User-Driven Development of a Pilot Data Management Infrastructure for Biomedical Researchers

Meik Poschen[1], Mhorag Goff[1], Rob Procter[1], Peter Halfpenny[1], Lorraine Beard[2], Jon 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

Meeting the challenges of curating digital research data is becoming ever more important in the face of the "remarkable growth of data-intensive research in all knowledge domains" (Blue Ribbon Task Force report, 2010, p.3). These challenges reflect the need to address the whole data lifecycle, from the creation of source data to the end point of publication, the complexities of dealing with a multitude of data types and formats, and the importance of ensuring that solutions (both technical and non-technical) are capable of being embedded in diverse disciplines, working practices and research processes. According to the UK JISC-funded Digital Curation Centre (DCC)[1], data curation means supporting data capture, management and dissemination along the data lifecycle by adding value (metadata, cross-references), enabling preservation (secure long-term storage and backup) and sharing via trusted archives and repositories, thereby reducing the effort researchers have to put into these processes.

UK research councils now recognise the need for better data curation procedures and have started to explicitly require detailed data management plans for research awards (see Jones, 2009, for an overview of funders' data policies); similarly, the NSF has recently announced that data curation procedures are a "scientific necessity" (Mervis, 2010). However, awareness of the importance of data curation remains low within the research community and there is a lack of robust technical infrastructures to support sustainable data curation by individual researchers, groups and institutions (Blue Ribbon Task Force report, 2010). Furthermore the different disciplinary research practices and cultures around managing and sharing of data have to be taken into account[2].

This paper presents the approach and requirements findings to date of the MaDAM[3] project which is funded under the infrastructure strand of the JISC Managing Research Data programme[4] from October 2009 to March 2011. The project has the following objectives:

- Develop a pilot data management infrastructure for Biomedical researchers at the University of Manchester along their data lifecycle supporting digital curation and data sharing.
- Engage and work closely with the pilot user groups to ensure the infrastructure is fit for purpose for the individual and domain specific research practices.
- Investigate how research data management services and infrastructure may be embedded within research practices of the University of Manchester. The pilot acts as a first step in analysing how a university-wide data management service can be introduced.
- Develop a data management plan and investigate activities to ensure the sustainability of service provision, including a cost-benefit analysis. The findings will be used as input to a wider strategic activity to address the needs of the whole of the University research community.

Methodology

User engagement in the MaDAM project focuses on an iterative user-driven development process together with collecting non-technical requirements and is grounded in the concept of co-realisation (Hartswood et al., 2007). The 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 and, in particular, individual and domain specific data lifecycles. The aim of this methodology is to bridge the gap in perception between developers and users of what makes a work-affording system that fits the actual needs of users in their research environment. It also fosters the in-depth gathering of other, non-technical requirements which are necessary to understand the researchers' work settings, their institutional context and relevant

[1] http://www.dcc.ac.uk/

[2] Goff, M. 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.

[3] MaDAM: Pilot data management infrastructure for biomedical researchers at University of Manchester

[4] http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

policies and procedures. Figure 1 shows the MaDAM methodology, depicting the main activities involved in eliciting technical and non-technical requirements within a cyclic, user-driven process.

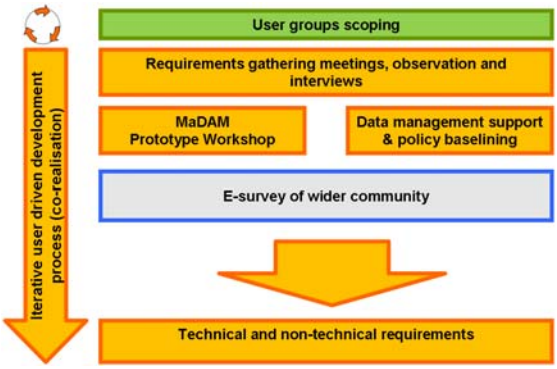

Figure 1: MaDAM method-flow

To find suitable users within the remit of the project, the team first met with a number of Biomedical research groups, gathering initial requirements to subsequently decide on whom to take on as pilot user groups (see next chapter). All information gathered in the recurring requirements capture and prototype evaluation activities with the pilot groups (see red-orange boxes in Figure 1) is constantly documented, circulated and re-evaluated within the project team to inform the development process and provide a rich picture of the users' needs and their research settings. A first e-survey of basic requirements of the wider Manchester research community has recently started to compare the specific requirements of the pilot domains towards a rollout in the university as a whole. A hands-on workshop was especially beneficial to evaluate the first prototype of a web-based data management infrastructure with users. The prototype[5] is based on the previous iteration of requirements and provides a navigation structure based on researchers' projects and experiments, centralized and backed up data storage, access rights, linkage and annotation of research data and a search function.

Pilot User Research Groups

The strong involvement of co-Investigators from Life and Medical Sciences from the MaDAM proposal phase onwards drove the focus on these domains and the existence of a pre-identified need in imaging sciences cutting across the disciplines provided the convergence on image data, a solution for which it is felt can be generalised to other data types. Furthermore, although the main research objects are images in various formats, resolutions, file sizes and diverse 'biographies' within specific research workflows, the data lifecycle also includes other data types such as text documents, diverse metadata, statistical data and outputs for dissemination. The MaDAM pilot user research groups are:

1) Life Sciences Electron and Standard Microscopy: The Life Sciences pilot group includes three sub-groups who all work with large quantities of imaging data in diverse formats. Within their specific research they use different methodologies and instruments (e.g. Standard, Cryo-Electron and 3D Tomography Electron Microscopes).

2) Medical Sciences Magnetic Resonance Imaging (MRI) Neuropsychiatry Unit: The research of this pilot group 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.

The work with the pilot user groups is further complemented by information and requirements gathered from additional researchers and PIs within the domain, IT and experimental officers as well as research and data policy managers.

Main Interim Requirements

In the paper we will present in detail the main findings around institutional context, researchers' working practices and workflows, technical requirements and data policies. The institutional context is such that responsibility for good data management is devolved to individual researchers. This entails lack of a consistent minimum standard or common set of agreed conventions; even though there may be some good practice where PIs of research groups set standards for their teams. As a result there are no back-up policies to guard against loss of data and no structured annotation of data. Storage solutions range from the strictly ad-hoc (e.g. using personal portable storage devices) to more organised shared network server storage space. This, however, is also employed as temporary storage and data needs to be purged regularly or backed up onto tape. Although many users refer to this as archiving, there is no annotation of data to aid retrieval and reuse, and none of the other selection, appraisal, sanitisation and curation activities associated with an active decision to preserve for the long term.

Transferring and sharing data is constrained by the need to use email or portable storage devices such as USB sticks or CDs with associated security, capacity and format issues. This results in limited

[5] Collins, S. et al. (2010): Towards a generic research data management infrastructure. Submitted to AHM 2010.

ability to share data within both local and distributed research groups. This is particularly problematic for PIs of research groups who lack an easy means of sharing, monitoring and reviewing of their team's work.

Using a folder directory paradigm entails high levels of redundant data due to duplicate copies of data being stored in different folders. For the above reasons search capabilities are limited and reliant on intelligent file and folder naming, remembering file names and for older data date ranges for the data in conjunction with contextual detail from lab books. Although researchers retain their data, often indefinitely, there are no archiving policies to guarantee long term curation in the absence of any designated institutional or local level archive. Some researchers are able to make use of public databases to deposit outputs, however, this constitutes a small proportion of the research data users would retain and many researchers say that there is no appropriate database for their type of research.

Challenges, Findings & Next Steps

Current approaches by researchers to long term preservation is underdeveloped because their basic needs for secure, trusted storage (and back-up) to support the research lifecycle are not yet being met. Existing institutional and faculty support for researchers, including IT Services, Research Offices and people managing the core facilities and scanners, directly and indirectly contribute to research data management. Engagement of these support structures will be essential to policy development and are critical to sustainability in terms of both buy in and the potential for capacity building in their services.

There is a need to develop the MaDAM solution to tie in with actual research practice for a range of researcher 'profiles' and workflows. This entails a need for balance between flexibility on functionality while reaping the benefits of consistent data management practice. In the Medical Sciences domain there is additional complexity due to the need to integrate and comply with prescriptive procedures for management of human data and deal with political difficulties around accessing data generated within the NHS. A research data management plan will be developed as a tool to link policy and support requirements. Next steps include working through a data management plan template with users to generate a data management plan for their own and their team's use, and to serve later as a basis both for a tailored, discipline specific and for a wider institutional data management policy.

The microscopy pilot user group have already adopted more standardised data management practices through involvement with MaDAM. We have recruited 'user champions' to drive cultural change within their disciplines and to perform an advocacy role for MaDAM within their communities, facilitating wider roll out of the pilot, and exploration of sustainability and scalability issues. Good progress has also been made in establishing the functional requirements for the prototype data management infrastructure. Technical support and sustainability is being addressed through Cost Benefit Analysis and financial modelling. A cultural change is needed for the proper support of domain specific data management plans, research practices and research management policies in general, and this, inevitably, will take time.

Acknowledgements

We would like to thank our pilot users for their time despite tight schedules and their invaluable and ongoing contributions regarding their work practice, requirements and evaluation activities. We wish also to thank the JISC for funding the MaDAM project.

References

Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010): Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final Report, February 2010. Retrieved May 24, 2010 from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

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

Hartswood, M., Procter, R., Rouncefield, M., Slack, R, Voss, A., Buscher, M. and Rouchy, P. (2007): Co-realisation: Towards a Principled Synthesis of Ethnomethodology and Participatory Design. In: M. Ackerman, T. Erickson, C. Halverson and Kellog, W. (eds.): Resources, Co-evolution and Artefacts, Springer.

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

Jones, S. (2009): A report on the range of policies required for and related to digital curation. DCC Policies Report, Version 1.2, Glasgow, March 2009. Retrieved May 24, 2010 from http://www.dcc.ac.uk/webfm_send/129

Mervis, Jeffrey (2010): NSF to Ask Every Grant Applicant for Data Management Plan. ScienceInsider, Breaking news and analysis from the world of science policy, 5/5/2010. Retrieved May 24, 2010 from http://news.sciencemag.org/scienceinsider/2010/05/nsf-to-ask-every-grant-applicant.html