apply to scientists in their role as data generators and consumers, and in doing so it encapsulates some of the changes within research. It is only presented in outline here and could be elaborated further.

It should be noted that digital rather than analogue data now dominates this whole cycle.

Figure 8 below shows our model of the traditional research process, tracing the movement and refinement of information from the original scientist through research activity which produces primary, raw data that is analysed to create secondary, results data; this is then evaluated, refined to be reported as tertiary information for publication. With the mediation of the pre-print and peer review mechanisms, this then goes into the traditional publishing process (whether in paper of electronic form) where primary research publications feed secondary publications (bibliographic and abstracting reference services), both in turn feeding tertiary publications such as reference works. These then feed publication archives. Information from the publication process cycles back through to the scientist and into research effort. It is also made available to a wider community of consumers: libraries, the public, industry and (international) peers. (Secondary data also feeds the patent application process, but this is not considered further here.) It should be noted that even at this level primary research data may need to be retained with minimal access, solely for the purposes of meeting legal retention periods or professional academic requirements (i.e. validation of research or professional conduct). Our findings indicate that this may not always be the case.

We call this minimal process **Level One Curation.**

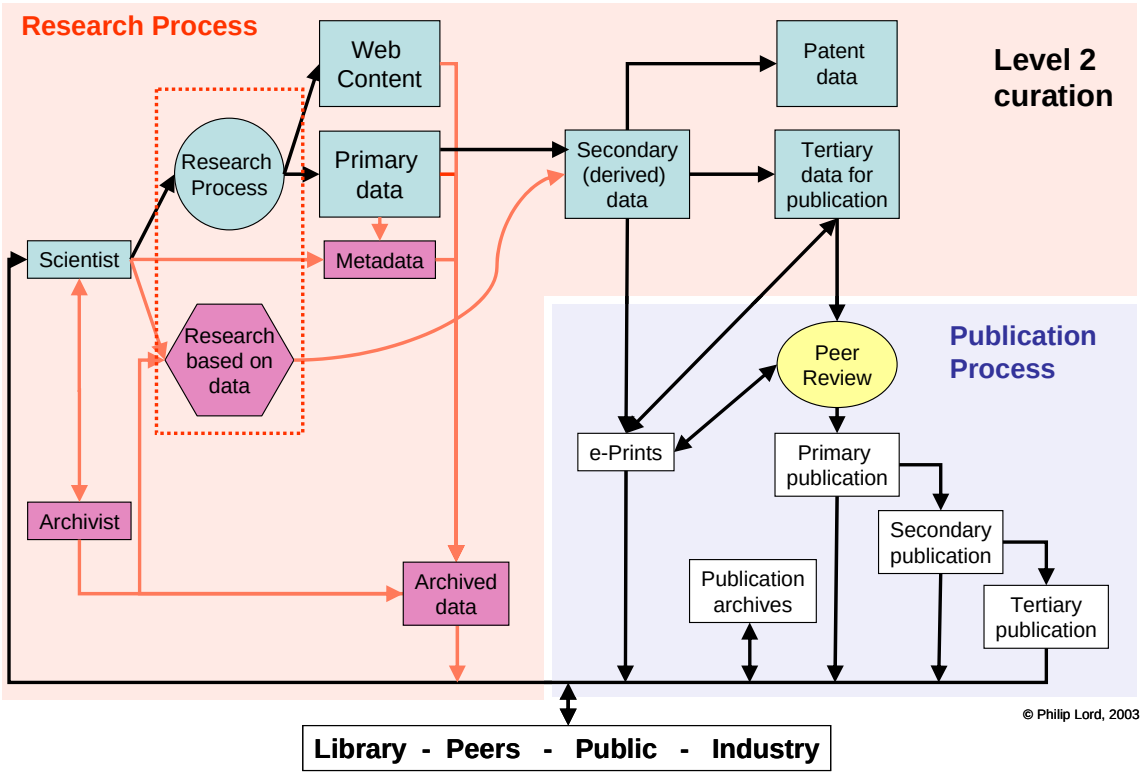
Figure 8: Traditional academic information flow – Level One Curation



In some areas, there has been a move to archive primary and secondary data under professional supervision. If we add a **data** archiving and preservation process to this model we then get a more complex picture for both the scientist and information consumers. The red lines in Figure 9 below show, in outline, the additional data flows and processes which are brought in, and also show the added ability to perform data-based research alongside traditional research, i.e. using data to make new discoveries or to obtain further insights which in themselves have value and which can feed the whole information cycle in their own right. This enhanced picture with data archiving we term **Level Two Curation**.

Another role comes into the picture, the data archivist. People in this role in general need to interact with the data generator to prepare data for archiving (such as generating metadata which will ensure that the data can be found, and can be rendered or used in the future). The role of the archivist is close to that of a curator – we have used the term "archivist" to emphasise the distinction with the next level. The dotted red box (Figure 9, overleaf) shows the enlarged research context. For the scientist the added roles of data preparation and the opportunity to do data-based science mean that new skills will be needed, with concomitant needs for training, motivation and reward. For consumers of data the assistance of archivists may be required in the context of future access, data rendering or re-use.

Figure 9: Information flow with data archiving – Level Two Curation

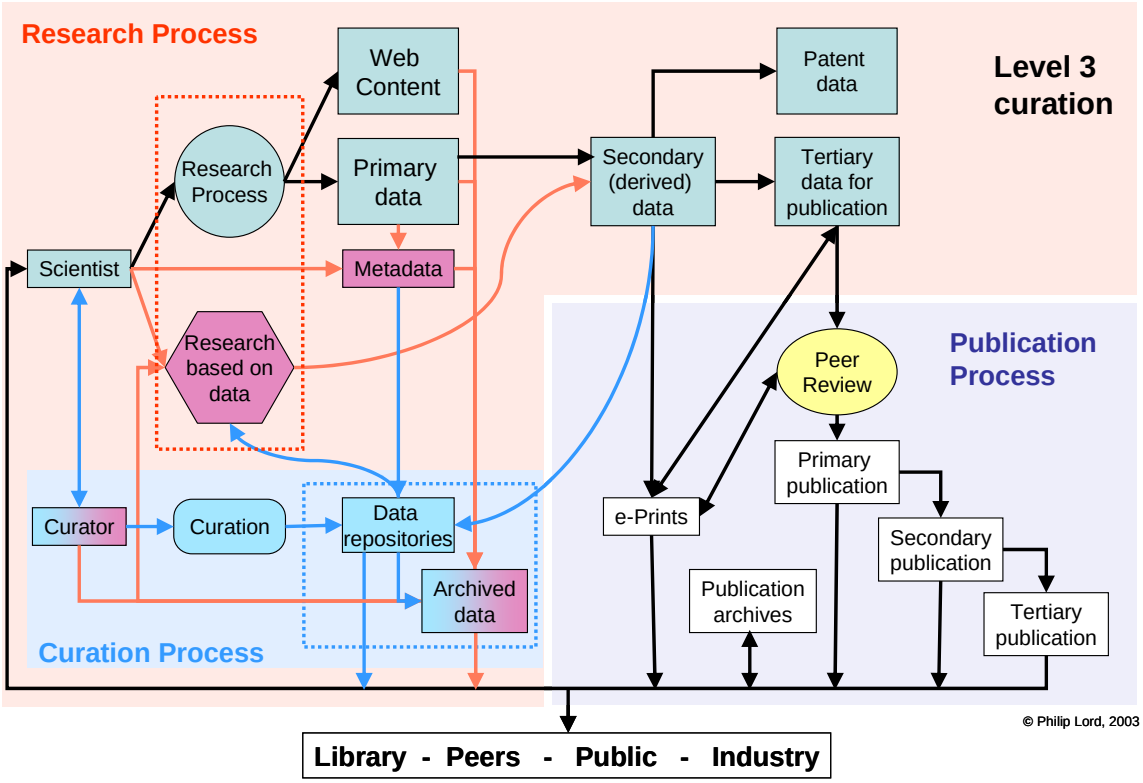


In circumstances where the data being created and is kept dynamic with annotation and linking to other research, we can envisage a further level of active curation added to this picture, as shown in Figure 10 below, indicated in blue.

In this case the information resources available to the scientist and the external world are further enhanced, at the expense of an even more complex environment for data consumers and data generators alike. The role of archivist becomes more active than in Level 1 – and we now substitute the term curator. The role implies a continuous review of information in their care, though they may still retain archival responsibilities; further requirements around this role are discussed below. Note that curation as a process is largely separated from mainstream research activities, though the degree to which this is the case will vary according to circumstances and domain. We call this **Level Three Curation**.

Curation at this level is usually associated with longer-term and open-ended data collection projects and dynamic databases, as described in section 4.1. In this model the curated data repositories and data archives may be considered as one (dotted green box), though this need not always be the case (for example, snapshots of the curated repository may be made and archived separately). The relationship between constantly curated, evolving datasets and those in static digital archives is one which still needs to be explored, through research and accumulation of practical experience. It is possible to envisage different degrees of curation within this level (Levels 3.1, 3.2, . .); we do not attempt to define all these possibilities here.

Figure 10: Information flow with data curation – Level Three Curation



All levels of curation will coexist, as implied by Figure 10.

The term "curation"

Until now we have used a working definition of "curation" (see page 11). Using this model it is now possible to elaborate further on the role covered by the term. The use of the word "curation" in this context is recent. As with any term, it is developing its meaning organically. Terminology in the fields of digital archiving, records management and now here, digital curation, is not yet stable, and one of our recommendations (B7) is for work to be done to agree territory and scope of meaning of these terms, to provide an important foundation for more efficient working and communication.

The term "curation" builds on our understanding of the word "curator", somebody who keeps something for the public good, whose value often needs to be brought out by the curator. There are two points to be made here. Firstly, this open context implies more support for explicit policies with regard to data sharing, and it has major implications for structuring and tools. Secondly, the digital curator is store-keeper but he is also closely linked to promoting new science, making sure that his user-base is solid, sufficient, and looking forward to identify new ways to serve present and future researchers. The digital curator should take an active role in promoting and adding value to his holdings, hold exhibitions, run joint events; he should manage the value of his collection.

Recommendation B6: Data curators should be closely linked to discipline research; their role should include active promotion of their holdings as well as acting as store-keepers and service providers.

The term "primary research data" points to something beyond. This is key to the value of the curated holdings: they should be visible through catalogues and national or international indices (virtual or otherwise), they should be well-documented and associated to related material, able to link to publications, pre-prints and other subsequent material. For some data, curation involves annotation. We believe that annotation is also relevant for curation of other data, and technology may facilitate links to relevant annotation which accrues over time, to create a growing knowledge base. UK research curation holdings should enable cross-searching and mining. Work on metadata, ontologies and the semantic grid is important, as is communication between these areas.

The term "curation" is now commonly used to refer to the work done on genomic and proteomic databases, annotating and managing annotations. The use of the term in this context extends this meaning: it covers a wider context than just archiving; it embraces the care of the record within scientific context and environment. This is particularly relevant for primary research data which, as the term implies, is part of an ever-widening chain – indeed, a chain which increasingly is a cycle.

Recommendation B7: An agreed ontology of digital archiving, preservation and curation needs to be developed; this could become part of the research programme (see also A10).

In the discussion that follows all levels of curation are discussed – reference will be made to level as necessary.

5.2 Data acquisition, planning, selection, and enhancement

The first requirement for curation is for data itself to be available to go into a repository. For this to happen, there must be an administrative framework, to provide mechanisms and channels, and there must also be a compliant cultural environment, to engage support from those involved. The cultural aspects we discuss in section 5.3 below, where we also cover grant conditions.

Planning

Repositories and curation centres would be able to plan a few years ahead to a certain extent by receiving information about new research projects at the inception of these projects, or earlier. Basic information is already available, and we believe it is possible that adjustments in existing systems could enable this.

One data centre interviewed emphasised they would like to be more pro-active, talking to their community, and leading in the development of standards and support for research projects. However, they don't know what data is coming when, in particular from non-

thematic programme research. There is a box on grant application forms for those generating research data to provide information about their data, but data centres are not sent this information, so they are unable to see what data is likely to be generated, whether it is important, or what support they might be able to provide to help the research teams.

In this context it is interesting to note the currently unique relationship that exists between the Arts and Humanities Research Board as a research funder and the Arts and Humanities Data Service as its nominated centre for data deposit. Under the terms of their joint C&IT Policy (computers and information technology), staff from the AHDS are involved at the earliest stage in assisting with technical appraisal of grant proposals (for example, data creation and archiving arrangements), and providing information and support for grant applicants and grant holders.

The Medical Research Council's Data Sharing and Preservation policy, currently in its consultation phase, will also require all grant applicants to set out a data sharing and preservation plan at application stage. ESRC and NERC already have policies in place, as stated earlier.

Possibly these are models which could be adopted more widely, and we suggest that a similar requirement be made for all grant applications. This information will (i) allow data centres to take steps to ensure potentially important data reaches them in due course, (ii) assess what data curation support might be required from project start, (iii) provide data centres with information for medium-term forward planning, and (iv) would increase awareness on the part of and curation support for the researcher.

Recommendation B8: Tools need to be developed to enable forward planning in repositories, and appropriate communication mechanisms should be established between funding agencies, researchers, and repositories.

Selection and enhancement

Not all primary research data needs to be retained or has long-term value. The reasons for retaining data and its potential value for generating new research will vary. The selection of datasets for retention, and the level of investment required in the curation of datasets, therefore needs to be identified and graduated accordingly.

Curation inevitably requires choices to be made, and thorough analysis is needed of:

- the reasons for selecting a particular dataset for retention;
- which categories of data have long-term value;
- the length of time data should be retained;
- and the appropriate level of investment that should be made in its curation.

We have set out in section 3.1 some of the primary reasons likely to be considered for keeping primary research data. The practicalities of maintaining access and preservation and adding value to data dictate that appraisal[46] is needed to focus limited resources to the best effect. Appraisal is not a straightforward issue or necessarily fixed at one point in time.

We cannot always know what is going to be of value in the future. However we may be able identify categories of data which do not need to be retained. A few useful ways of categorising data are summarised here: more are likely to exist and this should be an area for further research:

- **Reproducible data** –data which can be regenerated if needs must at acceptable cost.
- Data which is **not reproducible**, or only reproduced at very high cost, usually observational data.

An interviewee noted that the above tend to be **hypothesis-driven** and **descriptive** respectively. It was also noted that the volume of the former hypothesis-driven data currently tends to be greater than the descriptive, though this is changing as new recording and research techniques are employed.

Further clear, systematic criteria need to be established for data to be appraised. Some of these may vary according to discipline. We recommend that work be undertaken in this area, looking at simple check-lists - such as that suggested during our interviews for epidemiological datasets, shown in Table 11 below. We would also recommend investigating more sophisticated prediction algorithms.

Table 11: Epidemiological data set - possible retention criteria:

Criterion:
The nature of the questions being asked by the study
Whether it addresses only one question or many
Whether the question has been answered before
The richness of the data set
If it is a longitudinal study – "indicates an amber light"
Sample-related studies
Stability of the measures used
Possibility to go back to the population (e.g. for consent, ethical committee access.)
Uniqueness; value for possible future comparisons

[46] In the archiving community the term "appraisal" is used to denote the selection process and processes to determine where to keep information and for how long. It is being used here in a broader sense which covers other decisions and processes relating to added value and curation levels. In a digital environment this could be an iterative process over time.

Another interviewee looked at curation from a temporal point of view with similarities to management of computer storage and access: the older the data becomes, the "further the data may be put into remote storage" – perhaps following a path of lessening accessibility and investment as it ages and it is referred to less and less.

For data that should be retained, we set out above (5.1) different levels of curation with increasing levels of investment reflecting current perceptions of value and use. These may be essential concepts for funding bodies and repositories who wish to ensure further appropriate and effective allocation of resources. In our curation models the most expensive part of archiving is the "ingest"[47] phase – the preparation and appraisal of the data for the repository. For continually curated, active datasets the most expensive part is the cost of subject specialists for annotation and linking to other research.

Recommendation A6:	Criteria need to be established to determine what data we should keep, why and what level of curation is appropriate, together with mechanisms to monitor, validate and to modify them with accumulating experience.

Given that we cannot anticipate future value, it should also be a matter of course that most non-reproducible data not accepted into a national or other archival repository is kept securely for a minimum period, to decrease the risk of loss of data only later recognized as of value. A period of minimum retention may also be a requirement for legal or academic reasons, e.g. validation of research.

The level of curation and access during this period could be to agreed minimum standards perhaps administered at local level. During this period information on the data should go into a clearing house pool, with minimum metadata enabling location and evaluation but without incurring the higher cost of proper "ingest", access and user support.

In due course we suggest a project to research tools for unearthing dormant pooled data which with time has become of interest and has potential for enhanced levels of curation and re-use. Repositories and funding bodies will need to develop evaluation criteria and procedures for any such retrospective enhancement of datasets.

No system is going to be a perfect filter; we need to develop guidelines on how much we are prepared to lose.

Recommendation B9:	Data not accepted into a national or other archival repository should be kept securely for a minimum period. Information on this data should go into a clearing-house pool, with metadata enabling discovery and evaluation pending final arbitration of its future value or role.

[47] OAIS terminology - see ibid..

Recommendation B10: Research should be undertaken to create tools to evaluate dormant data for re-use.

5.3 Compliance - cultural issues

To get the most from e-Science needs a culture of providing resource (including time), and of data sharing. Interviewees believed there is a need to change the culture of the research environment in order to engage and motivate data creators in the activities entailed for them in data preservation and curation. There are significant cultural, psychological and organisation hurdles.

As noted above in regard to metadata collection, curation and preservation as currently practised requires investment by the researcher (primarily in the form of time) to prepare data for future curation. The input of the researcher in documenting and providing context for their data is often vital to its potential value for future research. Often the researcher's investment is needed later as well, as the data undergoes curation and preservation processes. This needs to be taken into account in curation mechanisms and the structuring of communications channels.

Recommendation B11: Research is needed to improve automated support tools to prepare data and metadata for archiving.

Incentives

Professional advancement comes from publication of papers, be that publication in journals, venerable or young, or increasingly on the web. (The use of the web by researchers to publish material makes it widely available, but one of the main risks here is that the material simply disappears over time.)

A number of interviewees suggested mechanisms to support recognition of citation of or access to datasets, in place of and in addition to citation of papers. Greater efforts to link and add value to information generated in the research process, through citation of datasets in pre-prints or in journal articles, and the ability to cross-search and access them in a common research information environment could raise the profile of the datasets; good curation of the datasets would thus also be in the researcher's long-term professional interest. A journal of datasets has also been mooted.

Research Councils, universities and other funders should actively encourage this process, and journal publishers should also be recruited to this effort.

Some institutional actions which could be implemented include:

- Requiring institutions to report on their curation activities (actions and benefits). This also creates accountability and may encourage them to identify benefits.
- Including management of digital assets (including primary research data) as a heading in future institutional risk-management assessment exercises.

Recommendation B12: Consider requiring institutions to report on their curation activities, including management of digital assets (including primary research data), as a heading in future research assessment and institutional risk-management assessment exercises.

The benefits of trusted, curated, grid-supported scientific data repositories need to be articulated to the wider scientific community[48]. It is important that scientists and researchers actively want to be part of the digital curation process. Incentives to do so – such as recognition for citation of datasets, promotion of scientific advance from collection-based science – are important. Ambitions in curation-enabled science should be encouraged. Recognition should be given to citation of datasets.

There should be awareness seminars for university and institute staff, explaining the opportunities, the risk of waste, and encouraging the creation of interest groups and discussion forums. The issue should be raised at meetings held by university vice-chancellors and institute directors, at an early juncture.

Recommendation B13: Seminars and forums should be held for institute directors, vice-chancellors, faculty heads; the benefits of trusted data repositories need to be articulated to the wider scientific and research community and their support staff.

Future grants should include a requirement to submit primary research data to a specified repository or body as mandatory, which in turn means that repositories need to be designated. Grant-holders should be provided with support to meet this requirement, and should be offered training where necessary.

Would rules and penalties help? Some research councils' conditions of employment or grant require researchers to submit their data to their data centres, but this is not entirely successful. One suggestion would be to withhold the last 10% of funding conditional on preparation of data for long-term retention and actual submission of data; while this would probably be effective, it addresses the problem after the data creation stage, when input is particularly useful; secondly it may require significant administration.

Greater compliance is likely to come from a combined approach addressing incentives to participate alongside any measures for enforcement of conditions.

The short term

Particularly within university-based research, projects see a high turn-over of staff, including within projects. This is the nature of research – staff move on, post-graduates and post-doctorate staff in particular. For data preservation and curation, however, this is an additional

[48] This issue is discussed in Research Libraries Groups and OCLC. "Trusted digital repositories: attributes and responsibilities", 2002.

problem and risk, as these people take knowledge about data (content, context, technical) with them. Some areas are less vulnerable than others, where good practices are driven by habits or by the nature of the work they are doing (epidemiologists, astronomers for example). The loss of tacit knowledge was an area of particular concern for several interviewees, including the most senior. Nowhere did we meet any systems or procedures which addressed this problem.

As we set out in chapter 4, the short term affects provision for primary research data in other ways.

More profoundly, however, because approval for funding tends to come from bodies for which research projects are top priority and by which they are measured, funding for repositories is inevitably regarded as taking money away from research projects. These repositories provide a community service. The "-omics" databases provide striking examples of their benefit. Funding, however, comes from a research rather than a research infrastructure source.

These considerations reinforce recommendation A3 relating to funding.

5.4 Curation activities

Data management, categorisation, preservation

General good data management practices on the part of computer users, at whatever stage of the data's life, but in particular when first generated and used, will reduce the burden of work required for preservation and curation.

It should also be possible in the future to attach annotations to data and publications - rather like slipping a piece of paper with a scribbled comment into a book, or writing a review in a newspaper about a book. These annotations should maintain persistent links to the underlying data object(s). These annotations might be informal additions, or provided formally (as for a curated genome database, for example). Provenance information is important for either type of annotation, to enable others to gauge the value and reliability of the annotation. This provenance information needs to remain attached to the annotation. This has implications for data architectures and is an important area for investigation.

The ability to maintain links between data, e-prints or articles, materials and annotations, provenance information, and the existence of a semantic web or grid using ontologies, brokered by portals, will build a growing encyclopaedia of information and knowledge of extraordinary value.

Data categorisation: As noted, our survey confirmed considerable heterogeneity in file format, structures and file sizes. Managing the curation of a range of formats and structures represents a large challenge, both to meet the technical problems and to hold down cost. On the other hand, there are commonalities in data, some as yet unidentified, and which may be shared over quite different subject domains (static/dynamic datasets are one example). These could be used in curation and curation management:

- **Static data** – once collected it is closed and needs no further curation (but may need preservation)
- **Dynamic data** – which will be under continual change, either accruing in content or merely subject to revision.

The two types of data have different management needs. Static data is usually assumed in models for long-term retention, but other datasets do have a long active period during which they are subject to curation processes in the wider interpretation. One interviewee noted that in the dynamic case, **structures** within the data might change as well as content and annotation, introducing another layer of complexity. This needs to be researched.

Recommendation B14: Research is needed to investigate further the curation of data subject to structural as well as content change.

Longitudinal data sets: We were reminded by interviewees that in disciplines like high-energy physics, experiments themselves can be spread over many years, possibly decades, so that even during the active, data collection phase of the data life-cycle technology change may occur and some of the preservation work within curation my be needed. Longitudinal studies are other common examples of data sets which can span long periods. Such data sets will require different curation management approaches, in particular with regard to access management (particularly where data is confidential) and preservation actions..

Versions: There can be different versions of datasets. The different versions of primary research data have differing relationships to secondary and tertiary data, which evolve differently over time.

Confidentiality and access management: A significant proportion of the data involves confidentiality issues. Confidentiality may involve anonymisation of data; it imposes access conditions. Access management entails authorization - we noted that this is a major area of work involving CCLRC, JISC and several e-Science projects; this is a particularly difficult area where clinical data is involved, extremely so over long periods of time and when the customer base may extend beyond national boundaries.

Some data is produced with industry support, which affects the timing of open release of the data and may also involve longer-term conditions. All these aspects must be respected and therefore managed as part of the data curation.

Heterogeneity poses major inter-operability problems. Mastering the inter-operability widens opportunities, as some of the e-Science projects are demonstrating. We note the significant amount of work being conducted in e-Science projects on inter-operability issues, to support collaboration in Grid environments. We would also comment, however, that inter-operability does not equate to preservation and should not be allowed to encourage inactivity with regard to the separate data preservation needs.

Primary research data **volumes** can be massive, and the range of record sizes within these volumes will range from one extreme to another. This poses handling and management problems[49] (which the Grid and Grid software tools will help to meet).

The curator's role

There has been concern with relation to existing archives that the level of re-use of data has been disappointingly low for significant parts of the collection. Without exception, those managing these archives referred to tight resources, preventing them from encouraging better and wider use of their holdings. Curation (as used here) is a relatively new discipline and profession. In addition to under-investment and lack of promotion, other contributing factors may be that past practices have led to inappropriate selection or curation. We hope that findings and recommendations elsewhere within this report will underline the importance of the modern curator's role in addressing these areas, and for appropriate future investment and practices.

Visibility and usability

The more visible and the more useful the data, the easier the curator's task will be. Portals, ontologies, mark-up languages, metadata, links all contribute to data discovery and visibility; data usefulness is a function of the ability to the use, manage, analyze the data in the first place (hence the role of tools) and of the quality of the data.

Though storage and curation might be local and distributed within an organization, there is a need for agreed metadata standards across and beyond the whole organisation. The same applies to structures in repositories (wider use of the OAIS reference model), and with mark-up languages (data creators should be aware of these, and agreed standards used).

5.5 Trust

This issue – can we trust the data we are collecting, accessing, caring for and preserving – emerges in a number of guises. At one level it is a systems issue – is the data kept on systems which are well managed and available; is availability continuing into the future? Do the systems have sufficient security, is access properly controlled?

Can we have confidence in calculations made in distributed environments? Different arithmetical architectures in different machines may lead to inconsistencies, and if the Grid middleware is not aware of these subtleties, or the user is not expert in numerical computation, results may be suspect, or there will be validation issues. Will a calculation made now give the same result as a calculation made in the future on a different architecture?

A point made particularly strongly by some interviewees was, can we trust the data itself? Data nowadays is rarely verified when it is entered. Data validation is an acute problem in clinical medicine, and a general problem throughout science except in certain areas (like high energy physics). Data we should now doubt may in the future be assumed to be correct. A

[49] See for example remarks relating to oceanographic data in Appendix 3.

"kitemark for data" was suggested, extending Professor Lievesley's suggestion in her 1998 report of a "kitemark" system for storage and preservation centres, a recommendation we ourselves make. These kitemarks might relate to levels of data cleaning, validation, and levels of curation undertaken by researchers and/or curators, or to the certified operational standards and practices of the repository.

Recommendation B15:	Consideration should be given to the adoption of quality markings ("kitemarks") for both data and for data repositories. This could be linked to development of codes of practice for curation centres.

Trust in data in the future may depend on knowing the business or research context in which the data was created; More than one person commented on the need for systems to capture workflows, even where these were of the non-prescriptive kind employed in research. This concern has echoes in the records management regulations set out in the FDA's 21 CFR Part 11 ruling[50]; this is a subject that current electronic laboratory notebook research may address.

With regard to the use of electronic laboratory notebooks, this should apply good record-keeping standards. There has been a considerable amount of work relating to electronic records and electronic laboratory notebooks in the commercial sector which is publicly available[51] and which should be consulted by public-sector research.

Another aspect of trust is what one interviewee referred to as the "social-infrastructure" tier in curation organisations, for example those who manage data such as SWISS-Prot[52], where trust is enhanced by the use of qualified biologists for data curation. The quality of a curated database depends on the quality and reliability of the people doing the work. It was stressed that the usefulness of the database is a function of its quality. The more the data is used, the better the return on investment. The better quality the data, the lower the risk.

Data with uncertainty is used less, not at all or, if used cavalierly (whether unconsciously or not), is a source of risk and contamination of knowledge and endeavour. Provenance information is essential; knowledge about or confidence in the level of provenance information is as essential as the provenance information itself. The provenance information needs to be long-term, because "data floating in cyberspace can have many lifetimes far from the control of the originator"[53].

The utility of data also depends on the ability of users to manage and analyze it. This means algorithms for mining data, visualization tools, user interfaces and portals to make data accessible. These will play a crucial role in accelerating research. E-Science projects are providing opportunities for case-based development of many such tools; several such projects

---

[50] ibid

[51] Such as that by CENSA.

[52] See: http://www.ebi.ac.uk/swissprot/

[53] John Rumble, Jr., National Institute of Standards and Technology, USA. Unesco abstract.

include archive elements, but these tend to be planned for later stages of projects, which have not yet been reached. We believe it would be useful to bring together project members and experts in the archiving/curation field to share findings and concerns, and by doing so draw out and draw attention to the relevance of these components of e-Science projects.

We recommend that codes of practice and minimum standards for curation-data centres be developed. This may be a role for the DCC.

Authenticity

This is closely related to provenance – how do we know in a digital environment that a record is what it purports to be and that its integrity has been preserved and it is linked to its creator? Again this was flagged as an issue by a number of people. The FDA's 21 CFR Part 11 ruling[54] addresses this for highly regulated environments; can the same objectives of assured authenticity be achieved in more free environments of academia? The pharmaceuticals industry is still struggling with the problem.

5.6 Areas for research

We conclude this section by gathering into one place (Table 12) the numerous suggestions made by interviewees for areas where further research is needed.

Table 12: Research topics suggested

Item (sorted alphabetically)	Notes
A top-level index – so we know where everything is	
Access to data	
An agreed vocabulary for curation	
Assumption of a fixed record	
Authenticity	
Blurring of metadata and data - implications	
Data volumes – how to handle	
Funding stability and levels – how to achieve	Mentioned twice
Intellectual property and copyright issues	
Liability issues	
Maintenance of confidence in data	
Metadata and the harvesting of it	
Metadata models for file formats etc.	

[54] ibid

Non-static datasets	
Ontologies – various aspects	
Ownership	
Preservation – particularly of multi-media formats	Mentioned 3 times
Provenance	Mentioned 3 times

Recommendation A10: Investment should be strengthened in those areas of curation research which will enhance data re-use; in particular we recommend focusing on those areas of research needed to establish trust in curated information.

5.7 Organisational components of curation

If it is possible to see any pattern at all in the views we gathered on this topic there is a consensus that curation activities need to be undertaken with (scientific) domain expertise. For those involved with genetic database curation, proximity to "the scientists" (preferably at the same location) for those carrying out database annotation was important. The location of storage functions could be decoupled from curation centres given adequate communications technologies, a facility enhanced by Grid technologies (provided the right service-level agreements are in place). This was referred to as a "levelled model" by two interviewees.

Though not in the science field, the experience of the AHDS was interesting: it started as a research project, and the organisation evolved with a decentralised model, serving several disciplines with a degree of autonomy and with a central administrative hub. This has proved less cost-effective as it has moved from project to service, as different solutions to storage and data organisation arose, whereas efficiencies could be gained by centralising what is common and can be shared between disciplines (such as software techniques, storage provider management) and leaving discipline-specific work with the domain experts. Re-organisation on these lines is now under way - "doing what is appropriate in the appropriate places".

Rolf Apweiler at Swiss-Prot reported that large numbers of people are working on curation in EMBL and Swiss-Prot/TrEMBL (50-100 curators). In his view this investment is extremely cost-effective for the discipline, 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, much is done centrally, allowing the organization to spend fewer resources and at the same time achieve higher quality. The same is essentially true in GlaxoSmithKline.

