The International Journal of Digital Curation

Issue 2, Volume 2 | 2007

Some Challenges for eScience Liaison

Graham Pryor,

eScience Liaison,

Digital Curation Centre, Edinburgh

November 2007

Summary

The Digital Curation Centre's promotion of expertise and good practice in digital data curation is no mere exercise in theory. Through its new eScience Liaison initiative the DCC has kept a close eye on its founding principle, that the necessity for the physical and life sciences to share access to digital research resources is due mainly to issues characteristic of eScience. This article describes some of the principal liaison activities that have been addressed within that community since the summer of 2007.

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.

eScience and the eScience Community

The UK's Digital Curation Centre (DCC) exists to provide a national focus for research and development in the field of digital data curation. Its approach to building this focus is through the promotion of expertise and good practice, both national and international, in the management of digital research outputs. Plainly, in order to enable a two-way flow of both expertise and techniques, one of the DCC's principal objectives has to be the strengthening of networks and collaborative partnerships within the eScience community, not least to establish a sound platform for developing the skills required for effective digital curation. As a mechanism for achieving that objective, since July 2007, a DCC eScience liaison role has been dedicated to this purpose.

But what is eScience and who comprises the eScience community?

Whilst most researchers today exploit electronic data resources and use IT-enabled processes almost routinely, this does not necessarily equate to the practice of eScience – nor indeed, to the broader term eResearch – yet attempting a definition of eScience that is simple, concise and distinctive seems to have defeated many who are immersed within it. Rather than retreat from the challenge, perhaps a more realisable approach would be to describe the characteristics that together define eScience and the community it supports.

Most particularly, eScience refers to the manner in which scientific (and other) research is executed across a geographically distributed environment – one in which the conduct of research has been enabled by the interconnection of high-speed computers, and where collaboration and the electronic sharing of data are core features of the research programme. Moreover, the defining essence of the term eScience, and what supplies its difference, is an implication that it will not only lead to faster and better research but that it will open up new areas of research that were previously unattainable.

eScience, then, has two inter-dependent aspects: the research undertakings that exploit this new environment, and the technologies and services that support its enabling infrastructure. As a consequence, cross-disciplinary ventures that are typically encouraged by the eScience environment will often include members of the informatics, computing science and data science disciplines.

The StORe and CARMEN Projects

Current illustrations of these two aspects of eScience can be found in the StORe and CARMEN projects, both of which have involved DCC interest and participation. StORe[1], a JISC-funded project that has now entered a second phase, was established to design and pilot a suite of middleware that would link seamlessly between electronic journal articles provided from output repositories, and the data sets from which they were generated, which are held in source repositories. The aim of this Source-to-Output Repositories project was to add value to the research endeavour through the im-

[1] http://jiscstore.jot.com/WikiHome

provement of opportunities for information discovery, and by providing the means for greater visibility and citation of research output. As such, StORe represents some of the features of the infrastructure 'hemisphere' of eScience; but it is interesting to note that, whilst the impetus for the project came from within the research library community, the StORe middleware was designed following an intensive survey of practising researchers within seven scientific disciplines, whose responses and requirements were used to shape the functionality of the middleware.

CARMEN (Code Analysis, Repository and Modelling for e-Neuroscience) is an EPSRC-funded project that aims "to create a virtual laboratory in which data on neuronal activity (electrical and optical measures) can be shared, stored, manipulated and modelled"[2]. CARMEN currently comprises a consortium of twenty academic investigators from eleven universities within the UK, as well as a number of commercial and international partners, although it is anticipated that its community will expand to include new partners as the CARMEN environment matures. The DCC's involvement with CARMEN is through a longitudinal study of the project, which will track the emergence of technological and organisational solutions to a range of discrete data handling problems, all of which are known to have frustrated cross-modal data sharing and integration in the neurosciences.

To an observer of the eScience community, the CARMEN consortium is an interesting entity. It is dedicated to the facilitation of research in neuroscience and it includes a significant body of neuroscientists; indeed, much of the conversation at consortium meetings is focused upon the activities that are core to neuroscience research. Yet this is computational neuroscience, and the consortium crucially includes members from the academic computing science fraternity. It is therefore at once both representative of eScience practitioners and the developers of eScience infrastructure, in a situation where the cross-disciplinary collaboration is also between different research domains.

CARMEN's primary interest in neurophysiological research is to enable the analysis of data from neuronal systems (networks of brain cells), as well as the development of models, to explain both the processes that form the character of these data and the high-level functions they express. Capturing and analysing these data is complex and expensive, with numerous techniques employed in a situation where models may be used to describe the output from many thousands of neurons. As is explained on the CARMEN Web pages[3], the data and models produced by small communities of specialist researchers cannot easily be integrated to contribute to the bigger picture, and data sets are discarded after the experimenter has completed an experimental report, or they are archived in a format that is not widely accessible.

These examples serve to reinforce the position that the DCC has taken in liaising with the eScience community, where it is recognised that the massive amounts of data being generated, transmitted and stored must be afforded appropriate levels of curation if the value being added by eScience to the trusted scientific record is to be optim-

[2] From http://www.carmen.org.uk/

[3] See the CARMEN Project Overview http://www.carmen.org.uk/?q=node/9

ised and preserved. Whilst the CARMEN team may have embarked on the provision of an ambitious and innovative solution, there remain many equally significant challenges elsewhere in the eScience community.

Emerging Requirements

In truth, there are not only challenges but also a range of new obligations. This year the UK research councils are investing three billion pounds of public money in research, a significant proportion of which is being expended under the eScience umbrella. On behalf of the public purse, the councils will expect a reasonable return on their not inconsiderable investment, and we are witnessing the emergence of data sharing and preservation policies aimed at ensuring that this is achieved. Whilst at the moment only some of these policies prescribe the deposit of research data in managed repositories, with others relying on various methods of encouragement, five out of the seven research councils are using financial levers to ensure that data management is taken seriously. Their portfolio of incentives ranges from the provision of a "follow-on fund", to be made available where data has been used to enable commercialisation, to the payment of final awards being contingent upon satisfactory data deposit and, in one case, the stipulation that a costed data management and sharing plan is a pre-requisite to any decision on funding.

Practising eScientists are finding themselves increasingly under pressure as they encounter the provision of resources being dictated by the requirement to capture and curate their data. It is a potentially daunting task, for not only are there huge sums of money invested, but the volume of digital data that is produced is vast. Describing his own research programme, one respondent to the StORe survey spoke of his data output as "one of the largest databases in the world! I think it's of the order of petabytes"[4]. This was not untypical.

The data produced from scientific research is also both complex and dynamic. Another StORe respondent described how "we define the format of our data. It changes in each step. We combine raw data from the detector with calibration constants to produce reconstructed data. We then produce event summary data and then analysis object data etc. etc." When extrapolating these experiences across a geographically distributed research team, and working at high speed over the Grid, the informality of more traditional research methods may be fondly missed.

Legal Issues

For the DCC, the prospect of bringing essentially technical solutions to support this new data landscape must be complemented by the provision of a safe route through the prevailing legislation, since the legal environment for digital curation is perhaps as complex as the data being produced. Some aspects are already commonplace: the StORe survey was informed by numerous expressions of concern over the need to bring clarity to issues of data ownership; the intellectual property rights (IPR) pertaining to deposited and shared digital data are the subject of current investigation by national agencies in the UK and elsewhere. As an eScience tool, StORe was welcomed as a means of expanding the horizons for data access; but by making data more easily

[4] Petabyte – usually understood as 10^{15} bytes (1,000,000,000,000,000 bytes)

available it also introduced new anxieties concerning the opportunities it gave to data predators, as well as the risk of premature dissemination of research results. The resultant crucial privacy issues had to be addressed as a main technical feature of the StORe middleware.

The DCC operates its own legal services unit, and has recently published a Legal Watch Paper on IPR and associated issues, but through liaison with the eScientist in the laboratory we are also encountering a host of practical questions in which technical and legal questions have become inextricably linked. These include the identification of methods for data validation, where there is a need not just for systems to manage data ingest effectively, but also a means of ensuring compliance with those individual and corporate responsibilities that govern the condition and provenance of data uploaded to a repository. In the case of developments such as CARMEN, which involve both system and data integration, legal advice will be essential in order: to explain any obligations arising from the deposition of code and protocols that are subsequently found not to be benign; to confirm that appropriate licences are in place to cover all the installations in a distributed community; and to ensure that the rights of access agreed and applied to raw data are being properly referenced and interpreted when access to synthesised data is being sought. That, of course, is not to mention the further legal nightmare that could arise from the mischievous use of a communal data environment to distribute illegal or offensive material!

DCC Support

The DCC's inculcation and support of good practice in electronic data management is not restricted to eScience projects. Working with national data centres through a series of regular visits, and through agreement in November 2007 to establish a forum of data centres and institutional repositories, the DCC is committed to facilitating the exchange of experience and knowledge. Commencing in Spring 2008, the forum will support six-monthly 'show-and-tell' workshops, when staff from data centres and university digital repositories will reveal their solutions to the riddles of data management and describe their issues of the moment, with the purpose of sharing best practice, achieving robust standards in data curation, and always with an eye to assist in avoiding the totemic 'recreation of the wheel'. The results of the DCC's *Data Centres Synthesis Study* are also due to be published at the end of 2007, which will highlight areas of common interest in advance of the first forum, focusing on issues that are recurrent across the disciplines and with descriptions of the technology and process solutions that have been implemented or are under development.

For the DCC, structured liaison with the eScience community may still be regarded as being in its formative period; yet I believe this brief article has demonstrated that engagement with the UK's eScience programme and with the research councils is already being pursued through a full programme of support activities. These are being undertaken in association with selected eScience and other data-generating projects, in partnership with the established professional services that exist to preserve and deliver research data and, finally, through the network of regional eScience Centres, with which two workshops are arranged each year to showcase the DCC programme alongside the actual experiences of eScience practitioners. Consequently, we have established a broad portfolio of activities for encouraging a national focus within the eScience community upon the requirements for good data curation. We would of course be happy to receive suggestions as to how that portfolio might be improved.

For further information on the DCC's eScience liaison programme please contact:
Graham Pryor, tel. +44 (0)131 650 9985 or email graham.pryor@ed.ac.uk