JISC

e-Research

e-Science Data Curation

November 2004

Digital data – a new knowledge base for research

> *"… These data … will be used for both diagnosis and prognosis…. 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."*
>
> **Professor Michael Ashburner, writing when head of the European Bioinformatics Institute, with acknowledgement to Jim Ostell.**

Research and science are being transformed by accelerating change in information technology, bringing huge increases in computer power and network bandwidth, and an explosion in data volumes and information. A new order of collaborative, inter-disciplinary research is opening, with increasing access to collections of primary research data and information – a new knowledge base, creating new opportunities and horizons for research and discovery.

However the same technology changes also put the data generated at risk. We face serious and complex issues of strategy and policy regarding the creation, management and long-term care of data.

For, some the word "curation" conjures up images of the painstaking preservation and presentation of fragile, artefacts in museums. For the digital age, however, the term has wider significance.

Tomorrow's digital artefacts (and many we are familiar with today) have no existence except as bit streams – whether as a simple text message or the output from some huge and costly experiment into fundamental physics.

This data is extraordinarily fragile, in particular over time. It depends on a hierarchy of constantly and rapidly shifting technologies – hardware, digital storage media, operating systems, applications software and middleware. It also relies on tacit knowledge external to the data. Without due care of the information coded as this insubstantial stream of bits, we will lose our heritage and our knowledge. It will become inaccessible, untrustworthy, or meaningless.

Conversely, the opportunities created not only by the data itself but by its creation are substantial – the ability to maintain links between data, materials and annotations, provenance information, and the existence of a semantic web, brokered by portals, will build a growing encyclopaedia of information and knowledge of extraordinary wealth. Examples of this can already be seen in data initiatives such as the human genome in biology or the virtual observatory in astronomy which are transforming their disciplines.

Promoting good curation and an information infrastructure to capitalise upon and preserve expensively gathered data means bringing together varied technical and managerial resources, and managing these over time. This activity needs to be supported by clear strategies for resourcing and support. In September 2002 the JISC Support of Research committee and the UK's e-Science Core Programme commissioned a report to examine current provision of and future requirements for the preservation and curation of publicly funded research data. This document introduces some of the findings, recommendations and issues discussed in the e-Science Data Curation Report. The full report is available at:
www.jisc.ac.uk/e-sciencecurationreport.pdf

A key finding of the report was confirmation of the need for generic support in the area of digital curation, and firm endorsement for the Digital Curation Centre (DCC). The report identified specific areas of need which the DCC would be able to address, together with areas for research.

A programme for progress and excellence

Unless care is taken of digital research data, the UK will miss the opportunity to capitalise on the e-science (or, more inclusively, e-research) revolution. The authors of the e-Science Data Curation Report believe that digital curation is an area of key importance in this regard. Key strategic recommendations made in their report include:

Strategic recommendations

1 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.

2 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.

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

5 Endorsement of the need for the Digital Curation Centre.

6 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.

7 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.

8 Educational materials, guidelines and policy documents for researchers need to be developed and publicised.

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

10 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.

Where we are now

Current provision for the archiving and curation of publicly funded research data is distributed among university-based centres, units operating within the Research Councils, and organisations attached to international research bodies.

The report found that the funding for repositories, archive services and related research was generally short-term in nature and had to compete with research projects

The UK's e-Science Core Programme and the Research Councils, however, have funded and are funding a number of relevant projects which will support data curation, data preservation and data access.

In early 2004 an award of £1.3 million per annum over three years was made by JISC and the UK e-Science Core Programme to a consortium headed by the University of Edinburgh, with the University of Glasgow, UKOLN and the Council for the Central Laboratory of the Research Councils, for the establishment of the Digital Curation Centre.

The Digital Curation Centre aims to promote expertise and good practice, both national and international, for the management of all research outputs in digital format. The Digital Curation Centre will also provide a national focus for research into curation. The Digital Curation Centre itself will not be a data repository but will work with existing centres such as the Natural Environment Research Council data centres and institutional repositories. Substantial work, however, remains to be done by institutions and funders to develop the emerging information infrastructure.

