Sustainability of Research Data Management Services

Peter Halfpenny[1], Rob Procter[1] and Alex Voss[2]

[1] Manchester e-Research Centre, University of Manchester
[2] School of Computer Science, University of St Andrews

Overview
It is now widely accepted that research data is a vital resource that needs to be systematically organised, securely stored, fully described, easily located, accessible on appropriate authority, shareable, archived and curated. This is both to comply with funders', publishers' and regulators' requirements and also to preserve the value of the material for future research by the data generators or by others. Fulfilling all the research data management tasks is a complex socio-technical challenge with, as yet, no widely-agreed, mature solutions. Moreover, in the face of the data deluge and a world recession, the scale of the tasks is accelerating while the financial and therefore human resources to undertake the tasks are decelerating.

This paper briefly reviews the drivers for research data management, the challenges it poses and explores possible pathways to sustainable solutions. It draws on two studies: the MaDAM project, funded under the infrastructure strand of the JISC Managing Research Data programme, with the objective of developing a pilot data management infrastructure for biomedical researchers at the University of Manchester (Collins et al., 2010; Poschen et al., 2010; Goff et al., 2010); and the Storage, Archiving and Curation (SAC) project led by Manchester Informatics which is charged with developing and implementing a research data management strategy across all four faculties of the University of Manchester. [Both projects are currently underway and their preliminary findings will be available to present at AHM2010 in September.]

Introduction
We live in an information age characterised by a deluge of digital data (Hey and Trefethen, 2003). The potential benefits are enormous, offering the opportunity to mount multidisciplinary investigations on a hitherto unrealisable scale into humankind's major social and scientific challenges, marshalling artificially produced and naturally occurring data of multiple kinds from multiple sources. But this newfound wealth of research data is valueless unless it can be systematically managed to make it discoverable and useable.

Current practice in data management and sharing
A recent survey of researchers in the social sciences (Procter and Voss, 2010) confirms that, at least in this discipline area, research data utilisation still lags far behind the vision of collaborative 'data-rich science'. The study revealed that 86% of respondents reported that research data they themselves collected was their essential or important source of data (N=1,062). Only just over half said that data shared by a colleague in their own institution (51%) or in another institution (53%) was an essential or important source of data. Less than a third identified data acquired from a national data service (30%) or an international data service (26%) as essential or important to their research.

Data management and sharing practices are often highly discipline-dependent, making generalisations difficult, however, these findings suggest that few researchers systematically manage the research data they produce and use, other than perhaps through some idiosyncratic file naming system and directory structure, personally constructed database, and occasional back-ups of our laptop onto USB sticks or optical disks. In the survey of social science researchers, 85% of respondents report storing their data on a personal hard-drive during the course of a project, 65% on a USB stick and 27% on CDs or DVDs. In contrast, only 33% use an institutional server. The large majority of respondent (80%) reported that they or a team member was responsible for storing and caring for the data after the project was completed, with only 21% using an institutional repository and 17% a national data archive, despite the ESRC's supporting the UK Data Archive for social science data.

External requirements
These local solutions to data management occur despite external requirements and university high-level policies insisting on a more systematic approach. In January 2007, the Research Information Network published a report on the 'Research funders' policies for the management of information outputs' (RIN, 2007), noting that there are significant differences in the extent to which funders see it as their responsibility to preserve and manage research data, with only two (NERC and ESRC) now

funding data centres, although some also support the Atlas Petabyte Data Store.[1] The difference between approaches by funders is compounded by the various agencies undertaking research (for example, JISC's Managing Research Data Programme), offering support (the Digital Curation Centre,[2] the UK Data Archive[3]), or proposing initiatives (UKRDS, 2008). Protecting personal information and sensitive data raises further issues that have, again, been addressed by a variety of agencies recommending a variety of practices: see for example, the ESRC's Secure Data Service and the UK Data Forum's National Data Strategy. Lyon (2007) seeks to provide some order to this plethora of recommendations in her summary of the various roles and responsibilities of researchers, institutions, data centres, re-users, funders, publishers and aggregators with respect to research data. Similarly, five principles are set out in RIN's 2008 report on 'Stewardship of digital research data – principles and guidelines'

The research data lifecycle

This cavalier approach to research data management is probably driven by immediate pressures of work that push housekeeping tasks off the agenda, and academic citation and reward systems that give little incentive to manage data well. However, essential to any successful management programme is its close alignment with the everyday practices of research scientists so that it fits seamlessly into the research data lifecycle. The most widely recognised account of this latter is that of the Digital Curation Centre.[4] To operationalise this generic scheme, it needs to be tailored to the particular activities and cultures of research groups.