Following from these observations, in this section we examine some of the general principles for structuring curation. Recommendations for specific organisations or specialities were beyond the scope of this study, and clearly no single model will fit all circumstances. We start by listing some considerations which need to be borne in mind:

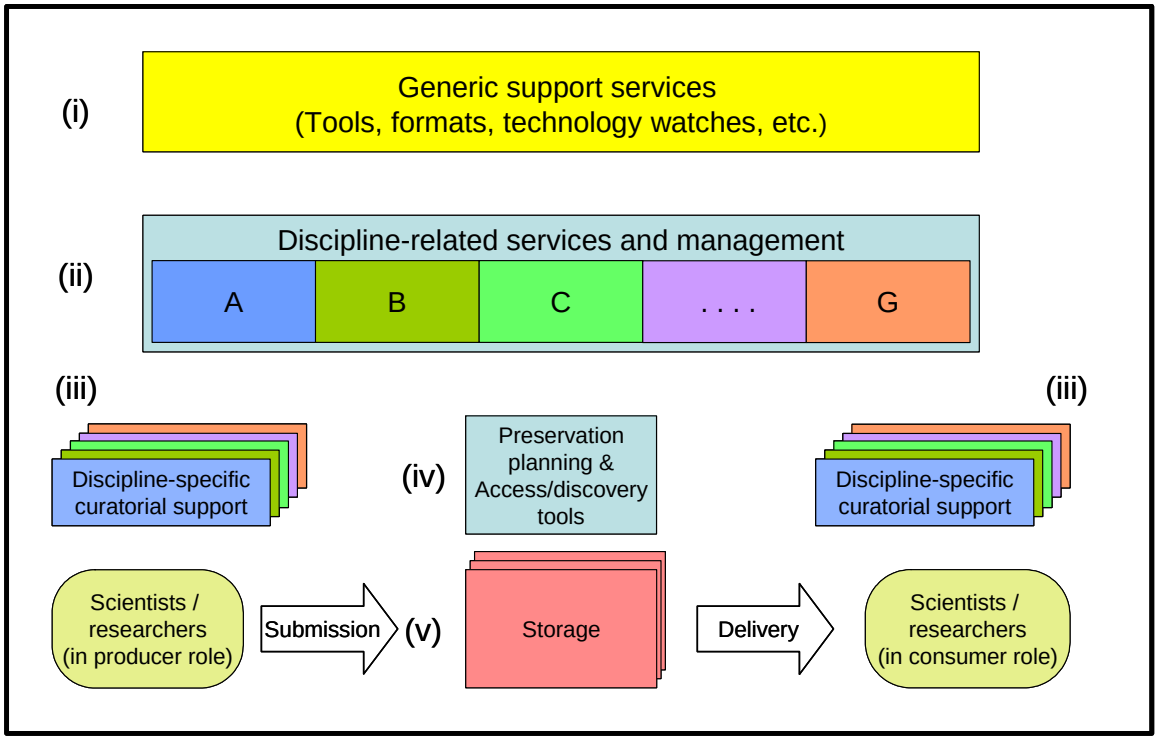
- Almost all institutions have their own computing centres already, though a computing centre is rarely a digital archive, still less a curated repository.
- Many institutions in the UK are already centres of excellence in relevant areas, usually centred around people, who in turn act as a magnet for others. There should be co-operation between centres (but no loss of autonomy). Aspects of curation activities vary widely according to data type, technologies underlying data, and (above all) discipline.
- Primary research data curation is likely to extend into or share similar concerns with other areas (such as e-learning/course-ware), and this should be taken into account in strategic and resource planning (flexibility should be maintained where possible). JISC and the proposed Digital Curation Centre are particularly well-placed in this regard to ensure appropriate linkages and generic services.
- On the whole (but not entirely), management of storage involves common skills and resources, whatever the discipline; (the "not entirely" relates in the main to extremes of data volumes, access frequency and bandwidth).
- Until there are reliable automated tools (possibly extending to automated scientific metadata harvesting), as noted already, data generators (submitters) will need good support; the same applies to consumers (who will need file management guidance and, for primary data in particular, assistance with technical issues relating to the re-use of information, visualization tools, etc); this support must have both discipline-specific and technical skills.
- On the other hand, there are many areas common to every discipline where generic tools and services can be provided, such as access to common file format specifications. Such is the role of the DCC discussed earlier.
- Some curation activities (genetic, protein database annotation, for example) require full-time staff physically located close to data generators. On the other hand, decentralized facilities provide economies of scale for machinery. Grids surely offer the ability to get the best of both worlds.
- We should build on the momentum and cross-disciplinary, cross-organizational co-operation engendered by the e-Science programme, though we note that the first and second sets of e-Science funding will be coming towards an end when any digital curation initiatives are in their early days.
- The OAIS[55] reference model provides a framework within which to build a suitable structure for repositories; we believe, however, that the model may need to be enhanced for curation activities. In particular the model has little focus on "pre-ingest" activities surrounding data creation and early stages of curation, or on dynamic datasets.
- Data is valuable; storage should be safe, secure, and managed to the highest standards to ensure retention of value and to promote trust in its provenance.

[55] ibid OAIS: Reference Model for an Open Archival Information System, now ISO 14721

Recommendation B16: Institutions, Research Councils and universities acting alone or in consortia should be encouraged to adopt OAIS-compliant archives and to institute methods to assess curation of academic digital assets.

Figure 11 shows the **organisational** components we believe are needed for a curation service: The individual components are described below; they may be geographically separated or together; they can be single or multiple.

Figure 11: Organisational components for curation



The following notes describe the main characteristics of the model, the Roman numerals indicating the various elements. The model makes no assumptions about the whether in any one instance (sub)components are shared or not, nor whether they are internal or outsourced (with the exceptions noted below). For clarity, links between components are not shown on the diagram.

i. A unit which provides generic support services serving all (sub)disciplines - such as providing tools which can be deployed across domains, basic training, technology watches, information on generically used file formats and software. This could be identified with the Digital Curation Centre.

ii. Discipline-related centres which provide overall management of curation for nationally/internationally selected datasets. Within this there may be units which specialise in the particular needs of sub-disciplines. Thus one could envisage a medical data curation centre, within which there may be specific speciality groups (such as for population data, genetics, clinical chemistry, etc.). These would have

contact with those working locally with scientist described in (iii). Staff here will possess both discipline-specific skills and curation skills.

iii. To satisfy the needs of users who need to interact with the system, these are (sub)discipline-specific support group(s) which work closely with the data generators and consumers. Again, staff here will possess both discipline specific-skills and curation skills.

iv. Provision of tools needed for rendering/access and ,discovery and evaluation of retained datasets, and also to preserve data and to plan preservation activities. We believe they will usually be provided by the organisations described in (i) and (ii).

v. Storage facilities, which may be local, central or outsourced according to circumstance, and which provide for the physical management of digital information. In general they can usually be independent of the disciplines whose data they hold, not needing knowledge of the intellectual content of the data. Such facilities provide storage media and its management, back-up, physical and logical security, network access, etc.

It is possible to map the OAIS functional components onto this organisation in specific instances.

We believe that it is not possible to build a monolithic structure for the whole of the UK primary research community; neither would it be desirable – though this may change in years ahead. While a monolithic option might appear to offer economies of scale, delivering curation to a highly heterogeneous client base would involve complex management which would probably be over-cemented to be flexible for the medium-term evolution in the activity. The components listed in Figure 11 need to be provided by the funding bodies and the institutions acting individually or in consortia, bearing in mind needs to:

- Enable decisions (which may cross domains) to be taken efficiently at the given levels;
- Facilitate clear policy-making across organisational boundaries (again, possibly across disciplines);
- Achieve economies of scale where possible;
- Fit into current organizational structures;
- Make use of existing facilities and resources where possible (we believe there are more of these than first meet the eye);
- Foster data sharing; and
- Cater for load-balancing of resources.

This is an area of complexity, involving a number of different domains and institutions, and there is a clear need for one or two small bodies or units to provide co-ordination guidance and expert input. This reinforces our recommendation A2 for the assembly of a curation task force.

6 Funding

Note: The following describes the situation at the time of writing – research funding structures are undergoing review.

Dual support system

The dual support system evolved as the means of managing public support for research in universities and higher education institutes. The public-sector funds come from two sources: (1) the education departments, administered by the Higher Education Funding Councils (HEFCs), and (2) the DTI's Office of Science and Technology (OST), administered via the Research Councils. HEFC money supports a basic level of research activity for university academic staff, library facilities, and the "well-found laboratory" in which work supported by Research Councils and other agencies is undertaken. Research Council and other agency support enables the selective support of lines of research, provides central facilities, provides access to international facilities, and can encourage particular fields believed to be of national importance.

Because approval for funding tends to come from bodies for which research projects are top priority, and by which they are measured, funding for repositories is inevitably regarded as taking money away from research projects. Under the dual support system, a substantial portion of higher education institute (HEI) funds depend on their research output. The result is that the HEIs focus on supporting as many researchers and generating as many research publications as possible, in order to sustain their future income.

There are interesting parallels here with the "research funding gap" identified in the 1997 Dearing report. This was the gap between what research carried out in higher education institutes is costing HEIs and what they actually receive in grants or contract payments to carry out the research. The implication articulated in the Parliamentary Office for Science and Technology's summary report was that "..[we are] losing the seed corn for future research and researchers". That research gap has been addressed in various infrastructure funds (see Figure 12 below).

The bottom line was that much of the research was not bearing the full economic costs of the work. The government's Science Budget 2003-04 to 2005-06 includes £120 million per annum from 2005-06 to enable the Research Councils "to pay a larger contribution to the full economic costs of research in universities". The "Investing in Innovation" report of 2002[56] notes that an important factor for the sustainability of the university research base will be for the full costs of research to be recovered by universities. Data preservation and curation activity involves resources (human) at project stage; if post-project data is held by universities (or other institute which was home to the original research), this may give rise to another funding gap.

[56] Investing in Innovation: A strategy for science, engineering and technology, July 2002.

One complicating factor with regard to the "research funding gap" was that universities in particular did not have accounting systems which were able to identify and cost the components of research.

Figure 12: University science research infrastructure funding (SRIF) 1999-2000 to 2005-06[57]

£ million	1999–00	2000–01	2001–02	2002–03	2003–04	2004–05	2005–06
	Joint Infrastructure Fund			SRIF(1)		SRIF2	
OST	50	125	125	125	250	300	300
DfES	25	50	75	150	150	200	200
Wellcome Trust	50	125	125	75	150	–	–
Total	**125**	**300**	**325**	**350**	**550**	**500**	**500**

(1) Does not include the £50 million per year that was dedicated to Research Councils' capital requirements.