The Government has recently outlined its Science and Innovation Investment Framework 2004-2014. This has recognised the importance of the information infrastructure for science and states:

> *"The growing UK research base must have ready and efficient access to information of all kinds – such as experimental data sets, journals, theses, conference proceedings and patents. This is the life blood of research and innovation. Much of this type of information is now, and increasingly, in digital form. This is excellent for rapid access but presents a number of potential risks and challenges. For example, the digital information from the last 15 years is in various formats (versions of software and storage media) that are already obsolete or risk being so in the future. Digital information is also often transient in nature, especially when published formally or informally on websites; unless it is collected and archived it will disappear. There are other challenges too, navigating vast online data/information resources determining the provenance and quality of the information, and wider issues of security and access."*
>
> **Science and Innovation Investment Framework 2004-2014, H.M. Treasury (2004)**

Development of the Digital Curation Centre and implementation of the other strategic recommendations in the e-Science Data Curation Report will be critical to achieving this vision and addressing the challenges and opportunities identified by the Science and Innovation Investment Framework.

Defining curation

Traditionally, curation is not only concerned with long-term care of books, paintings or other artefacts. It is also about maintaining their integrity and enabling and promoting their availability to appropriate audiences. The same is true for data. For data to remain useful, it may also need maintenance and enhancement. These, with the promotion of data to potential consumers, are two of the key roles of data curation. Curation also requires the skills of the archivist to ensure the data's continuing safe custody: as part of that, the data needs to be preserved against changes in technologies which will make the data inaccessible or meaningless.

The e-Science Curation report differentiated digital curation from archiving and preservation as follows:

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.

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.

Preservation: An archiving activity in which specific items of data are maintained over time so that they can still be accessed and understood through successive change and obsolescence of technologies.

Why keep data?

Underlying the phrase "data curation" is an assumption that data has value. In addition to the public accountability for and access to the outputs of publicly funded research, the e-Science Data Curation Report identified seven major reasons to keep data:

- 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
- Retention of expensively generated data which is cheaper to maintain than to re-generate
- Enhancing existing data available for research projects
- For compliance with legal requirements
- To validate published research results
- For use in teaching

When asked about the value of digital data, the publicly funded UK researchers surveyed, said 79% of primary data and 94% of published data is of value at project end.

Aspects of curation

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, as well as the quality of the data and the metadata (the data about the data). Whilst practice and experience in curation are increasing rapidly, areas of curation are still in a research and proof-of-concept phase. Much research and practical and exploratory activity is being undertaken in the UK, of world-class quality. Areas for further research, debate and action include:

- **Trust:** Will a calculation today give the same result made on an older computer architecture? Was the original data correct? Is it still an authentic copy? The greater the confidence we have in the data, the more we are likely to use it, and vice versa. Trust can be enhanced by the existence of qualified domain specialists who curate the data. The more the data is used, the better the return on investment; the higher the quality of the data, the lower the risk.
- **Utility:** As well as confidence that the data has maintained its integrity, we need certain information about the data – where it came from, how it was generated, for example – to enable future users to gauge the utility and reliability of the data, and indeed any annotation of the data. Data utility also depends on the ability of users to manage and analyse it; data mining tools and algorithms, visualisation tools, user interfaces and portals will play a crucial role in accelerating research.
- **Discoverability:** How will future users find data, in particular data they do not know exists, in other domains, or archived according to terminology which has fallen out of use? Data access is often organised through portals; how will those portals be organised? What tools will users need to read or use the data, and who will provide these tools?
- **Access management:** A significant proportion of data involves confidentiality issues. Ownership and rights management also need to be taken into account. These access questions can be particularly difficult across

national boundaries. All these aspects must be respected and managed as part of the data curation.

- **Selection:** What criteria should be applied when selecting data for longer-term retention? How do we know what we should keep? Who sets the selection criteria? How can selection be assessed, when, how often, by whom? Or should we keep everything?
- **Heterogeneity:** Not only is this data revolution creating a deluge of data, the data itself comes in very many different and often specialist formats, some created by the researchers themselves.
- **Complexity:** The data can be composite in nature, with links to external objects and external dependencies (such as calibration information), and be highly complex in structure. This complexity represents a significant challenge for the preservation of data.

The work being carried out and the tools being developed, such as those in the e-science projects, will contribute to the practicality, economics and thus viability of data curation. This work is also important for funders: on the one hand it lightens the cost burden entailed in keeping data, and on the other it can protect the value of data generated in research.

Good data management practices on the part of users will reduce the burden of work required for preservation and curation, and also for the acquisition of the data at the stage it enters a repository. While guidance is extremely important, this management will consume some resources by those generating data, and this needs to be reflected in funding. Again, tools which facilitate and encourage this process will mean that less funding is needed and a greater degree of compliance on the part of the user.

