Curation of Scientific Datasets: Trends, Current Initiatives, and Solutions

Malte Dreyer, Heike Neuroth
Max Planck Digital Library (MPDL), Munich, Germany
+49 (0) 89 38602 225
{malte.dreyer,heike.neuroth}@mpdl.mpg.de

Sarah Carrier, Jane Greenberg
School of Information and Library Science, UNC Chapel Hill, Chapel Hill, NC, USA
+01-919-962-8066
{scarrier, janeg}@email.unc.edu

Stephen Abrams, Patricia Cruse, John Kunze
California Digital Library, University of California , Oakland, CA, USA
+01-510-987-0425
{stephen.abrams, Patricia.Cruse,jak}@ucop.edu

Michael Day, Colin Neilson, Alexander Ball, Rosemary Russell
UKOLN, University of Bath, Bath, UK
+44 (0) 1225 386580
{m.day,c.neilson,a.ball,r.russell}@ukoln.ac.uk

ABSTRACT
E-Science and cyberinfrastructure developments present information professionals and researchers with significant curation challenges relating to the management of scientific datasets [1]. Among pressing questions are: What data should be collected for data curation? How can quality control be maintained? And, how can metadata be generated effectively? These and other challenges are made complex, given the diversity of methods by which data are produced, their heterogeneity, and the increasing scale and scope of scientific research projects. Available literature on the topic of data stewardship provides grounding for approaches addressing these problems, yet more work specifically relating to cyberinfrastructure and repository frameworks is required [2]. This international panel will report on current initiatives addressing the management of scientific data, focusing on advances and solutions in the curation of datasets. The reporting will take place in the context of recommendations from funding agencies and international councils [3,4,5], and models for data curation such as the DCC Curation Lifecycle Model [6]. The panel will provide recommendations for the scope and form of the effort required to address the challenge of scientific data curation and the implications for digital curation education.

Categories and Subject Descriptors
H.3.5 [**Online Information Services**]: *Data sharing, Web-based services;* H.3.7 [**Digital Libraries**]: *System issues, Standards*.

General Terms
Management, Design, Reliability, Standardization.

Keywords
Curation, scientific data, cyberinfrastructure, education

1. PRESENTATIONS

1.1 Research Data Curation: Problems and Challenges (Malte Dreyer, Heike Neuroth)
A chief motivation for data curation is discovery and re-use of valuable research data. Research universities and large organizations such as the Max Planck Institute face curation challenges due to the diversity and expanse of data produced. Specifically, they need to address issues of what data should be collected, quality control, curatorial responsibility, trust, and sustainability. An alliance of scientific organizations in Germany has been formed to collectively address these problems. The alliance includes the Deutsche Forschungsgemeinshaft (DFG, the German Research Foundation), the Fraunhofer Society, the Helmholtz Association of German Research Centres, and the Max Planck Society. All of the members have signed a joint national e-infrastructure policy initiative that has six priority areas; one is focusing on "Preservation and re-use of primary research data. [7]" An emphasis of our work is on open data. This presentation will give an overview of ongoing discussions in Germany underlying the alliance, issues and decisions made specific to data curation, and steps to support open access.

1.2 The Dryad Repository Application Profile: Groundwork Towards a Metadata Scheme for Scientific Data (Sarah Carrier, Jane Greenberg)
The Dryad Repository hosts datasets underlying papers published in the field of evolutionary biology and related sciences. Dryad's metadata architecture links data object metadata with publication metadata. The repository metadata team has developed an application profile with functional requirements that include long-term preservation of datasets, object retrieval and reuse, versioning, provenance tracking, instantiations, and the representation of complex relationships between datasets. Dryad's application profile supports the entire life cycle of a data object, starting with its generation, and ensures the long-term preservation of the metadata itself. The application profile is in compliance the

Singapore Framework for application profiles, a framework compatible with the Dublin Core Abstract Model (DCAM). This presentation will provide an overview of our application profile development work, with an emphasis on its support of curatorial tasks, and highlight challenges in complying with the Singapore Framework. Furthermore, we will illustrate the applicability of our work to other scientific endeavors and its integration with the Semantic Web. Issues addressed by the presentation will include the nature of scholarly collaboration in scientific domains, incentives for data sharing, the manner in which data is reused for research, and the central role that metadata plays in successful data stewardship.

