BRIEFING PAPER October 2006

Curating e-Science Data

The term 'e-Science' commonly refers to large-scale scientific collaborations carried out over the 'Grid', a technical architecture and infrastructure for co-ordinated and distributed sharing of data, resources and communications. e-Science methodologies increase capacity and capabilities, and are rapidly transforming not just science, but also medicine, engineering, and business; their effect is thus near-global. Typical and generic data types range from observational data, to large-scale experimental data, simulation, modelling, and design.

Curation of data collected and developed during these investigations is vital for post-analysis results verification, further experimentation and cumulative analysis. Yet despite its importance, usually only a very small percentage of outputs are properly managed and curated for re-use. Failure to properly curate means that investments are not maximised, research cannot be validated or reliably extended, and may even result in data loss and incorrect interpretation. Vigorous curation practices should be implemented to address these risks, ensure data provenance and integrity, and enable reliable re-use. The scale and importance of the research means a holistic and interoperable approach to curating research outputs is required. Ultimately, this is an issue that can only be addressed on a collaborative scale, like the Grid itself, and requires input from all stakeholders across the entire data life-cycle.

Short-term Benefits and Long-term Value

Curation enables confident and reliable re-use of data. It has immediate and short-term benefits for data creators, researchers, funders and data re-users, in that curation:

- Improves the quality of research data
- Provides access to reliable working data
- Allows conclusions to be validated externally
- Applies good record-keeping standards to data capture including in lab and field electronic notebooks, which enables scientists to draw conclusions from reliable and trustworthy working research data
- Enables large amounts of data to be analysed and developed across different locations by maintaining consistency in working practices and interpretations
- Manages relationships between different versions of dynamic or evolving datasets, and facilitates linkage with other related research and between primary, secondary and tertiary data
- Ensures valuable knowledge and data originating from short-term research projects does not become obsolete or inaccessible when funding expires
- Allows data sets to be combined in new and innovative ways, e.g. historic biodiversity data and GIS data can be combined to investigate trends in ecosystem development

Over the longer term, data curation:

- Facilitates the recording of evidence within the scientific process
- Maximises the potential of collected/analysed data and the initial financial investment
- Identifies and ensures valuable and non-reproducible data is preserved and can be re-used for temporally cumulative analysis
- Provides a framework to address technological obsolescence through reliable and controlled data migration and management of persistent, reliable metadata
- Enables provenance of data to be verified
- Allows future users to reliably re-use or draw upon older research data in new research

HE/FE and e-Science Perspective

"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."
JISC Senior Management Briefing Paper, *'The Data Deluge' (2004)*

http://www.jisc.ac.uk/index.cfm?name=pub_datadeluge

Curation in Practice

Responsibilities for good curation are shared between scientists and research groups, scientific communities (with support from funding bodies or commercial organisations), and (supra-)national research groups. These range across the life-cycle of the materials and often require involvement from more than one group. Curation is not, however, simply a technical concern and should also address organisational and cultural issues. These include:

- Ongoing communication between different sites to avoid problems caused by different systems and institutional goals
- Identifying staff at each site with a responsibility to promote curation and curation concerns
- Developing policies and procedures to address and define good practice in data creation, management, transfer, storage, archiving, preservation, access and re-use
- Training and education to ensure compliance with defined good practice

Immediate practical and technical aspects for scientists and research groups:

- Using Open Source Software and Open Standards to facilitate exchange and persistence of data through and across different systems
- Good annotation and creation of metadata to enable re-use of data
- Ensuring primary, secondary, and tertiary levels of research materials are linked and that links are persistent
- Using unique and persistent identifiers and a consistent citation format
- Identifying and selecting appropriate data for long-term curation and access

Longer-term considerations for scientific groups, communities, and funding bodies:

- Monitoring and updating old storage devices – obsolescence of storage devices has already been noted as a particular problem in astronomy and particle physics
- Identification or provision of significant and scalable repository facilities and investigation into potential preservation and access issues regarding data formats, data migration, and re-use requirements
- Establishing processes to validate and authenticate migrated data

The scale and distributed nature of e-Science research means collaboratively-developed, generic, and repeatable technical solutions are particularly beneficial and cost effective. Similarly, organisational and cultural challenges are more easily met by sharing effort amongst partners but must be tailored to the context, requirements, and resources of a given institution or group.

Additional Resources

Hey, T., Trefethen, A. E. *Cyberinfrastructure for e-Science*, Science, 308, 817-821 (2005)
Lord, P & MacDonald, A, *e-Science Curation Report* (2003) http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf
Szalay, A., Gray, J., *Science in an exponential world*, Nature, 440, 413-414 (2006)
Schopf, Jennifer M *What do we mean by the Grid and eResearch?* (2005) http://www.dcc.ac.uk/events/curl-sconul/presentations/j-schopf.ppt
e-Science Core Programme website: http://www.rcuk.ac.uk/escience/