The parallel (with the research funding gap identified in the Dearing report) only goes so far, however, in that we do not yet know what the costs of data curation will be (with the possible exception of a few indications, such as those provided by the curated databases maintained by the European Bioinformatics Institute), how they will be distributed over time, nor how they might be optimally configured. Curation for primary research data, however, spans a wide range of evolving and different volumes, growth patterns, relationships.

Data curation straddles research and infrastructure, both in terms of component activities and output.

Charities

The research funding gap has been and continues to be addressed through investment in the Joint Infrastructure Fund and the Science Research Investment Fund, to which the Wellcome Trust is a major contributor. Charities are increasingly important funders of research in universities and colleges (17% of total funding in 2000-2001; in biosciences and medicine, charities, led by the Wellcome Trust, provided more funding for research than the Research Councils). Their funding is outside the dual support system, though following the 2002 report for HEFCE into charity-funded research, HEFCE, institutions and charities are reviewing the possible contribution by charities to indirect costs of projects. We note that two major charities, the Wellcome Institute and the new cancer research institute, are actively involved in data curation projects.

Charging for access?

As we explained earlier we believe that the use of the word 'curation' is apt, as it looks outward and carries the sense of provision of a service for the good of the community. Of

[57] Source: DTI, Science Budget 2003-04 to 2005-06

course, this extends the stakeholder framework, with implications for any charging that might be applied (at the expense of a possible increase in complexity). At the moment, access to data held in existing data repositories or curated databases is generally free at the point of *delivery* in the UK for the academic sector, sometimes also for the commercial sector. There are some exceptions; a particular case is NERC, custodian of nationally important environmental data, which generates some £1 million to £2 million in revenues annually from commercial bodies accessing data. In such a broad report, this study has not sought to identify potential sources of or models for commercial income from curation of publicly-funded research data, given the breadth of domains covered, nor does this report assess reliability and eligibility.

As we also suggested, the returns generated by or on data curation are likely to be indirect, and they are also likely to redound to a party other than the data creator or direct funder, with indirect fiscal benefits for the government in question, which are difficult to track and quantify. However, as Figure 12 suggests, we believe the benefits will feed back into the system, as well as preventing waste.

Free-of-charge, unlimited access to data may be presented as an ideal, but is it feasible over the medium or long term? The funder(s) of curation/preservation may wish or need to make some charge to some or all users for the data or service provided, particularly if it is perceived that charges might help keep data updated and develop functionality, if appropriate. Many believe, and make cogent case, that more open pricing policy in the public sector generates greater economic benefits[58], in the form of increased economic activity and higher tax revenues. (Charging, of course, implies an additional activity for which resources have to be provided.)

The funder of data curation pays for the preservation and curation of the data. Depending on charging policy, he is also funding the user, through the provision of the data to the user. Actually providing the data may need to include a significant amount of curatorial support, in other words, it may entail quite a high cost of provision. If charging is applied, it may need to include the cost of, or contribution to the cost of, curatorial support when the data is provided to the user.

Multiple funding sources

Several biotechnology databases are supported by several sources which fund for limited periods. We came across many cases of multiple funding sources when researching this report. It is the case, for instance, for the UK Data Archive and EDINA. Multiple finance sources for data centres, datasets or databases may become increasingly common, particularly if the task becomes increasingly complex, and with an increase in inter-institutional or international collaboration which will be fostered by the Grid. This, however, may make sustained funding more difficult, particularly if there is a perception by one body or nation

[58] European Commission, 2002: memorandum on the proposed directive on the re-use and commercial exploitation of public-sector documents: This refers to several studies, European and American, which model the impact of different charging models, concluding that low-pricing models give the highest benefits for society as a whole".

that it is shouldering a disproportionate amount of the funding, or that it is subsidizing others and does not wish to do so.

Measuring benefit

Interviewees at policy-maker level indicated that funding is difficult to achieve where funders are unclear as to the benefits of what is funded. We heard of cases where archive funding had been questioned because of perceived low use of the archive, while on the other hand, we heard the frustration of many (including data centre managers) at having insufficient resource to generate more use of the holdings in question.

Time constraints prevented us from making a detailed analysis of the cost and benefit of example data centres within this study. When relevant e-Science projects are at a later stage it would be useful to carry out an analysis of cost components and outputs over time. Such analyses we believe would provide useful planning aids, helping to identify and endorse areas of focus for research and development of tools, and infrastructure planning, as well as contributing to decision-making with regard to allocation of funds.

In a cost-benefit analysis, various factors will have a direct and significant impact on the ratio of cost to benefit - tools facilitating use, provision of service, visibility of data, curation activities extending the reach of data, etc. It is likely that the tools under development (within e-Science projects, for example) and to be developed will substantially reduce the cost of curation.

In some cases it is still possible that the returns over time may prove to be less than the cost of creation of the data and curation combined - though a fairer measure would be the cost of curation of the data **net** of the cost of its creation.

6.1 Cost components

Within our study we looked more closely at the cost components of a few data centres or repository services. These work to varying business models (centralized national archive service, distributed national archive service, specialist, part of IT services provision) and fulfil different roles. Nevertheless, some interesting points emerged.

The accounts data available showed cross-subsidies - in other words, not all costs were included in their accounts for the unit's operation. Gaps related, for instance, to premises costs (including related overhead) or staff salary contributions. It was also frequently unclear whether certain operating cost overheads such as electricity were included in some accounts. In part this is because such costs are often provided by an institution as its contribution to a service and not included in accounts for other funders. However greater transparency may be necessary.

Cross-subsidies

More than once in our report we remark on cross-subsidies which are not shown in accounts we have seen for repositories. This is a source of vulnerability. When money is tight, those bodies providing the cross-subsidies, whether premises, headcount, administration, or

telecommunications infrastructure, may cut or restrict the effective subsidy, leaving the repository needing to fill the gap, which it will either do by acquiring corresponding funds for provision, or cutting its own provision. The removal of premises entails major risk, because of the likely need to relocate and staffing implications. Yet for these data centres, their raison d'être is to provide for the long term.

Recommendations B17: To address the risk represented by cross subsidies, activity-based costing analysis might be considered to identify particular areas of vulnerability. Vulnerability arising from cross-subsidies is a feature of the current short-term funding system (see A3).

The percentage figures quoted in the next paragraph relate to the figures provided in the accounts, not the actual full cost.

Components

Staff costs consistently represent by far the largest item, whatever the model (between 69% and 85%). Excluding staff costs, the largest items were maintenance (IT maintenance costs), travel, network costs. Publicity materials typically represented between 10% and 20% of non-staff costs.

Staff costs typically consist of five areas: management, operations and systems support, infrastructure support, domain-specific support, and research and development. The balance varies. Research funding for specific projects makes substantial if not exclusive contributions to research and development funding, usually **not** included in the accounts data we saw.

Without exception, data centre managers stated that while they meet demand requirements, their headcount is insufficient for the sort of contact with their scientific community commensurate with the level of service they would like to provide, and to promote wider and better use of their holdings.

The tighter the funding, the lower the proportion of domain-specialist staff.

Curation cost and curation benefit

The relationship of cost to benefit of digital curation is likely to prove not to be a straight line relationship. Over the next few years a more appropriate word than "cost" here would be investment – for example, in the form of increased expenditure in curation which goes into work to provide tools to improve the quality of the data, or support data appraisal by archivist or curator, or provide tools to improve data discovery capabilities. The investment in tools such as these will result in cheaper curation provision and preservation work in the future. This in turn will provide UK research with what the commercial world might call "competitive edge" – faster, more powerful, more extensive research resources.

The more data is used, the greater the return on the cost of the data. What one may see in the future, once tools and mechanisms are established, is that a marginal increase in investment in curation may produce a multiple of the benefits. Once a certain threshold of additional use is

reached, additional staff count may be required to support demand, and this would affect the relationship curve. Increased use may entail an increase in other costs as well, such as higher on-line storage costs. However, on the basis of current cost percentages and unless hardware and media costs multiply massively, increased usage would still improve the cost-benefit relationship.

Levels of investment: Analysis of cost in relation to benefit could be conducted at different levels of curation, ranging from the basic one of whether to invest in digital preservation/ archiving/ curation at all, to analysis at domain level, individual collection level, down to whether to keep a specific record or data set. Any such analysis would need first to address several questions, such as the point at which to start tracking cost, the time span covered, data quality and enhancement, and the measurement of benefit. In the context of this study, in many cases the overall purse paying for the data is the same (the taxpayer through the state).

Fragmented funding path: The existing cost "homes" for the various activities and thus cost components over the various stages (from data generation through to archiving or curation) are several and varied, which means that the funding for each section has to be fought for against different contenders for the same money, whose role and therefore funding criteria are different, and by stakeholders with different aims.

Time of analysis: The analysis may also look very different in the future: the activity of curation/preservation is made up of many sub-activities, many of which are cumbersome now but may become automated or routine in a few years. So the cost profile of data curation would be quite different to its profile today. Indeed, it is possible that data curation as a distinct activity will require significantly less in relative terms of resource needs despite increasing data volumes, if tools automating much of the process and provision can be developed.

7 Related issues: intellectual property; data sharing

7.1 Intellectual property rights (IPR) and copyright legislation

The umbrella term "intellectual property" became common in the 1960s, becoming more important in the 1980s to developed nations as their manufacturing power eroded. Use of the term exploded in the 1990s, boosted by increased numbers of company "spin-outs" from research (and also influencing the drafting of the European Database Directive discussed below). This is also known as technology transfer and this figures high on the UK government's lists of objectives for science and technology.

The European Research Area calls "the protection, management and transfer of intellectual property (IPR) [..] an increasingly important and strategic issue for all those investing in research and innovation. .... IPR is more than a legal matter, it is a key element in the transformation of knowledge into economic value. Improving IPR systems and their use is therefore an important dimension of European research policy and the creation of the European Research Area." The European Commission Research Directorate is therefore engaged in activities to "address the IPR-related needs of the European research community". These include:

- "Identifying, promoting and disseminating best practices for the use of IPR in the research and innovation process [...], and
- Promoting legislative adaptations of the IPR systems which may appear necessary in light of the changing research environment."

The European Research Area also states as a priority the stimulation of research using data and facilitation of access to data by academic and industrial scientists, stimulating further research and the commercial development of new products, processes and services. There is the possibility of conflicting aims, between open access and intellectual property rights. On the other hand, it is also possible that intellectual property rights can actually maintain accessibility of data for the research community.

A problem is that wording of intellectual property legislation - and also wording of individual patents - can restrict access to resources which are needed for other research. One famous example is that of the patent which Human Genome Sciences filed for and won for the CCR5 receptor, guessing the role of the gene, having compared its sequence with that of other genes whose functions were known. The following year, Parmentier and others, without knowing of HGS' patent filing, discovered that CCR5 is the portal entry into a human cell for the AIDS virus. In the words of the director of the Nuffield Council of Bioethics, intellectual property has a social cost.

