

The Data Deluge: Preparing for the explosion in data

Briefing Paper

November 2004

Vast increases in computing power now enable new forms of science and the creation of huge volumes of data.

This computing revolution brings enormous opportunities for research, science and technology. To exploit these opportunities, technical, strategic and organisational issues need to be addressed by research and education institutions.

The volume of data generated in research and by scientific instruments will soon dwarf all the technical and scientific data collected in the history of research. While differing in volume and characteristics, this data deluge has implications for the way in which research is funded and conducted, and for policies for research. It affects all academic disciplines, from humanities to engineering and science.

Until recently commercial databases have been the largest data collections stored electronically for archiving and analysis, but soon the volume of data in scientific and technical data archives will vastly exceed that of commercial systems. This watershed will bring challenges and opportunities.

The terms 'e-science' and 'e-research' are used to represent the increasingly global collaborations of people and of shared electronic resources needed to solve the new challenges of science, engineering, medicine, arts and humanities. These problems range from the simulation of whole engineering or biological systems, to research in bioinformatics and the environment. e-Research is enabled by huge improvements in computer networks, creating the ability to harness the power of linked computers and shared storage. The IT infrastructure that will make such collaboration possible in a secure and transparent manner is often referred to as the 'Grid'.

The scale of collaborative research ranges from local to global and crosses discipline boundaries. While many computational problems can be satisfied by the deployment of inexpensive networked computers at university, departmental and research group level, an increasing amount of research requires large-scale facilities provided at international centres.

The data itself is an increasingly complex mix of numerical information, text and images, creating correspondingly complex technical issues for its access, use and curation.

From data to information to knowledge

How we handle the vast outpouring of scientific, technical and other research data is of paramount importance. It will no longer be possible to manage manually the process of data examination to identify potentially interesting features and discover significant relationships between data.

More sophisticated automation will be required for the management of data and its storage. Automated knowledge management will be needed to explore and exploit data, at several levels. Metadata is key to this capability. We will need to develop search programmes to extract meaningful information from the metadata as well. Clearly, the quality of search engines will only be as good as the metadata which they reference.

Metadata is data about data. Everyday examples of metadata from the paper world are title, author, publisher and date of publication, or collections of metadata in the form of catalogues and directories. Metadata is vital to discovery within data and its subsequent exploitation.

Each of us encounters about a terabyte (1,024 gigabytes – about one library) of data each day directly through our own senses.

| | | |
|----------------------|------------|--|
| 1 million characters | 1 megabyte | A large novel |
| 1,000 megabytes | 1 gigabyte | Information in the human genome (1,000 novels) |
| 1,000 gigabytes | 1 terabyte | Annual world literature production |
| 1,000 terabytes | 1 petabyte | All US academic research libraries |
| 1,000 petabytes | 1 exabyte | Two thirds of annual production of information |

Collaborative e-research projects need to work with common definitions and standards. The existence of standards for metadata will be vital for working with data held in different formats across different databases or archival systems. In order to achieve this, communities and disciplines need to come together to define generally accepted metadata standards for their community. There are many cases where this is now happening, within user communities and in research and development specifically funded to address these issues.

The Data Deluge:

Preparing for the explosion in data

- In astronomy, sky surveys will generate hundreds of terabytes of data per year; these will be federated to create a Virtual Observatory
- The European Centre for Medium Range Weather Forecasting database in Reading held some 330 terabytes in 2002; since 1998 it has seen 82% growth in volume per annum

Data archives and curation

Preservation of data is another crucial aspect. There are many technical challenges to be solved to ensure that the data generated today can survive changes in storage media, devices and digital formats. In collaboration with key partners, JISC has funded a number of important projects and services in this area including the new Digital Curation Centre, the JISC Continuing Access and Digital Preservation Strategy and the Digital Preservation Coalition.

The Higher Education Funding Council for England is looking at the implications of the flood of e-research data for libraries over a ten-year time scale. Over this time, e-research data are likely to be annotated automatically and stored in a digital archive or library. What will the role of libraries be in this context? Could they become the organisations responsible for hosting and curating (digitally) all the research papers produced by institutions and for maintaining the library so that links continue to work? Could some institutions act as repositories for scientific or technical data on behalf of a number of institutions' 'collaboratories'?

What next?

The digital data deluge will have profound repercussions for the infrastructure of research and beyond. Data from a wide variety of new and existing sources will need to be annotated with metadata, then archived and curated so that both the data and the programmes used to transform the data can be reproduced for use in the future. The data represent a new foundation for new research, science, knowledge and discovery. As well as technical challenges, the management of the data deluge raises questions relating to roles and responsibilities. When considering investments in research and computing, and when formulating strategies and policies, the following should be considered:

- Institutional responsibilities for data curation and the promotion of data sharing
- Encouraging awareness and an understanding of the new methods of data-based science
- The need to establish and use standards to facilitate the linking and sharing of data within and between groups and institutions
- Measures to raise the awareness of researchers and other staff to the challenges and opportunities which will arise

This briefing paper has been produced with the help of Philip Lord and Alison Macdonald and is based on a paper written by Hey, A. J. G. and Trefethen, A. E. (2003) 'The Data Deluge: An e-Science Perspective', in Berman, F., Fox, G. C. and Hey, A. J. G., Eds. *Grid Computing - Making the Global Infrastructure a Reality*, chapter 36, pages pp. 809-824. Wiley and Sons.

Further information and resources

Digital Curation Centre

It is essential that valuable information created by and for researchers in the UK is stored in both a manageable and a durable format that will support access both for specialists and for the wider research community. The Digital Curation Centre will provide a focus for research into data curation issues and to provide training and advice on tools and best practice to support effective digital curation for education and research across all disciplines. This project is hosted at the University of Edinburgh and is co-funded by JISC and the e-Science Core Programme.
<http://www.dcc.ac.uk>

Text Mining Centre

Researchers in the biological sciences are generating vast amounts of academic literature and data held in electronic text format which are not easily assimilated outside, or even sometimes within, their subject areas. Improved use of this material depends on automated tools that can search large quantities of text and extract semantics from it. Intelligent text mining promises to enable researchers to excavate richer seams of electronic research material, including drawing up precise and tailored summaries personalised to the researcher. JISC is working in conjunction with the Biotechnology and Biological Sciences Research Council (BBSRC) and the Engineering and Physical Sciences Research Council (EPSRC)

to fund and develop a UK Text Mining Centre based at the University of Manchester. Potentially, the centre's work will extend beyond the biological environment.

www.nactem.ac.uk

e-Science Curation report

This report, sponsored by JISC and the UK e-Science Programme, discusses the wider strategic and policy questions, including funding, raised by the issues highlighted in this briefing.
http://www.jisc.ac.uk/uploaded_documents/e-Sciencefinaldraft.pdf

Digital Preservation Strategy

JISC is involved in many of the initiatives that are addressing the issues discussed in this paper. This includes the JISC Continuing Access and Digital Preservation Strategy and its implementation plan.
<http://www.jisc.ac.uk/e-sciencecurationreport.pdf>

Digital Preservation Coalition

JISC is a founding member of the Digital Preservation Coalition. The Digital Preservation Coalition (DPC) was established in 2001 to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally to secure our global digital memory and knowledge base.

www.dpconline.org