EVALUATION REPORT

MADAM – PILOT DATA MANAGEMENT INFRASTRUCTURE FOR BIOMEDICAL RESEARCHERS AT UNIVERSITY OF MANCHESTER

JISC RESEARCH DATA MANAGEMENT INFRASTRUCTURE PROGRAMME

MEIK POSCHEN, MERC – 30 JUNE 2011, FINAL VERSION

1. INTRODUCTION, APPROACH & USER ACTIVITIES OVERVIEW

1.1. INTRODUCTION

This report concludes the requirements gathering and evaluation activities of the MaDAM project at University of Manchester, which aims at developing a pilot infrastructure for the better management of data along the research lifecycle, from data capture to storing, preservation and dissemination, with a *formal end of project evaluation*. In this context it also gives an *overview of user activities* conducted in the second half of the project in this first chapter. For a comprehensive overview of the project, its deliverables and findings see the 'MaDAM Final Report' [1].

1.2. APPROACH

The approach taken for this formal evaluation is based on methodology and linked to findings described in detail in the 'MaDAM Requirements and Gap Analysis' [2], which includes a record of all related activities in the project until and including August 2010. During the project's lifetime developers and user engagement experts worked hand-in-hand within the project team and together with the pilot users in an iterative user-driven development and evaluation process, which included collecting technical and non-technical requirements. These qualitative requirements gathering and evaluation exercises included interviews, workshops and observations of the users' research practice and system use.

For the final formal evaluation two one-on-one interview and system walkthrough observation sessions have been conducted face-to-face with the two user champions representative for their research groups in a) Standard Microscopy (25 May 2011) and b) Neuropsychiatry (19 May 2011) at their respective workplace (see [1] and [3] for an account on pilot user groups, communities and disciplines). Both users provided significant feedback to the project team and have been very dedicated in testing MaDAM in their everyday research, and have therefore been especially driving the development of the system over the project's lifetime. The sessions lasted about one hour each and findings based on audio recordings and notes made by the author of this report are presented in chapter 2 together with the session outline. The evaluation sessions could not include the full Data Management Plan (DMP) form (in MaDAM also accessible under/called 'Project Details'), as it was completed to only 80 percent at the time (see under 1.3. IV & V), Furthermore the 'Archiving' and 'Publishing' functionality provided through the linkage with eScholar had to be completely omitted for evaluation, but was implemented end of June.

1.3. Fᴜʀᴛʜᴇʀ Usᴇʀ Aᴄᴛɪᴠɪᴛɪᴇs: Rᴇϙᴜɪʀᴇᴍᴇɴᴛs & Eᴠᴀʟᴜᴀᴛɪᴏɴ Fᴇᴇᴅʙᴀᴄᴋ Eʟɪᴄɪᴛᴀᴛɪᴏɴ

This section summarises further user activities and ways of eliciting feedback from users between September 2010 and June 2011.

I) Pilot users used the 'Feedback' feature within the MaDAM system itself to directly electronically log questions, bugs, and wishes for improvements and additional features while actually testing and using MaDAM since the second major iteration of the system.

II) This second release was introduced to our users in the 'Second Prototype Workshop' on 30 September 2011 where immediate comments and issues fed into the development process and users have been encouraged to use the prototype 'in anger' with their own data following the workshop and to feed back any questions or issues directly to the team. Within the cyclic development process this methodically already meant an overlapping of requirements gathering and evaluation feedback from this point onwards.

III) Further feedback in form of informal emails, chats and phone calls between users and the project team helped to verify given information and foster exchange and understanding (e.g. for desired features not within the remit of the project) towards making the envisioned user-driven process work.

IV) An additional workshop was held on 24 February 2011 to elicit requirements on the kind of metadata users would like to add when uploading data to MaDAM in the end refining the existing functionality of the metadata template and attributes (see chapter 2 for evaluation details on the same topic). It further touched on DMP questions on funding body compliances, data management plans, IP & copyright and preservation.

V) After enhancing the DMP part which included ingesting data from the University content management system on grants and grant holders (i.e. administrative research data information) the afore (in IV) mentioned workshop was complemented by two face-to-face sessions in June 2011. Again the user champions representing the respective groups in a) Standard Microscopy and b) Neuropsychiatry were chosen as the most useful candidates, this time by the MaDAM team's metadata expert (see [4] for documentation on DMP templates).

VI) The bug, issue and project tracking software JIRA has been used to collect, address and document all technical feedback and the project team decided on importance and feasibility of those entries and requests for further releases of the MaDAM system.

VII) During the second half of the project we continued to further complement the work with the pilot user groups with information and feedback gathered from additional researchers and PIs within the domain, IT and experimental officers as well as research and data policy managers (see [1] for details on external group's/researcher's interest).

VIII) Non-technical feedback and insights into the users research practice collected during the project's lifetime especially helped to compile a 'MaDAM Benefits Case Study' [3].

2.   Fᴏʀᴍᴀʟ Eᴠᴀʟᴜᴀᴛɪᴏɴ Sᴇꜱꜱɪᴏɴꜱ Mᴀʏ 2011

2.1.   Uꜱᴇʀ Dᴏᴍᴀɪɴꜱ

As touched under 1.2. the two user champions represent our two main pilot user research groups from two different biomedical domains at the University of Manchester:

a) A Life Sciences Standard Microscopy group working with large quantities of imaging data in diverse formats and resolutions.

b) The research of the Medical Sciences Magnetic Resonance Imaging (MRI)

Neuropsychiatry Unit involves primarily brain imaging data from a number of distributed MRI scanners, plus textual psycho-social data linked with those MRI scans.

In the following the users will be referred to as a) and b) where necessary according to their group belonging.

2.2. Sᴇssɪᴏɴ Oᴜᴛʟɪɴᴇ

The session outline drawn up by the author for the one hour MaDAM interview and system walkthrough/research practice observation at the user's workplace included the following topics and was used as a checklist:

- Main benefits of MaDAM

- System usability and system functionality

- System use and work/research practice (collaboration/sharing)

- Future storage requirements (user/group) and envisioned use

- MaDAM in the context of dissemination/publication

- Nice to have? Questions?

2.3. Gᴇɴᴇʀᴀʟ Wᴏʀᴋ Pʀᴀᴄᴛɪᴄᴇ Usɪɴɢ MᴀDAM (sᴏ ꜰᴀʀ)

At the time of this evaluation both researchers (and therefore their respective groups) use MaDAM in a similar way in transferring the whole and identical file structure of a project and/or experiment into the system via drag and drop from their desktop machine without any problems (for a) Mac OS X with Firefox, for b) Windows XP with Google Chrome) and at reasonable speed over fast University networks (the MaDAM system in itself is not perceived as to slowing these transfers down, but they naturally take some time as the data is in the GB range). They use real research data but due to the fact that MaDAM still has been a system under development they I) upload only part of their data and II) do not use the MaDAM system for actually analysing/processing their data at this point. This means for a) that the Standard Microscopists for now have only uploaded and backed up real data from finished experiments, the analysis (which includes duplicating files as source files have to be kept to be able to go back to them) is still done on the individual machines. The b) Neuropsychiatrists all the same do not use MaDAM for analysing their data, but they use it

4

for backing up pre-processed raw data (weekly backups are needed) while analysis is still done on the researchers' workstation. Metadata is added in both cases after the upload on project and folder basis (see 2.4. for further details on work practice with metadata templates/attributes).

2.4. METADATA

DMP/'Project Detail' metadata

Both users had no problems in creating a project and entering the 'Project Details' into the DMP form, for b) especially as there is a useful 'Help' attached to each item and data field.

General Metadata

In both domains the replication of the original file structure on MaDAM is essential to not loose track on a high level managing their data. Furthermore for a) the name of the data sets is still defined by and includes its date as the main high level identifier; b) always uses the same typical folders/folder names in this pre-processing phase while the root directory information is most important to know where the data actually came from.

Metadata via Templates and Attributes (work practice and usability)

As MaDAM is not used for processing/analysing data at this point the metadata is added via templates (which again include a list of attributes) only on a top level folder basis, cascading down into the sub-folders if needed; usually having the metadata visible at the top level is good enough. The a) group use pre-defined templates with only eight attributes for acquisition/provenance and analysis of the data; all have pre-set values and no open fields are used as they "wanted to have something a little restrictive" for the research group, as otherwise searching for terms and keywords will not yield proper results, especially if using/collaborating on data from different owners. The group of b) usually use one list which applies to the whole structure/data below and which is attached only to the high level project folder; the template has 12 attributes, enough to give useful information, but not too much to make it too complex. Overall templates and attributes are seen as crucial to identify and find data.

The usability of creating and applying templates and attributes is straightforward for a) at this point (over the project's lifetime this was one of the most discussed and continually improved

features); for b) it needs getting used to by working with it on a regular basis, "but then it's fine".

2.5. COMMENTS ON FUNCTIONALITY/FEATURES

- The 'Recycling Bin' is seen as a "neat feature" to make sure no data gets lost in deleting older versions especially when working with processed data live on the system in the future.

- The 'Latest Data' feature is perceived as very useful by b) to see what happens in the group, whereas a) at the moment cannot see a huge benefit of this feature, but might still in the future when collaboration really starts on the system.

- The 'File Count' feature is seen as quite useful to keep an overview and see if uploads did work – as with that many files users cannot check every file, "so that's a good control mechanism".

- The 'Calendar' is deemed as useful for future collaborative use (what has been uploaded when, potentially also review date).

- 'Create Note' and the 'email Link' are envisioned to be useful in the future when the system is used in collaboration with more users.

- The 'Project Usage' visualisation is seen as a nice feature, which might help managing projects later on.

- The 'Search' has been tweaked during the project's lifetime, but is seen as working fine now; it is not frequently used at the moment by the users (there has to be more data up on the system), but has good features to refine searches; especially searching for values is seen as quite important as well as the time stamp of data objects.

2.6. ESCHOLAR LINKAGE: ARCHIVING & PUBLISHING

For both users and their groups the linking to publications is generally seen as beneficial; the envisioned way of archiving via a button and through eScholar is also seen as potentially useful in the future as it includes user defined access (important!), provides searchable metadata for the public and also establishes a fixed link which can be used as a data reference in publications.

2.7.   Iꜱꜱᴜᴇꜱ

There are no major issues from users' towards usability and functionality of the MaDAM system. Navigation and templates have been mentioned as needing getting used to at points, but overall usability and features have been assessed as "I think it's fine the way it is" and "I don't think there are any major things to change".

The "nice to have" mentions include I) a hierarchical view for admin purposes; II) a message after login on how much space you use and how much you got available; III) automatically filtering out Mac OS X hidden files (preceded by a dot, manually already possible to sort before upload); IV) the search could additionally list the value of the attached attribute; V) versioning would be a good way to handle updating data on the system, especially when it is also used for the actual analysing processes; VI) for the 'Calendar Review Date' an email reminder would be very useful as otherwise you have to be logged into the MaDAM system to see the alert.

A potentially major issue is the future sustainability of MaDAM as a service: Both groups would like to use MaDAM in the future also doing analysis on the data, but especially b) accentuates that the system has to be sustainable for this, with decisions to be taken on how much space is available, will it be running as a real service and on which costing model for users (e.g. on how much space a group uses vs. on how much time one spends on the system, or ..)?

2.8.   Oᴠᴇʀᴀʟʟ Bᴇɴᴇꜰɪᴛꜱ

This section is depicting the benefits especially visible in the evaluation sessions; benefits have already been reflected on in detail under [3].