Organisational model

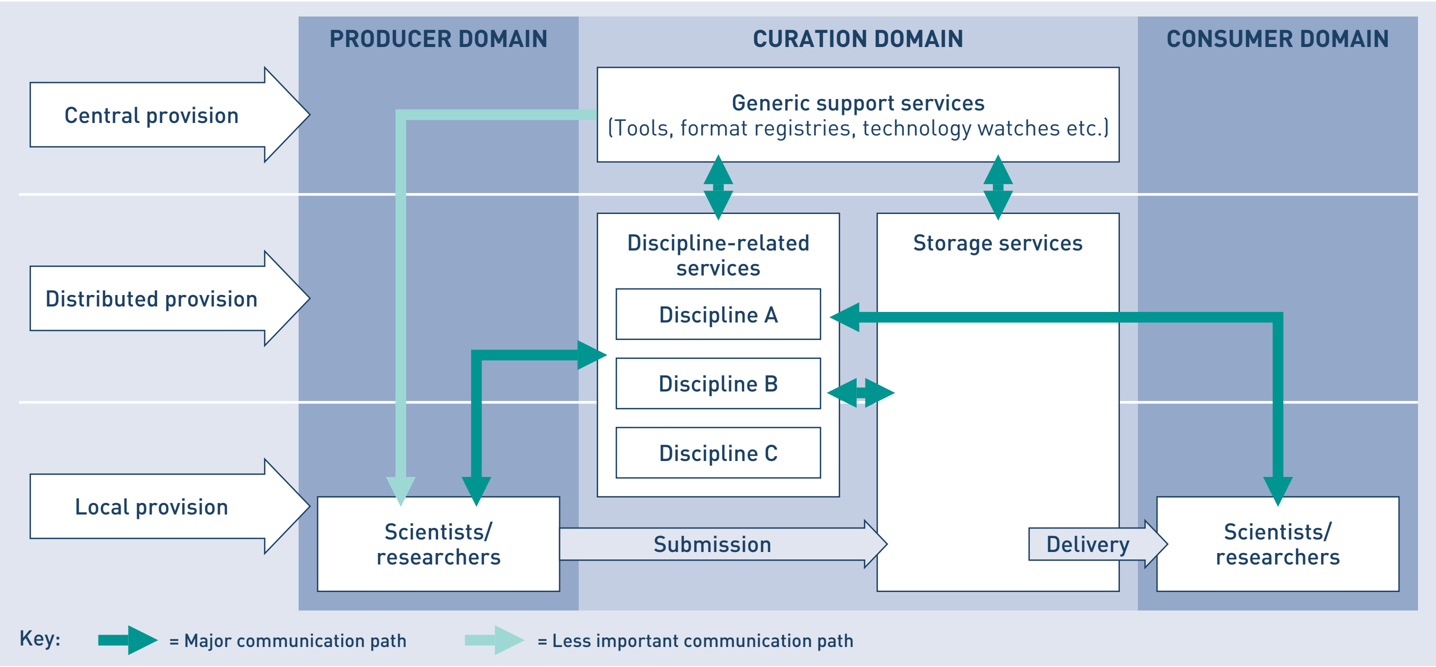
The diagram below provides a simplified model for data curation in the UK.

Viewed on the left is the data producer domain, where scientists and research workers in their various institutions create data, and who will submit this data to organisations providing curation services, shown in the centre column. Data curated within this curation domain will be accessed by researchers and others (consumers), as shown on the right. Initially consumers may be contemporaneous with the submitters, but eventually will be separated from the producers by many years.

Producer and consumer communities are seen as operating locally in their institutions, research groups or units. Some curation services will be best provided at this level too, whether related to the **content** of data (and therefore in general related to specific disciplines) or its **physical management and storage**. In other cases it might be better to provide such services in a distributed fashion – perhaps serving many users within a single discipline spread across different institutions. Other curation services might be more efficient if provided centrally, particularly those shared across communities. An example is the provision of tools to manage the preservation of data held in very common software formats.

Acknowledgements

The e-Science Curation Report and this summary were prepared by Philip Lord and Alison Macdonald of the Digital Archiving Consultancy. The authors wish to acknowledge the significant contribution made by Neil Beagrie, leader of the steering committee for the report, Dr. David Boyd, Fred Hopper and Professor Alan Rector of the report steering committee, and Professor Tony Hey.


PRODUCER DOMAIN
CURATION DOMAIN
CONSUMER DOMAIN
Central provision
Distributed provision
Local provision
Generic support services
(Tools, format registries, technology watches etc.)
Discipline-related services
Discipline A
Discipline B
Discipline C
Storage services
Scientists/ researchers
Submission
Delivery
Scientists/ researchers
Key:
= Major communication path
= Less important communication path