The implications of disciplinary practices for emerging modes of data sharing: a case study of Biomedical researchers

Mhorag Goff[1], Meik Poschen[1], Rob Procter[1], Peter Halfpenny[1], Lorraine Beard[2], Jonathan Besson[2], Simon Collins[3], June Finch[1], Tom Grahame[2], Mary McDerby[3]

[1] Manchester eResearch Centre (MeRC), University of Manchester

[2] The John Rylands University Library (JRUL), University of Manchester

[3] Research Computing Services (RCS), University of Manchester

Introduction

e-Research underpins a vision of a transformation of research practice that is predicated on increasing sharing and re-use of research resources, such as data (Jankowski, 2009; Procter et al., 2009). However, in most disciplines, existing data management practices, skills and infrastructure are simply inadequate to meet the challenges that meeting the e-Research vision raises. For example, without clear paths of recognition and reward, following an open access approach to data (Berlin Declaration, 2003) may be seen as a recipe for loss of intellectual capital and competitive advantage (Williams & Pryor, 2009).

The MaDAM project, which is funded under the infrastructure strand of the JISC Managing Research Data programme, has 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; Halfpenny et al., 2010). In this paper, we explore the reactions of biomedical researchers to the adoption of a common infrastructure and data management practices, with a particular focus on disciplinary practices and cultures, notably attitudes to more open modes of data sharing that projects such as MaDAM – through their capacity to encourage greater standardisation of data curation practices – will make possible.

Methodology and Overview

As part of the wider requirements gathering activity within MaDAM, we have been conducting detailed studies of biomedical researchers' work practices. Our approach draws on insights from participatory design (Greenbaum & Kyng, 1991) and ethnomethodologically-informed workplace studies (Heath & Luff, 2000), taking into account the situated, contexted nature of researchers' work practices. These insights have been acquired as a result of mixed interview and observation sessions with individual researchers allowing the project team to gain an understanding of researchers' practices live and in situ.

The MaDAM project involves two pilot user research groups in two domains. The Life Sciences group includes Electron and Standard Microscopy researchers who all work with large quantities of imaging data in diverse formats. Within their specific area they use different methodologies and instruments (e.g. Standard, Cryo-Electron and 3D Tomography Electron Microscopes). The research of the Medical Sciences Magnetic Resonance Imaging (MRI) Neuropsychiatry Unit involves primarily brain imaging data from a number of distributed MR scanners run by University, Wellcome Trust and NHS. This includes textual psycho-social data linked with MR scans.

Summary of Findings

Two main themes have emerged from our investigations: first, local personal and group reuse of data and second, concerns raised by the implications of the wider sharing of data as implied by moves towards open access policies.

We found that microscopy researchers were keen to retain all data for potential reuse and reanalysis, even data without immediate obvious re-use value. For example, data from failed experiments may be kept in order to benefit from a 'lessons learned' process and to potentially uncover patterns in the future that might lead to new research questions and projects:

> *"I often (well a few times a year) check old data sets either for comparison with new data, or to check whether they fit a pattern that is becoming clearer the more data we collect. Sadly I'm never sure which data sets I'll revisit, hence the hoarding. Even a bad data set may be useful if it shows an example of a feature which we start to focus on in future research."*
>
> *"It is entirely possible that many datasets would be over little or no value, perhaps the result of a failed experiment and would never be looked at. However, they should be kept, on a 'just in case' basis."*

It needs to be recognised, however, that in some fields of biomedical sciences this attitude to retention is overridden by the sheer impracticability of storing vast quantities of data. In certain fields it is easier and preferable to rerun an experiment, or reprocess raw data than to try to store it. One of the standard

microscopists deletes digital versions of film images and retains hard copies because electronic storage capacity is at a premium and at any time she can rescan the film to generate a digital image if required. Another researcher recounted an instance where a set of scans were not kept because the raw data was too large for any of the optical disks available. In short, researchers' attitudes to managing data are grounded in the practical realities of their field and, in rapidly developing fields, where new instruments and computational techniques allow for greater throughput and acceleration of the research process, data is potentially more disposable.

Discussions about making research data open access all initially elicited a unanimous "no, we wouldn't want to share our data" response. However, it was also evident that our researchers do deposit data in public repositories such as the Protein Data Bank[1]. Further probing revealed a more nuanced attitude towards open access. Our researchers do make their data open access; they just may not make *all* of their data open access. For example, one Cryo-Electron Microscopist deposits data in an open access subject repository; however, this would only be a subset of his data. Similarly, for public engagement, researchers cherry pick good examples of outputs, and these are shared, for example, on a research group website, without any of the associated data about methodology which would allow someone to replicate an experiment:

> *"The competition is very high, we don't let these data at least out of this university before it is published and then only selected [data]."*

This is the crux of the matter; as several researchers told us, their discipline is highly competitive and this entails a need to be the first to publish new research findings. It is important that any potential for individual researchers' data to be used to generate more publications and lay the foundations for new projects are exploited by the researcher and their team, not by someone from an external research group. Therefore, even with collaborators, such as where a neighbouring lab may provide biological materials, research data is only shared on a 'need to know' basis as results only and not methodology, with contribution acknowledged in authorship of papers:

> *"There are certain politics in science where sometimes you want to share some stuff and some which you'd want to hold back so that you can save it for another paper."*
>
> *"It's always a trade off but (..) in biology, it's different from the physical sciences where they put everything in open access, we can't afford to do that."*

Data sharing statements for those biomedical scientists whose research outputs have no natural home in a public repository can be as minimal as a statement that "data will be made available on request"; to-date, none of the researchers we interviewed had been asked for their data as a result of this. In effect, this overlaps with the commitment to provide data on request to back up results reported in publications, however, one researcher summed up the general position that if someone asked for some reagent which it had taken her a year to produce she would think carefully about sharing it without any collaborative agreement which would allow her to gain from the outputs of sharing it.

We find that sharing by researchers takes place in the context of sharing with colleagues they know and trust and whom they can rely on providing some reciprocal benefits to justify the risk and loss of control, i.e. social capital at work. Where sharing does take place, it is in the context of research group lab meetings for the purposes of monitoring and managing the workflow of the team, getting colleagues' views and contributions, or diagnostic opinion in the case of human research, and to plan future work.

Conclusions

As other investigations have discovered, the barriers to making data open access currently outweigh the drivers for many researchers, despite the policies of funding bodies, For example, our findings are These findings are commensurate with those of Williams et al. (2009), particularly in terms of concerns around sharing of data beyond networks of trusted colleagues and risks of not getting credit for findings, and the effort that has been invested. Hence, researchers are not willing to make their data open access without having some control over what is released, when and being able to place some constraints on access.

The MaDAM project will provide data sharing and access capabilities as an enabler to open access data when the cultural conditions are right. However, our investigations confirm that it is important to acknowledge that a number of institutional and cultural incentives need to be in place alongside the practical capability to make data open access before that capability is actively exercised by researchers.

[1] http://www.pdb.org/pdb/home/home.do

The challenge for the development of policies and governance for data management and curation is that specific selection, retention and sharing criteria cannot simply be mandated due to the differences in research practices and cultures across fields.

As Williams and Pryor (2009) have argued, such policies need to recognise the influences exerted by different research filed and cultures. Changes in researchers' attitudes to data sharing are unlikely to come about merely by being mandated by research funders. Instead, as our findings suggest, the way forward is more likely to involve building the infrastructure that will facilitate the informal data sharing already in evidence, while exploring what kinds of incentives at the group, institution and community level may eventually succeed in encouraging researchers to adopt more open practices.

Bibliography

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. Conference on Open Access to Knowledge in the Sciences and Humanities. Berlin, October. Retrieved June 7, 2010 from http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

Collins, S. et al. 2010. Towards a generic research data management infrastructure. Submitted to AHM 2010.

Greenbaum J. & Kyng M. (eds.) 1991. Design at work: Cooperative design of computer systems, Lawrence Erlbaum Associates, Hillsdale, NJ.

Heath C. and Luff P. 2000 Technology in action, Cambridge University Press.

Halfpenny, P., Procter, R. and Voss, V. 2010. Sustainability of Research Data Management Services. Submitted to AHM 2010.

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

Jankowski, N.W. (Ed.) 2009. E-Research. Transformation in Scholarly Practice. Routledge, NY, USA & Oxon, UK.

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

Procter, R.; Poschen, M.; Lin, Y.; Goble, C.; De Roure, D. 2009. Issues for the Sharing and Re-Use of Scientific Workflows. Electronic Proceedings of the 5th International Conference on e-Social Science, Cologne, Germany, 24-26 June 2009.

Williams, R. & Pryor, G. 2009. Patterns of information use and exchange: case studies of researchers in the life sciences. Research Information Network. Retrieved June 7, 2010 from http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf