From Data Deluge to Data Curation

Philip Lord, Alison Macdonald, Liz Lyon, David Giaretta

The Digital Archiving Consultancy Limited and the Digital Curation Centre

Abstract

e-Science – or e-Research - enables new forms and layers of research. It generates massive amounts of data, at different research stages. Yet the many technologies used also transform data and put its integrity at risk. Readability and usefulness are jeopardized not just by technical factors. Data's future quality – richness, trustworthiness –is a function of investment in it. But should all data be kept? What other issues arise, for whom? We highlight findings of the recent e-Science Data Curation report commissioned by JISC with the support of the e-Science Core Programme, and present the Digital Curation Centre, the first of its kind in the world, and its role in providing resources and support for digital curation and research.

1. Introduction

The volume of data being created is growing at an astonishing rate[i]. E-Science, or perhaps more inclusively e-Research, 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. This is the knowledge base of research.

There are challenges, however: these same technology changes and the flexibility in use of information technology tools put the very data they create and transform at risk and raise serious and complex issues of strategy, policy and practice regarding the creation, management, and long-term care of data – its curation.

A recent study[ii] commissioned by the JISC Joint Committee for the Support of Research showed that much needs to be done at all levels to enable the data which is being created by this revolution to remain available and valid to future researchers. And much is being done by the e-Science community, in projects, research and other initiatives, and which will be reported at the e-Science All Hands Meeting of 2004. As part of their response to this problem, the JISC and e-Science Core Programme are jointly funding the newly established Digital Curation Centre (DCC)[iii]. Its remit is to provide practical guidance and outreach concerning data curation, and to undertake research into digital curation. The DCC is the first initiative of its kind in the world, and is expected to become a centre of excellence in the area.

In this paper we highlight some of the technical, strategic and policy findings emerging from the e-Science Data Curation report and discuss the DCC's role in addressing some of the practical challenges to be addressed.

2. e-Science Curation

This is a relatively new field, and terminologies are not yet stable. We have used the following working definitions of three key activities:

- **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 other published materials.
- **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 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.[iv v]

2.1 Survey Findings

The e-Science Curation report surveyed and reported on the provision of curation for e-Science data in the UK, listing some 13 major findings, and making ten major recommendations for action at a strategic level.

Strategic and policy level findings are not presented in detail here, but in summary they showed that:

- Urgent action was needed for the UK to capitalize on the opportunities presented by e-Science.
- Action was needed to address a short-term funding regime which mitigates against the essentially long-term needs of data curation.
- Before data-based research can flourish, questions of trust in data (and data as it ages) need to be addressed, such as security, confidentiality, ownership, provenance and authenticity.
- Awareness of long-term data curation was generally low among research workers, and researchers need to be encouraged to engage more in the curation of their own data.
- Provision of services for curation tended to be patchy, but was more advanced in some areas – particularly areas concerning the bio-sciences and in "big", collaborative science such as astronomy and particle physics.

Areas for further research, debate and action include:

Preservation: How is data to survive the constant changes in information technology, which sees the rapid obsolescence of hardware architectures and software and file formats? How do we decide to keep what, and how? Various proposals have been made for addressing this problem, but the area remains one where more work is required, both theoretical and practical.

Awareness and compliance: The viability of data over the longer term depends on awareness. This means that the originators of data, or of data annotation need to be aware of the issues of preservation and curation, and they also need to be given practical guidance to be engaged in the process. Forums such as the e-Science programme and the All Hands Meetings are opportunities to spread the curation word, and to encourage our audience to do so too. Of course, there needs to be awareness at all levels.

Trust: As we noted above, in a digital environment it is not obvious how to engender trust in data which has been passed on to us. How can we be sure of its provenance, its quality, freedom from corruption, and its continued privacy and security (where that is an issue – as in medical science)? We need to determine to what extent these are real issues, and for which data. Work is proceeding on a number of fronts – examples include the work being done by Professor Buneman on databases and the provenance question, or the Qurator project, looking at tools to help discover and document the quality of information resources. However, we are still a long way from complete solutions.

Data selection: What criteria should be applied when selecting data for longer-term retention? Some data is obviously of unique value, but what else should be kept? Selection introduces uncertainties – how do we know what we should keep? Questions of costs and risks arise. Who sets the selection criteria? How can selection be assessed, when, by whom? Or should we keep everything, bearing in mind the costs of maintaining it (its curation)?

The work being carried out and the tools being developed such as in the e-Science projects will contribute to the practicality, economics and thus viability of data curation. Thanks to data grids, portals, defined taxonomies, ontologies, users will be able to discover data resources (which may include the metadata about data) without having to worry about loading the data, establishing its reliability, or not finding it in the first place because of a spelling error.

This work is surely 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.

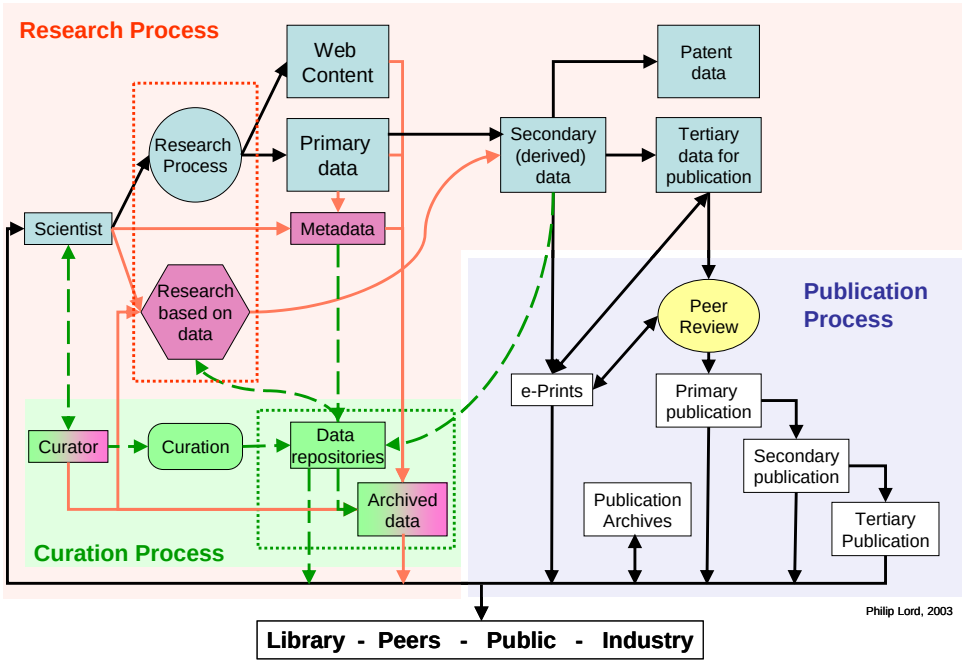
Grid infrastructure provides a distributed computing environment which facilitates the creation and analysis of large volumes of data from e-Research experimentation and applications. It creates opportunity for versatility in model, as well as opening the knowledge base described above to much broader research communities.

Data curation is an emergent field and an exciting one, with many current areas of active research.

2.1 Curation report recommendations

The report's findings led to endorsement of the creation of a Digital Curation Centre in the UK as part of the national provision of curation facilities. Since the report was drafted the DCC has become a reality, and its programme of work is described briefly in section 4 of this paper. Of the other nine strategic

Figure 1 Model of the Curation Process



recommendations made, three are of direct relevance to the DCC's programme:

- The production of research led-exemplars to demonstrate and promote benefits of curation should be co-ordinated by the DCC.
- National and international activities should be initiated to promote incentives which will foster a scientific culture of engagement in data curation.
- Educational materials, guidelines and policy documents for researchers need to be developed and publicized.

3. A curation model for e-Sciences

The accompanying diagram (Figure 1) shows a model of the newly emergent research knowledge cycle. This has three major components: research and data creation, publishing, and the maturing area of curation.

In this model the traditional cycle of research findings going through the publications process and back to consumers, the research community and other consumers (peers, libraries, the public and industry) is shown to the top and right of the information flow diagram (indicated in blue and white on the diagram, and referred to as Level 1 Curation in the report).

More recently, on the research side, this has been augmented by research methods based primarily on the re-use and interpretation of data held in archives (indicated by red in the diagram, and referred to as Level 2 Curation in the report). This somewhat enhanced cycle is exemplified by the work done by social scientists re-using data held in repositories such as those held by the UK Data Archive (UKDA) at the University of Essex; similar models also appear in the life sciences, and within the arts community too, with the Arts and Humanities Data Service (AHDS) – a distributed resource with a central base at Kings College, London.

Another example is the astronomical domain where there are two types of data collection which are common: observatory mode - where data is taken on behalf of the observer and is processed by the observatory system, as opposed to principal-investigator mode where the observer has hands-on control and processes data him/herself. The latter case is more likely to pose problems with archiving and curation.

We are now entering a phase where a third level of curation is demanded. In this matured situation, data repositories which are actively curated are a reality, rather than mere archival stores. This new part of the information cycle is depicted in the lower left of the diagram (in green). In this phase the data is not merely stored, but is preserved to overcome the technical obsolescence problem noted above, and is subject to revision and enhancement as necessary, perhaps augmented with tools to assist discovery, (re-)exploitation and presentation, such as the use of ontologies.

We note that accompanying this trend to curation there is a parallel movement of provision of enhanced bibliographic facilities in digital libraries, and even more significantly for the scientific information cycle, there is an increasing role for enhanced electronic pre-print services (e-Prints) and electronic delivery of completed articles. This trend has been described in other work sponsored by JISC in its initiatives under the e-Research Programme[vi] and in the Digital Preservation and Continuing Access Strategy[vii viii ix].

A good example of a curated resource at this level is the UniProt/Swiss-Prot Protein Knowledgebase. UniProt/Swiss-Prot[x] is an annotated protein sequence database, which was first established in 1986.

The knowledgebase contains curated protein sequence information that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases. It is a "one-stop shop" that allows easy access to all publicly available information of protein sequence annotation. It is maintained collaboratively by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). It employs approximately 100 scientist in the curation process. Release 43.6 (21-Jun-04) of the knowledgebase contains 153320 sequence entries, comprising 56,402,618 amino acids abstracted from 117,067 references.

4. The Digital Curation Centre

The DCC, was awarded funding from 1st March 2004. It is based at the National e-Science Centre in Edinburgh and the consortium comprises four partner institutions:

- University of Edinburgh (lead, Informatics, Law, Information Services and research institutes)
- University of Glasgow (HATII and Information Services)
- UKOLN, University of Bath
- Council for the Central Laboratory of the Research Councils (CCLRC).

The DCC aims to provide a comprehensive advisory service, a repository of user tools and knowledge base, outreach and dissemination activities including an e-journal and an innovative research programme.

The DCC is also forming an Associates Network to provide a forum for engaging with the communities of practice and with key organisations working in this area.

The Centre is currently gathering information and feedback from disciplinary representatives and users, which will inform the research and development initiatives of the Centre and will begin the process of building a user base and community network.

The DCC is also developing an “Approach to Curation” which will inform and provide underlying principles and technical standards for the curation activity. The DCC is monitoring existing architecture work and developments elsewhere with the aim of positioning the DCC research and development programmes within the wider landscape. Further information about the DCC is presented in a separate AHM poster[xi].

5. Conclusion

New avenues of research within which digital data and its continued care and enhancement are central are now emerging. We can expect to become part of the mainstream research in a few years. To take best advantage of this nationally and to contribute fully internationally, strategic and policy level recommendations have been recommended. These initiatives are required both on management and technical fronts. Action has already been initiated on some of these, most notably with the founding of the Digital Curation Centre this year, with the objectives of supporting the scientific community in taking best advantage of new opportunities.

References

[i] Tony Hey & Anne Trefethen, 2003. The Data Deluge. In: Grid Computing – Making the Global Infrastructure a Reality, Wiley, January 2003. Summary: JISC Senior Management Briefing 2004.

[ii] Lord and Macdonald, 2004. Data Curation for e-Science in the UK,. See: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

[iii] See: http://www.dcc.ac.uk

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

[v] Cedars, 2002. Cedars Guide to Technical Strategies. See http://www.leeds.ac.uk/cedars

[vi] eBank UK project, See: http://www.ukoln.ac.uk/projects/ebank-uk/

[vii] Jones, Maggie, 2003, *Archiving E-Journals Consultancy - Final Report*: http://www.jisc.ac.uk/uploaded_documents/ejournalsfinal.pdf

[viii] James, Hamish, et al, 2003, *Feasibility and Requirements Study on Preservation of E-Prints*: http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf

[ix] Parker, Elizabeth, 2003, *Study of the Records Lifecycle* (revised edition of original first published 1999) (Joint Information Systems Committee). See: http://www.jisc.ac.uk/index.cfm?name=srl_structure

[x] European Bioinformatics Institute, 2004. See: http://www.ebi.ac.uk/swissprot/

[xi] Giaretta, D. Robinson, B., Lyon, L, 2004. Curating for the Future – the work of the Digital Curation Centre.. AHM 2004 Poster.