The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

The DCC / Regional eScience Collaborative Workshop

Martin Donnelly,
eScience Liaison Support Officer,
Digital Curation Centre,
University of Edinburgh

June 2008

Summary

A report from the Digital Curation Centre (DCC) / Regional eScience Collaborative Workshop, held at the eScience Institute in Edinburgh on June 12, 2008. The DCC is committed to forging and strengthening links between the digital preservation and eScience communities via its eScience Liaison team. This event was the fifth in a series of workshops which are being arranged throughout the UK, with a view to helping the DCC learn more about the eScience perspective on data curation, and to giving eScience practitioners an opportunity to influence the current and future work of the DCC. The workshop took data sharing as its central theme.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The Digital Curation Centre (DCC) supports data custodians in the storage, management and preservation of valuable electronic information, with a view to promoting active data enhancement and prolonging long-term access and usefulness. In July 2007, the DCC's Community Development team added eScience Liaison to its remit, aiming to develop the lines of communication between the research and preservation communities, to enable a bidirectional flow of expertise, techniques and best practice which will lead to a well-trained and confident community of data curators, and to identify and benefit from potential synergies between different disciplinary approaches (see also Pryor, 2007).

An important component in this work is the relationship between the DCC and the National eScience Centre (NeSC)[1] and eScience Institute (eSI).[2] As the national eScience Hub and link between the UK's regional eScience Centres, NeSC plays a key role in the network-building endeavour by making eScientists aware of new developments and tools created by the DCC, and in helping the DCC identify innovative eScience projects for potential case studies, testbed experiments, and other forms of collaboration.

This series of one-day workshops enables the DCC to learn more about the eScience perspective on data curation, as well as giving eScience practitioners an opportunity to influence both the current and future work of the DCC. This particular event took data sharing as a key concept, and focused on real examples of data curation challenges and solutions from within the eScience community, consisting of presentations, demonstrations and discussion. The programme was directed at researchers currently involved in aspects of data collection and management, data accessibility and long-term data curation, particularly scientists involved in eScience research projects, those engaged in the process of digital data management, and those wishing to explore opportunities to engage with the DCC and the broader debate around the management of research data.

Report

Welcoming delegates to the event, DCC Associate Director & e-Science Liaison **Graham Pryor** outlined the aims of the day, together with a summary of related DCC activities. A key goal of the DCC in its second phase is to engage with the eScience community more deeply, establishing researchers' curation needs via collaborative events and information sharing. There is a growing need for universities to hold primary data as well as publications, with the primary impetus for this coming from the funding bodies; however, relatively few researchers have the time or the expertise to carry this out adequately or effectively. Pryor defined the DCC's developing role as interlocutor between researchers and their funders, with liaison and information exchange as key inbuilt processes, and as a forum for the sharing of good curatorial practice. One development already being planned is the Edinburgh eScience Exchange, which will provide a portal to the members of the local eScience Community, and could act as a sponsor for further collaborative events.

[1] National e-Science Centre http://www.nesc.ac.uk/

[2] e-Science Institute http://www.nesc.ac.uk/esi/

Pryor ended his introduction with a list of possible themes and discussion topics, which included: current activities in data storage, preservation and curation, and the policies guiding them; types of data, formats and metadata; repositories and content stores; ingest procedures; and other, discipline-specific challenges.

The first presentation came from **Colin Neilson** of UKOLN at the University of Bath and **Esther Conway** of the Science and Technology Facilities Council (STFC), who provided an overview of the DCC's SCARP project,[3] investigating current curatorial practice across a wide range of disciplines with a view to improving understanding of the changing needs of the community. Neilson began by offering a view of curation which mixed traditional, human expertise with machine-driven text- and data-mining, exemplified by the Wikiproteins project.[4] The SCARP team engages in immersive disciplinary studies (Architecture and Engineering, Medical and Social Sciences, Biology, and large Scientific Datasets and Archives), and part of their work has involved mapping the DCC Lifecycle Model to specific disciplines, identifying context-specific differences as well as common ground. Only through application can the model develop and adapt to a range of varying factors such as scale, organisational complexity, working methods, and desired end products.

Conway outlined two case studies with a common purpose of investigating the digital landscape, the first of which is linked to the CASPAR project[5] and covers semantic information, file formats, stakeholder analysis, human factors and intended outcomes. The second case study constituted an immersive engagement with the British Atmospheric Data Centre,[6] involving inspection of ten of the Centre's 147 datasets, manually reading 1000 files and noting over 3000 OAIS relationships. Conway noted that digital curation does not exist in a vacuum; like anything else, it is subject to environmental changes, exemplified by the focus on atmospheric impacts on radio communications during World War II. The group noted the lack of a joined-up European policy on data-sharing, which she felt likely to hinder efforts until that lack is addressed.

Conway ended by sharing the results of a scoping study on digital repositories, which found implementation of Fedora[7] to be prohibitively time-consuming for the Centre, with an estimated programmer workload of around eighteen months. Given the resources available, the University of Southampton's EPrints is likely to be the best solution; it was found that the standard EPrints distribution can be up and running within a week.[8]

[3] Sharing Curation and Re-use Preservation http://www.dcc.ac.uk/scarp/

[4] Wikiproteins http://proteins.wikiprofessional.org/

[5] CASPAR Preservation User Community http://www.casparpreserves.eu/

[6] The British Atmospheric Data Centre (BADC) http://badc.nerc.ac.uk/

[7] Fedora Commons http://www.fedora-commons.org/

[8] Open Access and Institutional Repositories with EPrints http://www.eprints.org/ Conway pointed out that time should be also allowed for functional adjustments to the 'vanilla' version, particularly with regard to subject classification of digital support information objects, which are vital for reuse of scientific data.

Trevor Carpenter then spoke about the curation needs of the Scottish Funding Council (SFC) Brain Imaging Research Centre (SBIRC)[9] and the SINAPSE pooling initiative.[10] SINAPSE is a consortium of six Scottish universities which pursues collaborative research in brain imaging, and offers support and connections to clinical research networks. Carpenter outlined the complex legal and regulatory framework within which SBIRC operates, comprising the Data Protection Act (DPA), the NHS Research and Development (R&D) Ethics Committees, and the UK Department of Health's Medicines and Healthcare products Regulatory Agency (MHRA). The DPA and R&D committees both forbid the reuse of data without consent, and place strict time limits on its retention, while the MHRA requires data to be kept indefinitely for audit purposes. SBIRC proposes to resolve these disparate requirements by using pseudonymous identities with access to subject identifying data controlled by the relevant data controller. An exemplar dataset was the Lothian Birth Cohort of 1921; Carpenter suggested that permission to retain the original assessment data in unanonymised form for around 80 years might have been very difficult to justify had the DPA been in place in the early twentieth century.

Having sketched the regulatory environment, Carpenter spoke about the Centre's curatorial practice. The current policy is for the raw data, which is held in standards-based formats, to be archived in its original form, with responsibility generally held to rest with each project's principal investigator for any derived data or publications. Archiving and curation can be seen as impossible tasks by PIs, and take-up of version control software and open source licences has been disappointing; this partly suggesting a lack of awareness surrounding the importance of good practice, but it is also a reflection of the career pressures under which PIs work. The Centre is beginning to develop standard operating procedures to cover the archiving of project data, the creation of data dictionaries, and the documentation of analysis procedures. One approach being investigated is the development of data management techniques which will mean that the metadata required for curation are created during the active phase of the project, rather than post-hoc.

The presentation concluded with a DCC 'wishlist' of sorts, including advice on the longevity of different data storage media types, and training materials for digital curation which could feed into researcher training programmes, and collaborative involvement in the development of standard operating procedures.

After lunch, **Bob Mann** gave an overview of the University of Edinburgh's Wide-Field Astronomy Unit (WFAU),[11] concentrating on the Unit's past, present, and future approaches to curating sky survey data. The Unit's collections span from scanned images of photographic plates to born-digital images which are captured in Hawaii and Chile, transmitted to England for data cleaning/ artefact-removal, and then piped up to Scotland where the data are held. He also described the global Virtual Observatory Alliance initiative, which aims to make all of the world's astronomical archives interoperable.

Mann offered concrete examples of the rapidly increasing scale of the task facing the WFAU: the existing WFCAM archive ingests in the region of 20 terabytes of

[9] The SFC Brain Imaging Research Centre http://www.sbirc.ed.ac.uk/

[10] Scottish Imaging Network: A Platform for Scientific Excellence http://www.sinapse.ac.uk/

[11] Edinburgh: Wide Field Astronomy Unit http://www-wfau.roe.ac.uk/

image data per annum; the proposed Large Synoptic Survey Telescope (LSST)[12] will capture approximately this amount every night. These data volumes are too great for user download; the key challenge is therefore to make these very large datasets useable. Other challenges included the synthesis of data and knowledge through interlinking archives and the online literature, adding value to the archives via third-party-user-created metadata, and keeping staff abreast of developments (Donnelly, 2005).

The presentation concluded with a list of the astronomy community's needs which the DCC should aim to meet, namely: policy advice for dealing with the funding bodies; technology briefing advice; a clearer outline of the advantages for various stakeholders of collaborating with the DCC; and for further training/ advisory materials to be created and made available online.

Jeff Christiansen of the Medical Research Council Human Genetics Unit (MRC HGU) then spoke about the Edinburgh Mouse Atlas of Gene Expressions (EMAGE),[13] which holds *in situ* data of mouse embryos linked to digital images. The Atlas combines data from multiple sources: over 150 journals; a manually created and curated index of papers and articles from the latter; several large-scale screening projects; and dataset submissions from several laboratories in Europe, North America, and Australia.

EMAGE links virtual reality (VR) models of embryos to an ontology describing the anatomical parts of the embryo at different stages of development. Each record is in two parts: (i) a human-created text; and (ii) a spatial annotation to the VR model. This means that data can be interrogated and mined in two broad ways: spatial-based or text-based analyses.

Having outlined the project, Christiansen highlighted some of the obstacles it faces, notably the access to raw data and reuse of image data from journals where copyright is held by academic publishers. A key benefit of EMAGE is the ability to compare many similar images concurrently, but this functionality is currently hampered by the inability to embed copyrighted images within the Atlas. Only three of the journal publishers are open access (using a Creative Commons Attribution Licence); reaching re-use agreement with the others has to be done individually, and can be extremely time-consuming and expensive for small research teams.

Christiansen ended with an update on the activities of the annual International Biocuration Meetings, which provide a valuable forum for information and experience sharing, and the proposed formation of an International Society for Biological Curation.[14]

[12] The Large Synoptic Survey Telescope (LSST) http://www.lsst.org/
[13] EMAGE http://genex.hgu.mrc.ac.uk/Emage/database/
[14] See http://www.biocurator.org/ for links to the International Biocuration Meetings, and planning progress on the formation of the new Society.

Robin Rice of the EDINA national data centre[15] and Data Librarian of the University of Edinburgh spoke about the University's part in the DISC-UK DataShare project,[16] and in the development of a Data Audit Framework.[17] The DataShare project's aim is the collaborative development of tools and methodologies for academic data sharing within digital repositories (such as DSpace,[18] Fedora, and EPrints), while remaining conscious of the developing policy and technological environments in which the data stakeholders operate.

Rice's presentation incorporated a list of 17 potential obstacles for data sharing in institutional repositories, and a shorter list of five broad benefits. The key message of these lists was that the major barriers are primarily cultural or methodological as opposed to technological, from which it might be inferred that the old ways of thinking about and relating to the custodianship and ownership of analogue resources tend (perhaps predictably) to persist into the digital era.

The Digital Audit Framework has been developed as a response to recommendations made by DCC Associate Director Dr Liz Lyon in the JISC-commissioned report *Dealing with data* (2007), with development work led by another DCC Associate Director, Professor Seamus Ross of HATII at the University of Glasgow. The final part of Rice's presentation summarised lessons learned from the pilot implementations of the five-stage framework across four UK universities, addressing issues of time resourcing, scope and granularity, methodology, and the widespread divergence between policy and practice even in highly data-literate institutions, suggesting that greater automation may be required in order to maximise standardisation and predictability.

In the final presentation of the day, the DCC's **Mags McGeever** spoke about the legal considerations for data sharing , identifying two principal areas of relevance: (i) intellectual property rights (specifically copyright and the Database Right), and (ii) data protection.

McGeever began by noting how many of her themes had been touched upon over the course of the day, confirming their relevance. Continuing on from Robin Rice's observation that the barriers to data sharing tend to be methodological rather than technological, the group heard about potential challenges posed by the legislation together with some of solutions to these challenges). The laws which govern intellectual property linked to electronic data tend in the main to be extensions of those which governed (or 'protected') their analogue forebears, but with a wealth of new factors involved, most notably their potentially dynamic/ ever-changing nature, as reflected in the relatively recent Database Right (1996), an intellectual property right developed exclusively in order to protect databases.

[15] EDINA http://edina.ac.uk/

[16] DISC-UK DataShare Project http://www.disc-uk.org/datashare.html

[17] JISC: Data Audit Framework Development Project http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/dataauditframework.aspx

[18] DSpace http://www.dspace.org/

Beyond intellectual property, researchers (especially those involved in medical or social science disciplines) must be aware of their rights and responsibilities with regard to data protection. The laws covering this seek to maintain a balance between the rights to individual privacy and the benefits which may be gained from legitimate use and analysis.

Conclusions

Summing up, NeSC Deputy Director **Dave Berry** offered his reflections on the day's presentations, noting that while the growth in Open Source and Open Access approaches can circumvent some potentially tricky legal issues – particularly when the developers are collaborating for the common good – this is not always suitable given the prevalence of IPR-centric revenue models.

Thanking the speakers for their insightful contributions to the debate, Berry ended his summary with a provocative question in two parts: what is the environmental impact of data curation, and how do we determine at what stage this impact outweighs the data's worth?

The next collaborative workshop event is expected to be held in Autumn 2008. For more information on this event, or any other facet of the DCC's eScience link-up, contact Martin Donnelly[19].

Acknowledgements

This report is part of the Digital Curation Centre's (DCC) Community Development programme, which is supported by a grant from the United Kingdom's Joint Information Systems Committee (JISC).

References

Donnelly, M. (2005). *Digital Curation Centre case studies and interviews: Wide Field Astronomy Unit (WFAU)*. Retrieved July 29, 2008, from http://www.dcc.ac.uk/resource/case-studies/wfau/

Lyon, L (2007). *Dealing with data: Roles, rights, responsibilities and relationships*. Consultancy report. Retrieved July 29, 2008, from the UKOLN Web site http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#2007-06-19

Pryor, G. (2007). Some challenges for escience liaison. *International Journal of Digital Curation 2 (2)*, pp. 105-110. Retrieved July 29, 2008, from http://www.ijdc.net/

[19] Martin Donnelly e-mail: martin.donnelly@ed.ac.uk