It is this tailoring that is at the heart of the MaDAM project, which involves intensive requirements gathering within a small number of research groups, close study of developers responses to the requirements, and rapid feedback of users' experiences of prototype systems in an iterative process of co-realisation of a data management infrastructure that is embedded as part of the normal functioning of the research teams involved. Alongside this is an investigation of the viability of different financial models, a benefit realisation strategy, and a cost-benefit analysis of different financial models designed to sustain the system in the long run. The outputs of the MaDAM project will provide insights and working demonstrators to the SAC project as well as to the wider JISC community. In particular, MaDAM provides a potential model for instituting a university-wide data management service in a highly constrained funding environment. Instead of struggling to finance a 'big bang' approach that starts by assembling a developer team to produce a generic infrastructure and then impose it on researchers, a bottom-up, phased roll-out, with researchers buying into an infrastructure that evolves over time within broad guidelines, may prove to be more financially feasible.

The roll-out of a series of MaDAM-inspired demonstrators could also provide the incentive for investigators to sign up to systematic research data management, as research groups realise first the short-term gains of improved access to their own, systematically stored data, and then begin to gain longer-term benefits such as increased citation of their work by linking research data and publications in the university's institutional repository, eScholar: see Piwowar et al. (2007).

Financial models for sustainability

The 'Keeping research data safe: phase 2' study[5] has developed a cost model for preserving research data, and provided in depth case studies. It shows that there are considerable cost benefits to current researchers in the short term as well as long term benefits. MaDAM and SAC go further in investigating how sustainable ongoing financial support of the data management infrastructure by the University can be achieved. Funding models being investigated include:

Direct cost recovery

This model recovers costs directly from research awards under fEC guidelines. Investigators would include the cost of research data management infrastructure in proposals as estimated by the extent and type of anticipated storage capacity and any tailoring of the infrastructure needed to meet special requirements (such as particular curation needs and compliance standards). This model can deliver an excellent customer-driven service provided accurate and transparent accounting systems assure customers that they are receive what they have paid for.

---

[1] http://www.e-science.stfc.ac.uk/services/atlas-petabyte-storage/atlas.html
[2] http://www.dcc.ac.uk/
[3] http://www.data-archive.ac.uk/sharing
[4] http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf
[5] http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads

Free at the Point of Use
This model provides research data management infrastructure free at the point of use (FATPOU), paid for either by the indirect costs included in research proposals costed under fEC or through block or QR funding from the funding councils. The fEC route is probably easier to make accountable through transparent costing of the university's research infrastructure, though could be resisted by research councils already troubled by wide differences in universities' indirect costs. Similarly, the funding council route could open up debates about dual funding again. In both cases, the university's strategic goals and external compliance requirements will have to be balanced against competing demands on funds.

Beyond institutional data management services
Like other research-intensive UK HEIs, the University of Manchester is committed to developing a strategy that is capable of delivering a research infrastructure that is fit for purpose and capable of meeting the needs of its local research community as research practices become more data-intensive (Hey et al., 2009). However, it must also be recognised that the e-Research vision also anticipates research becoming more collaborative and research teams more widely distributed. To meet this challenge, UK HEIs will need quickly to develop strategies and infrastructure solutions that enable the federation of individual data repositories and the virtualisation of data services. This adds a further layer of sustainability issues that will be briefly explored in the full paper.

Bibliography
Collins, C., et al., 2010. Towards a generic research information and data management infrastructure. Submitted to AHM 2010.

Goff et al., 2010. Understanding the impact of disciplinary practices upon emerging modes of research collaboration: a case study of Biomedical researchers. Submitted to AHM 2010.

Hey, T., Tansley, S. & Tolle, K. 2009 The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research

Hey, T. & Trefethen, A. 2003. The data deluge: an e-science perspective. Grid computing: making the global infrastructure a reality, pp. 809–824. Wiley.

Lyon, E., 2007. Dealing with relationships with data: roles, rights and responsibilities, UKOLN, University of Bath. See http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report.final.pdf

Piwowar HA, Day RS, Fridsma DB, 2007, Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Poschen, M. et al., 2010. Requirements for the User-Driven Development of a Pilot Data Management Infrastructure for Biomedical Researchers. Submitted to AHM 2010.

Procter, R. and Voss, A. 2010. Digital Tools and their Use in Social Sciences. Submitted to AHM 2010.

RIN, 2007. Research funders' policies for the management of information outputs.

RIN, 2008. Stewardship of digital research data: a framework of principles and guidelines.

UKRDS, 2008. The UK research data service feasibility study: report and recommendations to HEFCE.