1.3 A Micro-Services Approach to Data Curation (Stephen Abrams, Patricia Cruse, John Kunze)

Data curation is a set of activities aimed at maintaining a balance of usability and authenticity of data objects over time. Rather than centering these activities around a preservation repository, we see them spread across a range of access repositories. Relatively quiescent, or even "dark," storage systems are still important tools, but selected curation and preservation services can and should be applied to any repositories with sufficiently highly valued data assets. It follows that such services are inherently not repository-bound. For example, a naming micro-service could supply preservation-ready identifiers for newly born data objects originating in any number of laboratories within an institution or a discipline; an identity micro-service could then host the basic metadata bindings to give descriptive reality to the named object. Among those bindings, deliberately curated data should generate technical metadata as a matter of course during processing first by a characterization micro-service, to supply early feedback on well-formedness, and second by a fixity micro-service, to generate checksums to help in change detection and version management. This presentation will review key components of the micro-services approach to data curation and note some of our current challenges. We will also comment on the impact of this topic on data curation education and preparing professionals.

1.4 Disciplinary and Institutional Perspectives on Digital Curation (Michael Day, Colin Neilson, Alexander Ball, Rosemary Russell)

Abstract models like the DCC Digital Curation Lifecycle embody the concept that the curation of research data cannot be considered in isolation from the wider contexts of scientific research and practice. One aspect of this is the need for curators to engage with the teams and individuals that are responsible for creating data. Some recent studies [5, 8] have begun to identify how curation roles and responsibilities are shared across all of those institutions and individuals that play an active part in the ongoing stewardship of research data, including scientists, institutions, data centres, funding bodies, and the users of third-party data. This sharing of responsibilities for curation emphasises the importance of collaboration, and the need for generic technical and organisational frameworks to support it. In practice, however, the data curation cultures of different research disciplines (and sub-disciplines) are extremely diverse, posing significant challenges for those trying to develop generic (or institution-based) solutions. This presentation will explore these issues with reference to detailed disciplinary case-studies of curation undertaken for the Digital Curation Centre as part of the DCC SCARP (Sharing Curation And Re-use Preservation) project and a feasibility study conducted by UKOLN into the potential for developing a generic metadata application profile for scientific datasets.

2. REFERENCES

[1] Hey, A.J.G., & Trefethen, A.E. (2003). The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox, & A.J.G. Hey (Eds.): *Grid Computing - Making the Global Infrastructure a Reality*, pp. 809-824, Hoboken, NJ: John Wiley & Sons.

[2] Karasti, H., Baker, K.S., & Halkola, E. (2006). Enriching the Notion of Data Curation in E-Science. *Computer Supported Cooperative Work*, *15*, 321–358.

[3] International Council for Science (ICSU). (2004). Scientific Data and Information: Report of the CSPR Assessment Panel. Available: http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf

[4] Joint Task Force on Library Support for E-Science. (2007). Agenda for Developing E-Science in Research Libraries. Available: http://www.arl.org/bm~doc/ARL_EScience_final.pdf

[5] National Science Board. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. Available: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

[6] Higgins, S. (2007). Draft DCC Curation Lifecycle Model. *The International Journal of Digital Curation*, *2*(2), 82-87.

[7] Alliance of German Science Organisations. (2008). Priority Initiative "Digital Information." Available: http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/allianz_initiative_digital_information_en.pdf

[8] Lyon, L. (2007). Dealing with data: Roles, rights, responsibilities and relationships. Available: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf