The International Journal of Digital Curation
Volume 7, Issue 2 | 2012

Development of a Pilot Data Management Infrastructure for Biomedical Researchers at University of Manchester – Approach, Findings, Challenges and Outlook of the MaDAM Project

Meik Poschen, June Finch, Rob Procter, Mhorag Goff,
Manchester eResearch Centre (MeRC),
University of Manchester

Mary McDerby, Simon Collins,
Research Computing supported by IT Services for Research,
University of Manchester

Jon Besson, Lorraine Beard, Tom Grahame,
The John Rylands University Library,
University of Manchester

Abstract

Management and curation of digital data has been becoming ever more important in a higher education and research environment characterised by large and complex data, demand for more interdisciplinary and collaborative work, extended funder requirements and use of e-infrastructures to facilitate new research methods and paradigms. This paper presents the approach, technical infrastructure, findings, challenges and outlook (including future development within the successor project, MiSS) of the 'MaDAM: Pilot data management infrastructure for biomedical researchers at University of Manchester' project funded under the infrastructure strand of the JISC Managing Research Data (JISCMRD) programme. MaDAM developed a pilot research data management solution at the University of Manchester based on biomedical researchers' requirements, which includes technical and governance components with the flexibility to meet future needs across multiple research groups and disciplines.

International Journal of Digital Curation (2012), *7*(2), 110–122. http://dx.doi.org/10.2218/ijdc.v7i2.234

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. The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

A few years ago, a special issue of Nature on 'Big Data' (2008) explored the implications of the huge and ever growing volumes of data being created in science, emphasising the necessity for researchers to adapt their practices and institutional settings to deal with this 'deluge' and exploit the opportunities this presents. Jim Gray postulated a "Fourth Paradigm for scientific exploration" (Hey, Tansley & Tolle, 2009) based on data-intensive research methods and stressed the need "to do better at producing tools to support the whole research cycle – from data capture and data curation to data analysis and data visualization" (ibid), to the end point of preservation and dissemination. The insight grew that management and curation of digital research data has become more and more important in the face of the "remarkable growth of data-intensive research in all knowledge domains" (Blue Ribbon Task Force report, 2010). Doing this effectively means taking into account the multitude of data types and formats, and ensuring that technical and non-technical solutions fit within diverse working practices, research processes, cultures and disciplines. Funders are recognising the need for open data, better digital curation procedures and policies. UK research councils[1] now require data management plans for research awards (Jones, 2009) and the NSF[2] has declared data curation procedures to be a "scientific necessity" (Mervis, 2010). However, until recently, practical awareness of the importance of research data management and data curation within the research community has been comparatively low, resulting in a lack of robust e-infrastructures to support and sustain digital curation in the research environment (Blue Ribbon Task Force report, 2010).

One of the (ongoing) initiatives to raise awareness and improve data management and curation for research and higher education institutions in the UK was launched by the JISC in 2009 under its 'Managing Research Data programme'[3]. In this paper we look closely at one of the projects in the infrastructure strand of this programme, the MaDAM project at the University of Manchester[4] (Finch, 2011; Poschen et al., 2010).

The aim of MaDAM was to implement a pilot data management solution for biomedical research data, with a focus on diverse image data, supporting the whole data life cycle and everyday research processes, taking into account the existing institutional landscape. The solution comprises of technical (hardware and software) elements and also policy, process and governance frameworks for research data management. The user groups selected for the pilot were researchers in the Life and Medical Sciences at the University of Manchester (UoM). MaDAM also provided input to a wider strategic activity at UoM to address the needs of the whole of the research community. In the context of the project, 'curation' is understood in broad terms as activity to manage and organize a collection of data along its lifecycle, which would cover a range of researcher activities designed to preserve and add value to data. The DCC defines Digital Curation as "maintaining, preserving and adding value

[1]Research Councils UK (RCUK): http://www.rcuk.ac.uk/

[2]US National Science Foundation (NSF): http://www.nsf.gov/

[3] JISCMRD programme: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

[4]MaDAM overview: http://www.library.manchester.ac.uk/aboutus/projects/madam/ (MaDAM website), http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmi/madam.aspx (JISC MaDAM page)

to digital research data throughout its lifecycle”[5] as an active management process to ensure that the data’s research value is maintained and that it remains accessible in the long term. This paper presents the MaDAM project’s approach, technical infrastructure, findings, challenges and outlook, including a brief overview and future development of the recently started successor project, MiSS.[6]

Approach

Project Objectives

The practical rationale behind MaDAM, against the background of an absence of an institutional repository or a comprehensive strategy for the dedicated management of research data at the time, was threefold:

1. Researchers need to be supported to manage their data effectively and comply with legal and funder policies.
2. Funders want to ensure public money spent on research is maximised, which means ensuring research data is open when possible and preserved for reuse and potential validation.
3. The potential future value in data assets needs to be preserved.

Hence, a technical and governance solution was to be produced based on researchers’ requirements and their working practices with the flexibility to meet future needs across multiple research groups and disciplines, taking into account the institutional landscape, funders’ requirements and existing internal and external policies. Furthermore, routes into sustainability were to be explored, with the findings informing the development of a wider research UoM data management strategy.

Methodology

Grounded in the concept of co-realisation (Hartswood et al., 2008) and backed by experience from previous activities (see e.g. Poschen et al., 2008) the MaDAM project employed an iterative user-driven development process, together with collecting non-technical requirements (Poschen et al., 2010). Figure 1 shows the elements of the MaDAM methodology, aimed at identifying technical and non-technical requirements:

1. Assessment of current working practices and data lifecycles, policies and procedures of the pilot groups (including support staff, e.g. research administrators, central and faculty IT services);
2. Assessment of the research (e-)infrastructure (i.e. technical instruments);
3. Development of a technical data management infrastructure plus supporting measures, such as data management plans.

The project began with a user groups scoping process involving a number of meetings and interviews to identify researchers in the two chosen domains, Life and

[5]Digital Curation Centre (DCC): http://www.dcc.ac.uk/digital-curation/what-digital-curation/

[6]MiSS (MaDAM into Sustainable Service) project website: http://www.manchester.ac.uk/miss/

Medical Sciences, who would benefit from taking part and be able to dedicate the necessary time and effort.

All information gathered in the recurring requirements capture and prototype evaluation activities (see Figure 1) was iteratively documented and evaluated within the whole project team to provide a rich picture of users' needs and their research settings (see Goff, 2010a; Poschen, 2011b). Interviews and meetings were also conducted with other stakeholders in the research environment as mentioned above. This led to a better understanding of the intricacies of the institutional landscape and settings, relevant internal and external policies and related legal and ethical issues (see Goff, 2010b).

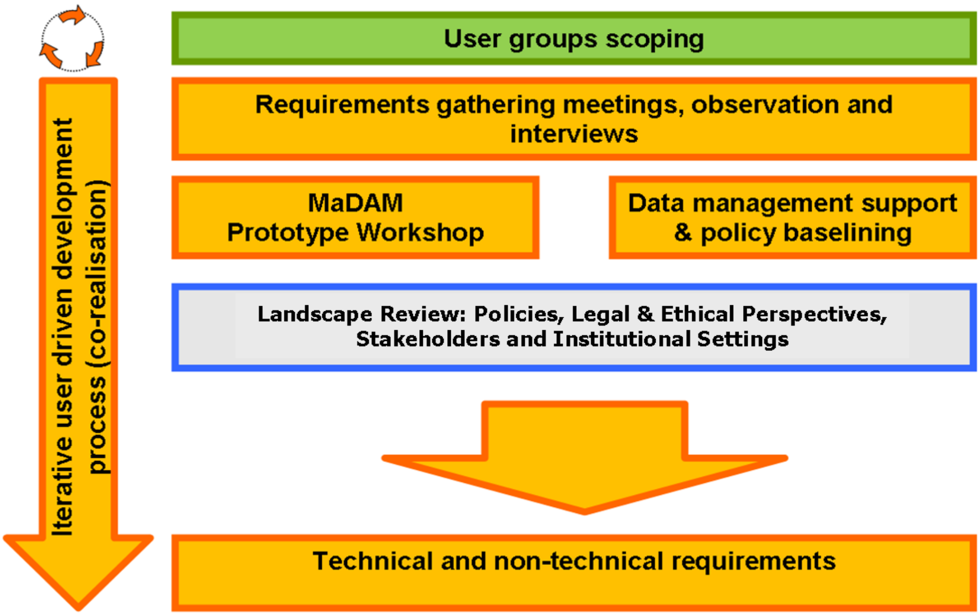

Figure 1. MaDAM Method Flow.

The following activities were employed to elicit requirements, keep user groups engaged and gather feedback throughout the development process.

- Four rounds of face-to-face and group interviews to refine understanding of users' needs, the data lifecycle and work practices and, from mid-project onwards, observation and usability testing to evaluate the system.
- Interviews about data management plans, funder and policy requirements.
- More informal email exchanges and chats with users to further verify information and foster understanding.
- Three 'Prototype Workshops' to discuss major iterations of the MaDAM infrastructure. After the second workshop in September 2010, users were encouraged to use the system with real research data. Continuous feedback from this point onwards meant that requirements and evaluation went hand-in-hand within the iterative development process.

- A feedback feature within the MaDAM system enabled users to electronically log bugs, questions, request additional features, etc., while testing and using it.
- The bug, issue and project tracking software JIRA was used to collect and document all technical feedback, with the project team deciding on importance and feasibility for further releases.
- Diverse meetings and interviews with other project stakeholders to gather information for the landscape review and for insights into data management policies and plans.

User Community

The pilot user groups consisted of researchers in the Biomedical Sciences, specifically standard and electron microscopy users in Life Sciences and Magnetic Resonance Imaging (MRI) users in Medical Sciences. In the proposal stage of the project these domains had been flagged as having a clear need for research data management support, particularly because of the large and diverse imaging data sets produced. The Life Sciences group included eight active core users, plus some occasional users. The Medical Sciences group consisted of five users.

Life Sciences Microscopy Users

The Life Sciences researchers make use of a number of standard and/or electron microscopes to generate raw data. Although there are data and metadata generated in the preparation of samples (and in the case of the Medical Sciences users, the preparation of clinical trials and subject metadata; see next section) the user community had few issues in managing this complementary data, being low volume and low complexity. The point where data management and research workflow intersect, from the MaDAM project point of view, starts at the instrument with the researcher physically fetching the raw data from the microscope or scanner (a single microscope run can create anything from 1GB to 200GB). This has to be done because the instruments are 'firewalled' to insulate them from external networks for security reasons, and the researcher must transfer the data using a portable device (e.g. USB or optical media) to their own PC.

At this stage, the whole data set becomes the researcher's responsibility and, in the absence of practical guidance for data management, every researcher has their own processes for back-ups, file management, annotation, metadata capture, storage locations and media for the short, medium and long term. Raw data are manipulated and analysed through a series of steps, using a variety of computational, statistical and other analysis techniques and software to produce various interim versions of processed and analysed data, up to the point of creating outputs for publication and other forms of dissemination, such as website material for public engagement.

Medical Sciences Magnetic Resonance Imaging (MRI) Users

Researchers working in the Neuropsychiatry Unit (NPU) within the Imaging Sciences research group in the School of Medicine make up the second user group. Their research involves using multiple MRI scanners across the campus, within local hospitals and research facilities, some of which sit within the NHS network.

NPU researchers carry out scans on human subjects to discover the mechanisms which underlie common mental illnesses such as schizophrenia and depression to establish links between psychiatric symptoms, psycho-social conditions and brain activity to better understand brain function. The studies are usually not for clinical purposes; however, they are sometimes associated with clinical trials of drugs. Most studies involve taking functional scans of subjects' brains to look at blood flow in areas of the brain over a period of time, in response to a particular drug, stimulus or set of conditions, or for comparison of brain function in healthy versus non-healthy volunteers with a specific psychiatric condition.

As scans are 4D or 'functional' (i.e. a series of images taken over time as the fourth dimension) rather than "structural" (a single scan image), the volume of raw data generated by the MRI scanner for each subject is fairly large (usually one study consists of 20-40GB of data). The images in a study are further complemented by textual psycho-social data in the form of interviews, questionnaires or assessment reports. This data is kept in the usual office-type formats and is quite lightweight in size.

The data workflow broadly mirrors that for the Life Sciences users, with the key distinction being that the Medical Sciences users have to manage confidentiality and other ethical issues related to human data. Scan images are anonymised at source during the process of transfer from the MRI scanner to a local PC, from which the researcher will retrieve the images. However, the researcher maintains the ability to trace back to the subject's personal details if necessary by referring to a hard copy 'key' kept locked on site with the scanner.

Key Characteristics of User Groups in Both Domains

Both pilot user domains use advanced instrumentation to generate data, entailing:

- Data output in different formats. Depending on the equipment manufacturer, some entail use of proprietary formats and therefore there is a need to convert formats and possibly capture format specific metadata;
- Research constrained by procedures around use of the instrument;
- Users transferring data from the local instrument network to a PC via portable storage device to protect instrument settings/software integrity;
- In general, fairly large volumes of data generated in each session;
- Greater automation in the usage of instruments. This opens the option for microscopists to run and monitor sets of samples on the microscope remotely (as compared with MRI users who need human subjects to be present);
- Developments in processing and analysis techniques, which imply increased data throughput through the research process and therefore greater volumes of data to manage.

 doi:10.2218/ijdc.v7i2.234

Technical Architecture and Software Functionality

Based on the requirements and evaluation activities, a web-based technical solution has been developed (for a more detailed technical overview see Collins et al., 2010). Overall, the system should be robust and easy to use, provide secure access, support annotating and structuring of data, and deal with high throughput of large data sets and a large number of files.

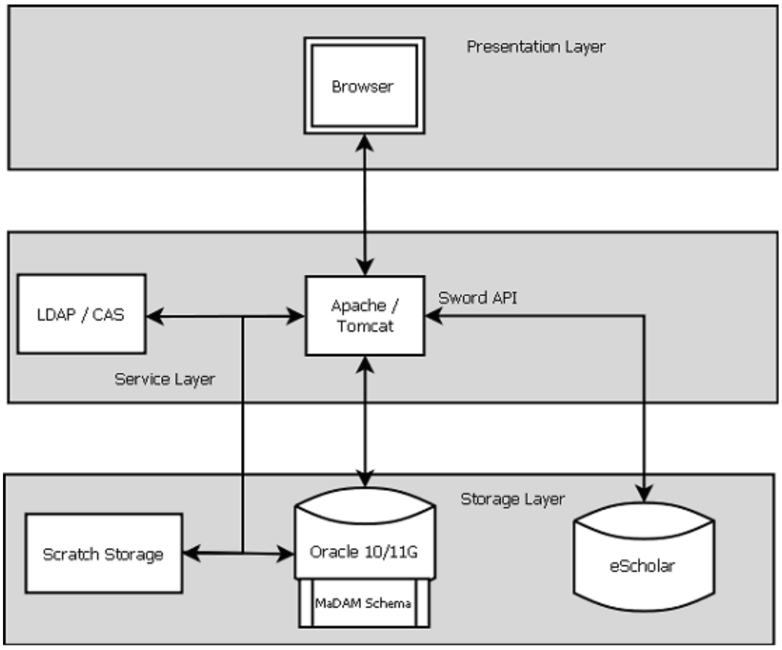

Figure 2. Architecture for the MaDAM data infrastructure.

The pilot data management infrastructure developed consists of two tiers as shown in Figure 2. A relational database management system (RDBMS) handles the actual management of the research data, while the dissemination and long term archiving solution utilises the UoM's eScholar[7] which is based on the Fedora Commons digital repository (through the SWORD API4 a subset of the data in the database will be archived in eScholar). In functional terms, the infrastructure is a cross between a Content Management System (CMS) and a lightweight Laboratory Information Management System (LIMS).

Figure 3 shows the final version of the web-based user interface following user testing and feedback. Hierarchical entities can be created and defined, then annotated and enhanced by uploading data. "Experiment", "Sample" and "Publication" are common example entities. The folder structure is based on the notion of projects and users can manage access rights to single files, folders or defined groups of people. Pre-defined and configurable lists are provided for tagging information and metadata next to entities, adding context to data and supporting search functionality. Entities can be (multi-)linked to other entities (e.g. an experiment may then belong to many projects) or objects outside the system. Users are able to browse and preview data, an important feature to (quickly) identify objects. An archiving button creates a record in

---

[7]UoM's publication and dissemination repository: http://www.escholar.manchester.ac.uk/

eScholar (for dissemination) and tags the data as ‘archived’ and read-only. The user may choose if the data will be publicly available, embargoed for a time or stay private.

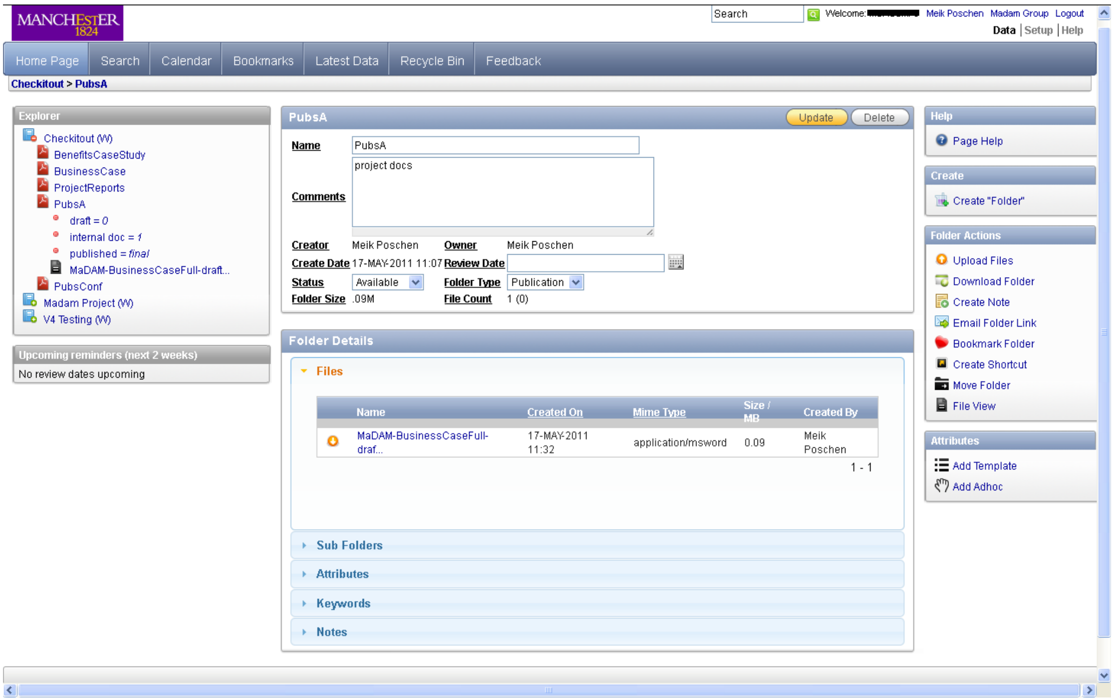

Figure 3. Web-based MaDAM System, final version.

When creating a project, a Data Management Plan (DMP) has to be populated to make sure internal and funders’ requirements (e.g. on data retention policies) are adhered to, making use of the eDMP tool (an integrated data management plan tool for this purpose, also providing guidelines). The DMP can then be maintained by the user throughout the project. MaDAM has further explored ways to integrate administrative data directly into its system via a link with the Manchester Research Information Management (RIM) database to make information easily available via ‘drop-down’ menus and ‘autofill’ features, easing the process of DMP creation for the researcher. As a proof of concept, this link worked in a prototype version but could not be fully implemented and evaluated within the project’s lifetime.

Findings and User Benefits

Secure Storage Space, Back-up and Day-to-Day Data Management

For the MaDAM pilot user groups, the main benefits lay in the support of their day-to-day storage and data managing needs. MaDAM provides a central storage location with automated back-up, freeing up researchers’ PCs and local storage for ‘work in progress’. Trust in the security of the data is critical for the user community to have the confidence to delete data from their PCs. An immediate benefit was the reduced risk of data loss from not backing up, as well as making data more discoverable as its storage location will now be known.

The pilot service enables easier sharing of files within the research group or project team, using a function which sends an email link and/or by allowing group-based access to the data. Membership and access is controllable by an administrator of choice (often the PI) and read-write permissions by the data creator. This opens up opportunities for better coordination, more timely feedback and collaboration, as well as eliminating the need to spend time preparing data for sharing internally. A function which allows data files to be linked with multiple projects eliminates the need to store multiple copies of data in different locations.

Capabilities for capturing metadata at all levels, from project down to data files, and in a number of forms provides flexibility for researchers to use a range of formats from self-defined and fairly constant experimental variables, e.g., instrument names which can be selected from a drop down menu, to free text annotation. Metadata templates can be set up containing key attributes which need to be captured for all projects, experiments and data files.

Making metadata fields flexible and user-defined allows researchers to record the metadata most relevant for their specific research area. A free text 'notes' field enables more idiosyncratic and context-specific annotation for various purposes, including being able to flag data for review, for important 'notes to self', or adding information about the project rationale, which may be useful as an aide memoir or for other researchers. Achieving an appropriate balance between flexibility and imposed structure will be important for wider uptake and scalability.

Institutional Settings, Policies and Data Management Plans

The absence of coherent data management policies when the project began made it difficult for researchers to make use of existing sources of information on policies or procedures. As pointed out in the 'MaDAM Benefits Case Study':

> "This lack of a supporting framework for policy at University level is [..] due in part to the differences in needs, culture and politics of the different faculties and disciplines which operate almost as distinct entities and which may be difficult to reconcile. For the MaDAM pilot research groups, this means their work practices regarding data management procedures or plans are quite diverse: mostly it is down to the single researcher, sometimes to the PI, to set at least a minimum of standards." (Poschen, 2011a).

With requirements on the DMP level gathered, the integrated eDMP tool could only be implemented as a proof-of-concept within the project's lifetime. It was seen as very useful, but longer term usage is needed to further evaluate and refine it.

Dissemination, Re-Use and Open Data

For most of the project users' views on long term preservation, linking their datasets directly to publications, sharing them outside of their project team and re-using data other than their own were underdeveloped – in spite of the topic being raised on several occasions by the project team. In one of the last evaluation sessions, one of the user champions asked for enhanced functionality to be able to link a file as a 'best example' to a publication representing a specific experiment or dataset.

MaDAM's link into eScholar as the preservation and dissemination endpoint was therefore not a particular demand during the project's lifetime and, despite being implemented, would need further use before it can be evaluated.

Summary of User Benefits

User benefits resulting from requirements captured, evaluated, implemented and honed over the project's lifetime can be summarized as providing an accessible, easy-to-use platform for the organisation and annotation of day-to-day research data. Making trusted, secure storage available to reduce risks of data loss and establishing mechanisms to enhance data, add metadata and make research objects visible and searchable, thereby making existing resources easier to discover and reducing redundancy of data, proved to be crucial user demands. Other important needs include facilitating better data sharing (securely and owner-controlled), maintaining media format accessibility for long-term reuse and enabling compliance with legal and funder obligations (via DMP).

Other benefits pertain to raising awareness of diverse data management issues at the individual researcher level, the local institution and the wider research community. UoM has further benefited from getting insights into researchers' requirements and integration of the pilot with existing but separate systems as the University moves towards a future service for managing research data.

Challenges, Outlook and Future Development

The main challenges for the project lay in:

- Making the best use of users' limited time for recurring interviews, usability sessions and workshops;
- Managing the expectations of UoM and external users who voiced interest in the project, as we were naturally limited in what we could achieve over the project's lifetime. Enthusiastic users had to be actively engaged with and MaDAM prompted interest from outside parties with research data management needs;
- Ensuring that solutions would fall inline with current working practices, including the caveat that cultural change takes time and different domains have different needs towards digital curation and open data;
- Dealing with a diverse and fragmented landscape with policies from external (funding) bodies and relevant internal policies around ethics, information security and data protection as separate themes;
- Engaging with existing institutional and faculty support for researchers, including IT Services, research offices and people managing the core facilities and scanners, who directly and indirectly contribute to research data management; engagement of these support structures is essential to

policy development and is critical to sustainability in terms of both buy in and the potential for capacity building in their services;

- Recognising a significant future challenge in managing storage capacity by balancing the facility to store research data centrally with initiation of a review process in which researchers will evaluate and re-evaluate their data at appropriate intervals to instigate disposal of unwanted data, whilst retaining data of continuing usefulness.

MaDAM has been successful in addressing the needs of its user groups and in developing a pilot infrastructure, which is live, maintained and actively utilised. MaDAM's outputs and findings, together with being part of an initiative for a sustainable University-wide Research Data Management Service (RDMS), helped secure funding for the recently started successor project MiSS (MaDAM into Sustainable Service).

MiSS will be building on the experience, outputs and infrastructure of MaDAM, although it is more a transitional project than a continuation, which will move the pilot into a sustainable service within the University's new technical framework. This means the technical infrastructure will be rebuilt, making the system more generic for a University-wide service while still being open and tailorable with an associated plug-in framework to allow specific requirements across all domains. Besides the technical service, MiSS will include a Research Data Management Policy, along with a supporting Service, and integrate with the necessary human infrastructure to address the Research Data Management needs across UoM. Five champion user groups (in Engineering and Physical Science, Life Sciences, Medical and Human Sciences, Social Sciences/Humanities) from all four faculties are involved to drive the requirements and will work closely with the project team.

MiSS also opens up the opportunity to follow up on MaDAM activities concerning the DMP and eScholar integration evaluation:

> "Further evaluation would be needed to explore the linkage of DMP to University databases/systems further and how DMP are actually used in everyday research. In addition, integration with eScholar for the archiving and publication angle should be further evaluated in concrete use. Finally, a further and ongoing look into proper use, collaboration and uptake would beneficial to evaluate MaDAM as it is now mature enough to be used day-to-day as it was envisioned." (Poschen, 2011b).

Finally, we also plan to revisit MaDAM's pilot user groups post-project within MiSS to track issues arising from routine use (e.g., how requirements evolve), including data management planning, collaboration, sharing and publication.

Acknowledgements

We would like to thank our pilot users for their time and their invaluable ongoing contributions to the project. We wish also to thank the JISC for funding the MaDAM project and the successor, MiSS (MaDAM into Sustainable Service).

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. Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Collins, S., Beard, L., Besson, J., Finch, J., Goff, M., Halfpenny, P., Grahame, T., McDerby, M., Poschen, M., & Procter, R. (2010). Towards a generic research data management infrastructure. Paper presented at the UK e-Science All Hands Meeting 2010, Cardiff, UK. Retrieved from http://www.allhands.org.uk/sites/default/files/2010/TuesT4CollinsTowardsaGeneric.pdf

Finch, J. (2011). *MaDAM final report.* Manchester, UK. Retrieved from http://www.merc.ac.uk/sites/default/files/MaDAMFinalReport.pdf

Goff, M. (2010a): *MaDAM requirements and gap analysis*. MaDAM Project Report, Manchester, UK. Retrieved from http://www.merc.ac.uk/sites/default/files/MaDAM_Requirements%20_%20gap%20analysis-v1.4-FINAL.pdf

Goff, M. (2010b): *MaDAM landscape review: Policies, legal & ethical perspectives, stakeholders and institutional settings.* MaDAM Project Report, Manchester, UK. Retrieved from http://www.merc.ac.uk/sites/default/files/MaDAM_Landscape_Review-v1.1-FINAL.pdf

Hartswood, M., Procter, R., Slack, R., Voss, A., Büscher, M., Rouncefield, M., & Rouchy, P. (2008). Co-realisation: Towards a principled synthesis of ethnomethodology and participatory design. In M.S. Ackerman, C.A. Halverson, T. Erickson & W.A. Kellog (Eds.), *Resources, Co-evolution and Artefacts: Theory in CSCW*, Springer.

Hey, T., Tansley, S., & Tolle, K. (2009). Jim Gray on eScience: A transformed scientific method. In T. Hey, S. Tansley & K. Tolle (Eds.), *The Fourth Paradigm: Data Intensive Scientific Discovery.* Microsoft Research, Redmond, WA. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

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

Mervis, J. (2010). *NSF to ask every grant applicant for data management plan*. ScienceInsider. Retrieved from http://news.sciencemag.org/scienceinsider/2010/05/nsf-to-ask-every-grant-applicant.html

Nature. (2008). Community cleverness required. *Nature, Special Issue on Big Data Editorial, 455*(7209). Retrieved from http://www.nature.com/nature/journal/v455/n7209/full/455001a.html

Poschen, M., Daw, M., Procter, R., Turner, M., Hanley, T., Slack, R., Hall, A., Jones, M., Jones, R., Le Blanc, A., Place, E., Rogers, N., Rowley, A., Schiebeck, T., Steer, D., & Williams, C. (2008). User-centered development of a virtual research environment to support collaborative research events. Paper presented at the 4th International Conference on e-Social Science, Manchester, UK.

Poschen, M. (2011a). *MaDAM JISC RDMI benefits case study.* MaDAM Project Document, Final Draft. Retrieved from http://www.merc.ac.uk/sites/default/files/MaDAM-BenefitsCaseStudy-FINAL_draft.doc

Poschen, M. (2011b). *MaDAM evaluation report*. Final Version. Retrieved from http://www.merc.ac.uk/sites/default/files/MaDAM-EvaluationReport.pdf

Poschen, M., Goff, M., Procter, R., Halfpenny, P., Beard, L., Besson, J., Collins, S., Finch, J., Grahame, T., & McDerby, M. (2010). User-driven development of a pilot data management infrastructure for biomedical researchers. Paper presented at the UK e-Science All Hands Meeting 2010, Cardiff, UK. Retrieved from http://www.allhands.org.uk/sites/default/files/2010/WedT4PoschenUserDriven.pdf