Professor Paul David speaks of the risk of "over-fencing of the public knowledge commons" in science and engineering arising from "ill-considered government support for expanding legal means of controlling access to information for the purpose of extracting private

economic rents". He continues, "This would bring adverse long-run consequences for future welfare gains through technological progress, and re-distributional effects further disadvantaging the present economically less advanced countries of the world". One subject of concern relevant to primary research data is the European Union's 1996 Database Directive[59].

Under this directive, European Union member states were required to protect databases by copyright as intellectual creations and to create a so-called "database right", which prevents unauthorized extraction or re-use of (the) contents of the database. Article 1.2 of the directive defines the object of protection (a database, electronic or paper) as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means". Thus a collection of works (such as an anthology) can qualify as a database. A database can consist of subject matter that is neither work nor data, such as sound recordings. The elements comprising the database must be "independent" and "individually accessible by electronic or other means" (so a string of images such as a film does not qualify). The protection granted also covers "the materials necessary for the operation or consultation of certain databases such as thesaurus and indexation systems", although it does not cover the software which drives the database.

Recommendation B18: Senior policy makers should be aware of possible negative repercussions of legislation aimed at supporting entrepreneurial activity.

The database right protects the "sweat of the brow" of the producer of the database - the skill, effort, money invested. The investment must be "substantial", and can be qualitative or quantitative. The protection lasts for 15 years from date of completion of making the database, or of first making available to the public. In addition, any "substantial change" extends that protection, so any regularly updated database effectively enjoys semi-permanent protection.

Only limited exemptions are allowed, however. These did not include traditional limitations to rights such as the right of quotation or library privileges, or re-use of government information.

Sir John Enderby, as Vice-President of the Royal Society, pointed out that in many cases in scientific databases the initial facts were probably obtained at a cost "many orders of magnitude greater than the investment in compiling the database" - an investment which may well have been funded by the taxpayer, who should not be asked to pay again, whether directly or indirectly[60].

Case law involving national legislation implementing the directive gives a picture of uneven interpretation and stringency in interpretation, and does not help clarify the meaning of some

[59] European Union, ibid.

[60] Open letter to Mr. Stuart Booth of the Patent Office, 31st May 2002, about the European Commission's review of the implementation and effects of the European Database Directive.

of the terms of the directive - there is no indication as to what "substantial" investment means, for example.

A review of the directive is being conducted by a Belgian law firm, which has been asked to look at implementation of the directive in national legislations and at the impact of the directive. At time of writing we have not been able to access any information as to the current position of the review within the Commission.

7.2 Ownership and rights management

Ownership of primary research data generated in publicly funded science and retained is likely to change at least once in its lifetime. It is important that rights are clarified and managed over the data's lifetime.

Usually ownership of primary research data after project end either stays with the original funder(s) (sometimes a commercial entity) or passes to the researcher. The curating body and/or repository for the data is likely to be neither original funder nor researcher. In that case, the owner of the data should grant the curating body the right to carry out actions pertaining to its function. A further issue, however, is (i) whether the owner is entitled to claim some enjoyment of financial benefit arising from re-use of the curated data, and if so, to what degree; and (ii) should any such benefit be net of the cost of curation over the interim? There are traditional legal conventions which apply here (definition of use, for example), but it will be important to address these issues in advance to avoid problems later.

A further point is that curation metadata added by the curating body will probably be owned by that body or its funder, unless otherwise agreed.

7.3 Data sharing

CODATA (the Committee on Data for Science and Technology of the International Council for Science) in November 2000 formulated six "principles for science in the internet era" to support "full and open access to data needed for research and education". The same themes permeate the work conducted within the OECD's various public domain data projects[61].

A recent OECD report[62] described the benefits of data sharing as follows:

> "Open access to, and sharing of, data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies

[61] Such as the report prepared for the OECD which provides a discussion of the issues, "Data Sharing Policies", Wouters, Paul, 2002.

[62] Promoting Access to Public Research Data for Scientific, Economic, and Social Development OECD 2003 see http://dataaccess.ucsd.edu/Final_Report_2003.pdf

> on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new data sets when data from multiple sources are combined.
>
> Sharing and open access to publicly funded research data not only helps to maximize the research potential of new digital technologies and networks, but provides greater returns from the public investment in research.
>
> Improving and expanding the open availability of public research data will help generate wealth through the downstream commercialisation of outputs, provide decision-makers with the necessary facts to address complex, often trans-national problems, and offer individuals the opportunity to better understand the social and physical world in which we all live."

As we also suggested, the returns generated by or on data curation are likely to be indirect, and they are also likely to redound to a party other than the direct funder, with indirect fiscal benefits for the government in question but which are difficult to track and quantify. Information and communications technology has dismantled geographic barriers to digital information. It makes multi-disciplinary and international collaboration possible, and enables access to data anywhere (subject to authorization and technical issues). However, if the beneficiaries of services funded by a body in one country are global, will this affect that body's charging policy? At the moment, the general line taken in the UK is that research data is generally free at the point of access to researchers and students.

Efficient data sharing, national or international, implies compatibility of structures and tools, which implies a need for technical co-ordination. This represents a particular challenge with regard to primary research data, which is characterized by its heterogeneity, and the need to maintain links between derived and primary data.

Recommendation B19:	The international links established within the e-Science programme, e-Science projects, JISC's Digital Preservation Focus and other JISC programmes should be continued, and opportunities for further collaboration should be encouraged with relevant fora and bodies in the USA (for example, the NSF) and Europe.

Paul Wouters' 2002 report on data sharing policy for the OECD lists limitations on data sharing found in his survey of Web documents:

- Need to safeguard the rights of individuals and subjects
- The rights of individuals to determine what information about them is maintained
- Legitimate interest of investigators, for example, materials deemed to be confidential by a researcher until publication in a peer-reviewed journal
- The time needed to check the validity of results
- The integrity of collections

- Data released to the public that could lead to the identification of historically and scientifically valuable archaeological sites [which] could invite looting and destruction
- Data enabling the identification of the location of rare botanical species outside the United States could lead to unwanted bioprospecting and could damage the relationships between researchers and the host community
- Information related to law enforcement investigations
- National security information.

To this list should be added the privileged period of fair use granted to commercial bodies which funded the research to use data following the research project, for a limited time.

8. Summary, timelines and recommendations

8.1 Summary of main findings

The **recommendations** we make are listed in section 8.3 below. The following summarises the strategic level **findings** described in the previous chapters:

1. Confirmation that the data revolution presents significant challenges and opportunities. However, our surveys show that the UK is not fully prepared to capitalize on the opportunities and urgently needs to address this.

2. There is a lack of a government-level, overall strategy for data stewardship and data infrastructure to which science administrators can refer, still less to support the researcher in their evolving roles and duties with regard to data curation.

3. Existing data centres are usually supported by sponsors whose primary funding focus is research projects.

4. The current short-term funding models for the provision of curation are antithetical to its long-term nature and needs.

5. There will be an exponential increase in data volumes from e-Science over the next decade as planned new scientific instruments and experiments come on stream. However, to benefit fully from this major investment, further action is needed to support the curation of the data that will be generated.

6. Not all primary research data needs to be retained or has long-term value. Its potential value for generating new research will vary, and the level of investment required in the curation of datasets therefore needs to be identified and graduated accordingly.

At a policy level we found:

7. Provision of curation is patchy, and more advanced in some disciplines than others.

8. Where retention and curation of primary research data is a requirement set by funding bodies, the majority of researchers surveyed stated this requirement was not funded. Where guidance is provided, researchers frequently felt that it was out of date or inadequate.

9. Awareness of the issues – particularly data longevity difficulties – is generally low among researchers. Consequently the good practice needed to assure data longevity is rare, putting valuable resources at risk.

10. For curation to be effective the researcher needs to be engaged in the curation of his or her own data and working in partnership with curators. But few incentives or procedures are in place to ensure that this engagement is achieved.

11. Whilst practice and experience in curation is increasing rapidly, areas of curation are still in a research and proof-of-concept phase. Much research and practical, exploratory activity is being undertaken in the UK, and its quality is world-class.

12. The data revolution raises many issues of trust which must be addressed before data-based research can flourish – issues of security, confidentiality, ownership, assured provenance, authenticity, data and metadata quality.

13. There is little interaction and sharing in curation experiences between science-based industry and the academic sector. Within the next decade the curation of digital content and data is likely to be critical to science- and engineering-based industries and to knowledge-based economic activity.

8.2 List of recommendations

This list is divided into two sections, major, strategic recommendations, numbered with an initial "A" and those which are of a tactical nature and are numbered with an initial "B". Strategic references are also listed in the Executive Summary. The references column provides section references to where the findings are discussion support the recommendation.

Figure 13: Full list of recommendations

No.	Strategic Recommendations	References
A1	Strategic-level advocacy for data curation is needed and should be assigned to a respected and influential champion so that strategic objectives are clearly articulated, to set the UK's curation agenda over the medium term, and to enhance the UK's standing, contribution and opportunities in this area.	**4**
A2	A curation task force made up of curation experts, practising researchers and research administrators should be established to inform and guide this agenda. This task force should work closely with and inform the work of the new UK Digital Curation Centre.	**4**
A3	The mismatch of short-term funding against the long-term needs for data retention needs to be addressed by providing new specific, long-term funding stream(s) for data centres and curation, thus ensuring that there is a strategic approach to data stewardship which addresses holding information indefinitely, makes it widely available and encourages cross-disciplinary usage, including linking to other digital information.	**4.2**

No.	Strategic Recommendations	References
A4	Funding bodies should consider supporting research-led exemplars of curation to demonstrate and promote the benefits of curation for new research.	**3.1**
A5	Our findings endorse the need for the Digital Curation Centre and we recommend its establishment as part of a national provision for data curation in the UK.	**4.4**
A6	Criteria need to be established to determine what data we should keep, why and what level of curation is appropriate, together with mechanisms to monitor, validate and to modify them with accumulating experience.	**5.2**
A7	A programme of activities, both national and international, should be initiated to promote incentives which will foster a scientific culture of engagement in data care.	**4.13**
A8	Educational materials, guidelines and policy documents for researchers need to be developed and publicised.	**4.7**
A9	There should be increased investment, knowledge transfer, and cross-sector partnerships with knowledge-based and science and engineering industries to capitalize on UK expertise in data curation. This should be led by the DTI.	**4.8**
A10	Investment should be strengthened in those areas of curation research which will enhance data re-use; in particular we recommend focusing on those areas of research needed to establish trust in curated information.	**5.5**

	Tactical recommendations	
B1	Measures to quantify direct and indirect benefits from curation should be developed.	**3.1**
B2	Clear terms of reference regarding the limits of data validation are needed for repositories.	**4.2**
B3	Methods to capture tacit knowledge need to be researched and then introduced, particularly for staff moving off projects.	**4.2**
B4	The responsibilities of the various parties to the curation process should be articulated and communicated to researchers and scientists and those responsible for curation.	**4.13**

	Tactical recommendations	
B5	Where conditions in grants for research stipulate that data should be cared for after the project end, methods to track compliance to these conditions should be introduced.	**4.13**
B6	Data curators should be closely linked to discipline research; their role should include active promotion of their holdings as well as acting as store-keepers and service providers.	**5.1**
B7	An agreed ontology of digital archiving, preservation and curation needs to be developed; this could become part of the research programme (see also A10).	**5.1**
B8	Tools need to be developed to enable forward planning in repositories, and appropriate communication mechanisms should be established between funding agencies, researchers, and repositories.	**5.2**
B9	Data not accepted into a national or other archival repository should be kept securely for a minimum period. Information on this data should go into a clearing house pool, with metadata enabling discovery and evaluation pending final arbitration of its future value or role.	**5.2**
B10	Research should be undertaken to create tools to evaluate dormant data for re-use.	**5.2**
B11	Research is needed to improve automated support tools to prepare data and metadata for archiving.	**5.3**
B12	Consider requiring institutions to report on their curation activities, including management of digital assets (including primary research data), as a heading in future research assessment and institutional risk-management assessment exercises.	**5.3**
B13	Seminars and forums should be held for institute directors, vice-chancellors, faculty heads; the benefits of trusted data repositories need to be articulated to the wider scientific and research community and their support staff.	**5.3**
B14	Research is needed to investigate further the curation of data subject to structural as well as content change.	**5.4**
B15	Consideration should be given to the adoption of quality markings (“kitemarks”) for both data and for data repositories. This could be linked to development of codes of practice for curation centres.	**5.5**

	Tactical recommendations	
B16	Institutions, Research Councils and universities acting alone or in consortia should be encouraged to adopt OAIS-compliant archives and to institute methods to assess curation of academic digital assets.	**5.6**
B17	To address the risk represented by cross subsidies, activity-based costing analysis might be considered to identify particular areas of vulnerability. Vulnerability arising from cross-subsidies is a feature of the current short-term funding system (see A3).	**6.1**
B18	Senior policy makers should be aware of possible negative repercussions of legislation aimed at supporting entrepreneurial activity.	**7.1**
B19	The international links established within the e-Science programme, e-Science projects, JISC's Digital Preservation Focus and other JISC programmes should be continued, and opportunities for further collaboration should be encouraged with relevant fora and bodies in the USA (for example, the NSF) and Europe.	**7.3**

In the following matrix (Table 12) we show where we believe the strategic-level recommendations (the A series in the discussion above) address the findings listed in 8.1.

Table 12: Where recommendations will make an impact on the findings

Key: ✓✓ Have a major impact ✓ Have an impact		Strategic recommendations									
		A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**
		Strategic-level advocacy and set UK agenda.	A curation task force to inform and guide this agenda.	Provide specific, new funding stream(s).	Fund exemplary applications of curation.	Support for the Digital Curation Centre.	Determine criteria for what data to keep and why.	A programme to encourage incentives for researchers.	Educational materials, guidelines to be developed and publicised.	Invest in cross-sector partnerships with industry.	Investment in research to establish trust in curated information.
Findings											
1	Confirmation there is an issue, but the UK is not fully prepared to capitalise on the opportunities.	✓	✓			✓		✓		✓	✓
2	There is a lack of a visible, central strategy for data stewardship.	✓✓	✓			✓		✓		✓	✓
3	Curation is competing with research funding.			✓✓							
4	The current short-term funding models for the provision of curation are antithetical to its long-term nature.	✓		✓✓							
5	Protect high investments in instrumentation and experiments by curating the data produced.	✓	✓	✓	✓	✓	✓	✓	✓		✓

Findings		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
6	Graduated investments in different datasets needed.		✓		✓		✓✓		✓	✓	✓
7	Provision of curation is patchy.	✓	✓	✓		✓		✓	✓	✓	
8	Mismatch between funding requirements to keep data and funding made available. Also poor guidance.	✓	✓	✓		✓	✓	✓	✓		
9	Lack of awareness by scientists is putting valuable data at risk.	✓			✓✓	✓	✓	✓	✓✓		
10	Few incentives for scientists to co-operate.	✓	✓		✓			✓✓	✓		
11	Some areas of curation are still in the research and proof-of-concept phase.		✓	✓	✓	✓✓	✓			✓	✓✓
12	Issues of trust must be addressed before data-based science can flourish.				✓	✓		✓	✓		✓✓
13	There is little interchange and sharing between the curation experiences of science-based industry and the academic sector.	✓	✓			✓✓		✓		✓✓	

8.3 Timelines

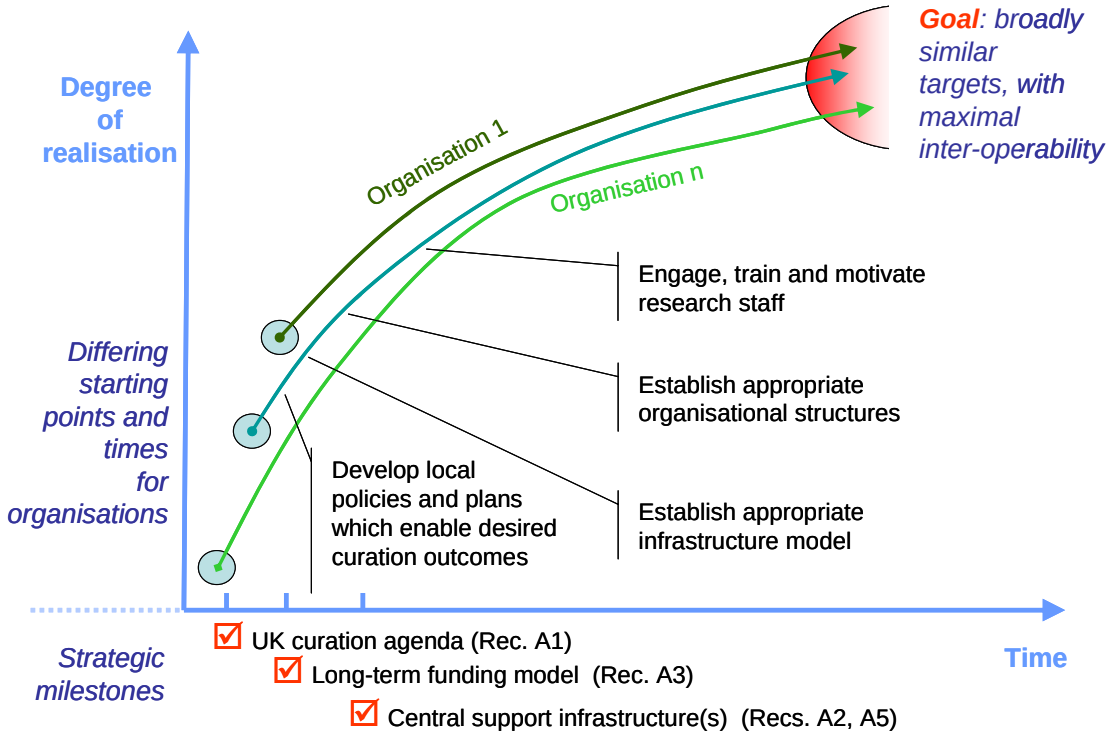
To enable data curation to become part of the way in which science is done, in practical terms it is also necessary to establish:

- local, organisation-level policies and procedures operating within the wider national agenda;
- appropriately curated data collections, with persistent links to related publications and other related resources;
- technological infrastructures to store and make data accessible using appropriate architectures;
- appropriate measures to ensure rewards to, and compliance from, the research community, including awareness raising and training programmes.

We use the word "appropriate" because definition is a matter for others, according to science, the IT characteristics of the underlying data and processes, and according to policy. Since our report covers such a wide organisational field, with different structures – seven Research Councils, the higher educational establishments in the UK university sector, specialist research institutes – we have not attempted to identify one plan which will establish data curation practices across all of these; furthermore, different organisations are at differing degrees of advancement towards curation (applying different terminology), with different driving factors. Organisations and research domains are at different points on what might be called the "curation road".

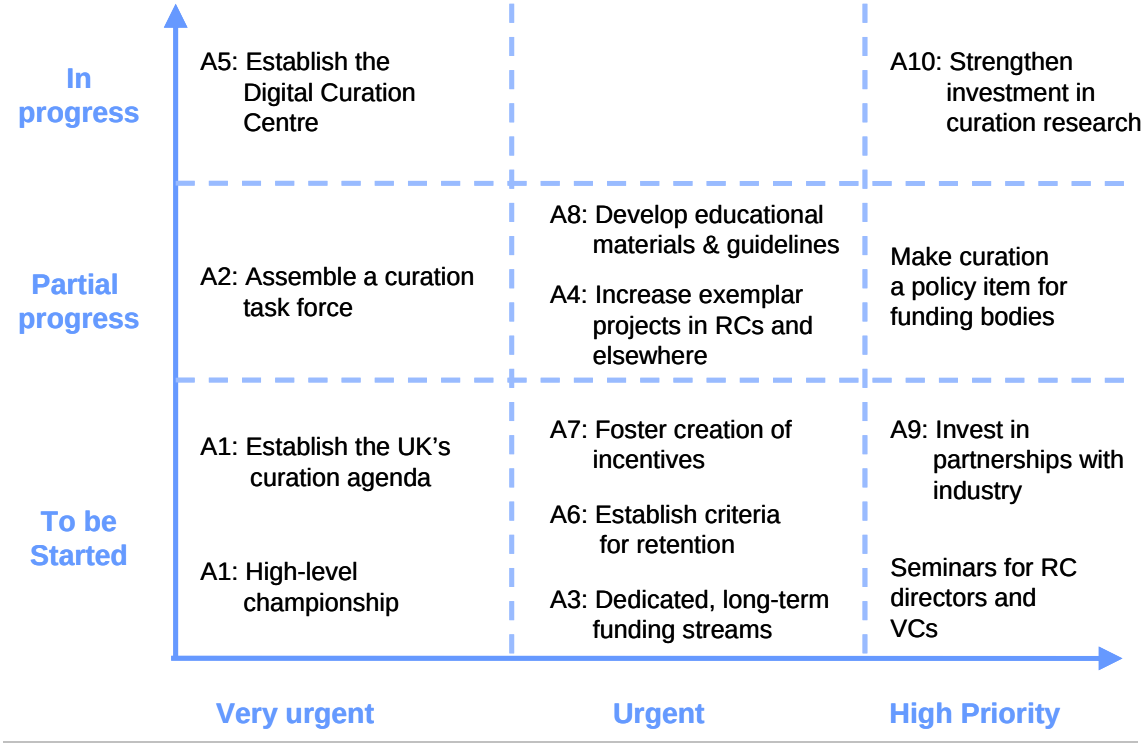
However, some general pointers show how different organisations may strive towards an ideal common goal with maximal interoperability. Figure 14 overleaf illustrates this notionally. It is predicated on fulfilling our major, strategic recommendations, particularly those concerned with defining the UK's data curation agenda (the goal), establishing a stable curation regime, and establishing central support facilities.

Figure 14: Organisations' routes to fulfilling the curation agenda



Targets and milestones should be set in a general framework for the whole of the UK academic science sector, as set out in our main recommendations that are listed in the Executive Summary.

Figure 15: Prioritisation of strategic recommendations



We have suggested some relative priorities for these in the diagram above, which summarises these along axes of urgency (x-axis) and current degree of completion (y-axis). Numbers in parentheses indicate the recommendation numbers in the executive summary: two others actions have been added: making curation a policy item within the funding bodies and informing and preparing the leaders of the Research Councils and the universities.

8.4 References

Atkins, D. et al, 2003, Revolutionizing Science and Engineering Through Cyberinfrastructure: National Science Foundation, <http://www.cise.nsf.gov/evnt/reports/atkins_annc_020303.htm>

Atkinson, M., Crowcroft, J., Goble, C. et al. 2001. Computer Challenges to emerge from eScience.

BBSRC Annual Report 2001-2002.

Beagrie, N., 2002, A Continuing Access and Digital Preservation Strategy for the Joint Information Systems Committee (JISC) 2005-2005. <http://www.jisc.ac.uk/uploaded_documents/dpstrategy2002b.rtf>

Berners-Lee, T. with Fischetti, M., 1999. Weaving the Web, Harper Collins.

Buneman, P. and Steedman, M. 2002. Annotation – the new medium of communication. (Extract of workshop).

CCCLRC Annual Report (full) 2001-2002.

CCCLRC Quinquennial Report. 2002.

Cedars, 2002a. Cedars Guide to Intellectual Property Rights . See <http://www.leeds.ac.uk/cedars>

Cedars, 2002b. Cedars Guide to Technical Strategies. See <http://www.leeds.ac.uk/cedars>

Consultative Committee for Space Data Systems, 2001. Reference Model for an Open Archival System. (Recognized as ISO 14721, 2003).

Corti, L. and Wright, M. 2002, MRC Population Data Archiving and Access Project. Medical Research Council, report prepared by the UK Data Archive, University of Essex.

David, P.A., 2000. The digital technology boomerang: new intellectual property rights threaten global “open science”. World Bank Conference Paper.

De Roure, D., Jennings, N. and Shadbolt, N., 2001. Research agenda for the semantic grid: a future e-Science infrastructure.

Department of Trade and Industry, Business Finance and Investment Unit 2002. R&D Scoreboard 2002. <http://www.innovation.gov.uk/projects/rd_scoreboard/introfr.html>

Department of Trade and Industry, December 2002. Science Budget 2003-04 to 2005-06.

DLM Forum (1997). Guideline on Best Practice for Using Electronic Information, <http://europa.eu.int/ISPO/dlm/documents/gdlines.pdf>

Dollar, Charles M. 2000. Authentic Electronic Records: Strategies for Long Term Access, Cohasset Associates, Inc., Chicago. (ISBN 0-9700640-0-4)

Duranti, Luciana, et al , 2002. The Long-term Preservation of Authentic Electronic Records: Finding of the InterPares Project, <http://www.interpares.org/book/index.htm>

EPSRC Annual Report 2001-2002.

EPSRC Funding Guide 2002.

EPSRC Guide to Good Practice in Science, 2002.

EPSRC Research landscape 2002.

ESRC Annual Report and Operating Plan 2001-2002. ESRC Strategic Plan 2001-2006.

ESRC Data Policy, April, 2000.

European Commission, 2002. Proposal for a Directive of the European Parliament and of the Council on the re-use and commercial exploitation of public-sector documents. Brussels.

European Commission. Working paper: Workshop report on managing IPR [intellectual property rights] in a knowledge-based economy – bioinformatics and the influence of public policy. Rapporteur Stephen Crespi. November 2001.

Financial Times (Kelly, J. Bahra, P. and Milton, U. et al.) 2002. UK Universities 2002 Survey.

Follett, Sir Brian, 1993. Joint Funding Council's Libraries Review Group: Report (The Follett Report). See < http://www.ukoln.ac.uk/services/papers/follett/report/>

Food and Drug Administration, 1997. “Electronic Records; Electronic Signatures”, Federal Register; Vol. 62, No. 54, 13430 – 13466.

Food and Drug Administration, 2002, <http://www.fda.gov/ohrms/dockets/GUIDANCES/DGUIDES.HTM>

Foster, I. and Kesselman, C. (editors), 1999. The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, San Fransisco. ISBN 1-55860-475-8

Foster, I., 2003. Scientific American, April 2003, p66.

Hedstrom, M., 1998. "Digital Preservation: A Time Bomb for Digital Libraries." Computers and the Humanities (31), no. 3, 189-202

Hey, A. and Trefethen, A, 2003. The Data Deluge. In: Grid Computing – Making the Global Infrastructure a Reality, Wiley, January, 2003.
Also: <http://www.research-councils.ac.uk/escience/documents/DataDeluge.pdf>

Higher Education Funding Council for England. Annual Report 2001-2002.

INSAR, 2002, Model Requirements for the Management of Electronic Records (MoReq), European Union, 2002, (ISBN 92-894-1290-9).

International Standards Organisation, 1994. ISO 10303-1, 1994, Industrial automation systems and integration -- Product data representation and exchange -- Part 1: Overview and fundamental principles

International Standards Organisation, 2001. ISO DIS 14721, Space data and information transfer systems -- Open archival information system -- Reference model, <http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>

International Standards Organisation, 2003. ISO 15836, 2003, The Dublin Core Metadata Element Set. (Also ANSI Z39.85)

JISC CEI Interim Preservation Strategy 1998-2001. <http://www.jisc.ac.uk/uploaded_documents/intpreservstrat.pdf>

JISC, Five-year strategy 2001-05, <http://www.jisc.ac.uk/pub01/strat_01_05/exec.html>

JM Consulting, report to HEFCE, 2002. Research relationships between higher education institutions and charities.

Jones, Maggie and Beagrie, Neil, 2001. Preservation Management of Digital Materials: A Handbook. The British Library, ISBN 0-7123-0886-5

Kenney, Anne R. and Stam, Deirdre C., December 2002. The State of Preservation Programs in American College and Research Libraries: Building a Common Understanding and Action Agenda.

Lievesley, D., and Jones, S., 1998. See: <http://www.ukoln.ac.uk/services/papers/bl/blri109/datrep.html>

Lord, P. 2002, A Survey of Information Technology Vendors, Digital Preservation Coalition, <http://www.dpconline.org/graphics/reports/>

Lorie, R. A., Long Term Preservation of Digital Information, JCDL'01, Roanoke, Virginia, June 2001.

Lorie, Raymond, 2002. The UVC: a method for preserving digital documents proof of concept, IBM/KB Long-Term Preservation Studies Report Series 4, ISBN 90-6259-157-4

Lyman, P., Varian, H.R., Dunn, J., Strygin A., and Swearingen, K. 2000, "How Much Information?" School of Information Management and Systems at University of California at Berkeley. See: <http://www.sims.berkeley.edu/research/projects/how-much-info/summary.html>

Macdonald, A. and Lord, P., 2003. Digital data curation task force: report of the task force discussion day, 26 November 2002. See: <http://www.jisc.ac.uk/uploaded_documents/CurationTaskForceFinal1.pdf>

Mann, Williams, Atkinson, Brodlie, Storkey, Williams, 2002. Scientific data mining, integration, and visualization. Report of workshop held at the National e-Science Institute, October 2002.

Matthews, B. and Sufi, S., editor: Kleese van Dam, K. 2001. The CCLRC scientific metadata model.

Moore, R.W., 2000. Knowledge-Based Persistent Archives. Proceedings of La Conservazione Dei Documenti Informatici Aspetti Organizzativi E Tecnici, in Rome, Italy, October, 2000.

Moore, R.W., 2001. Knowledge-Based Grids. Proceedings of the 18th IEEE Symposium on Mass Storage Systems and Ninth Goddard Conference on Mass Storage Systems and Technologies, San Diego, April 2001.

MRC Annual Review 2001-2002. See: < http://www.mrc.ac.uk/prn/pdf-annual_review_01to02.pdf>

MRC, 2001. Research funding strategy and priorities 2001-2004. See: < http://www.mrc.ac.uk/txt/index/strategy-strategy.htm>

MRC. 2002. Vision for the future. See: < http://www.mrc.ac.uk/pdf-vision_draft.pdf>.

National Audit Office, 2002. Delivering the commercialisation of public sector science. Report by the Comptroller and Auditor General. HC580. Session 2001-2002.

NERC Annual Report 2001-2002.

NERC Data Policy Handbook, Version 2.2, December 2002.

OECD, 2002. Measuring the information economy 2002.

Office of Science and Technology, 2001. Science Research Priorities 2001-02 to 2003-04. A note from the Office of Science and Technology to heads of UK higher education institutions.

PPARC Annual Report 2001-2002.

Public Records Office, 1999. Functional Requirements for Electronic Records Management Systems.

Reading University, Rules for the use of University computers and data networks (current).

Research Councils. 2001 Quinquennial Review of the Grant Awarding Research Councils: Implementation.

RLG/OCLC Report, May 2002. Trusted digital repositories: attributes and responsibilities. Research Libraries Group, Mountain View, CA. www.rlg.org.

Rothenberg, Jeff, 1999. Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. Council on Library and Information Resources. January 1999. ISBN 1-887334-63-7.

Rothenberg, Jeff, 1995. Ensuring the Longevity of Digital Documents. Scientific American, volume 272 Number 1, January 1995, 42-7

Rothenberg, Jeff. 2000. An Experiment in Using Emulation to Preserve Digital Publications. Koninklijke Bibliotheek ISBN 9062 59 1442

Royal Statistical Society, 2002. Preserving & Sharing Statistical Material, UK Data Archive, University of Essex, ISBN 0-906805 -02-3

Wouters, P. 2002, Data Sharing Policies, NIWI-KNAW, Amsterdam.

Wouters, P. and Schroder, P., 2002, Policies on Digital Research Data: An International Survey, NIWI-KNAW, Amsterdam.