Uploading large amounts of data works in a proper and expected way as the integrity of the data is kept and the whole file structure successfully replicated; added metadata makes these datasets searchable, data better findable and the whole project better manageable, saving time and storage space.

b) states that MaDAM definitely makes work easier for him (as a researcher) and his group (when collaborating with colleagues in his unit and in his position of administering the units data). Backing up data every one to two weeks, using a research log of what has been uploaded when makes him more proactive about the process of handling the research data

in terms of "what I've been working on in the last couple of weeks" and to secure the data. He also could take data off his system and put it up on MaDAM, saving storage.

2.9.   Groups' Storage Requirements and Future Use

a) and his group would need 2.5 TB (until the end of the year), this includes all current back ups ("projects that are open somehow", i.e. not immediately worked on, but still not fully completed) plus duplication through analysis (current open projects which are worked on now would only use up to 1 TB); (note: the group PI's storage requirements are unclear as his data is on a different server: it's "probably a lot"); a) himself has 1.3TB which he could probably reduce to 0.8TB in weeding out duplicates – BUT those duplicates are part of the process, when data is changed (i.e. analysed and tweaked) in a number of different approaches, the originals have to be kept and the useless duplicates can only be deleted after it is obvious they are obsolete (which doesn't include the notion of revisiting 'failed results' later as a new method in itself); the aforementioned shortcuts for 'best files' would also reduce duplicates within the MaDAM system. The Standard Microscopists aiming at using MaDAM for the actual working data as the system is fully backed up now and mature enough; it is planned for the whole group to actively use MaDAM as soon as possible (over the project's lifetime only two of the group's users have used the system regularly) as the system makes real sense for collaboration with the whole research group. The group is interested in the future of MaDAM: a hopefully larger rollout and therefore sustainability would be preferred as it secures the existence and further enhancement of MaDAM.

The Typical project data size for b) and his group is 20-40 GB; the group is at least producing half a TB a year, which consists of 100GB raw data plus analysis data; group includes 12-15 people who are regularly creating/using data; at the moment about 3TB "backed up stuff" could be added, which is lying around on diverse data media – but this would take a week to upload it and add metadata, so it would only make sense to do this if the sustainability of MaDAM would be clear.

NOTE: The numbers for data storage demands from the users' perspective are numbers *before* data is securely backed up on the MaDAM/eScholar server, which basically doubles these numbers for actual storage space.

3.   REMARKS

Setting up a project and/or archiving it always takes up some initial time for transferring the data (in the GB range) into the MaDAM system and adding metadata to this data on the system then takes up some more time. But the users are happy with these initial costs in time when they know data is secure, and the MaDAM system will be sustained; they still save time for live data they are working with in their everyday research which by their accounts by far outweighs the time investment at the start and they save valuable local space for projects they are archiving (for details on those benefits also see [3]).

The feature of providing metadata records for publishing via eScholar has gained more attention over the project's lifetime. Until the lasts three months of the project this was basically a non-issue for our users; as expected did the actual use of MaDAM, maybe updated funder/journal requirements and the exchange between users and project team lead to more interest and insight in this area. The most prominent example is an "important requirement" for a "best file" shortcut within an experiment to be able to link it to publications voiced by the Standard Microscopists.

Out of scope within the project's lifetime: 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. Additionally the 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.

4.   REFERENCES

[1] Finch, June (2011): MaDAM Final Report. Manchester, June 2011. Available under:
http://www.merc.ac.uk/sites/default/files/MaDAMFinalReport.pdf

[2] Goff, Mhorag (2010): MaDAM Requirements and Gap Analysis. MaDAM Project Report, Manchester, August 2010. Available under:
http://www.merc.ac.uk/sites/default/files/MaDAM_Requirements%20_%20gap%20analysis-v1.4-FINAL.pdf

[3] Poschen, Meik (2011): MaDAM JISC RDMI Benefits Case Study. MaDAM Project Document, Final Draft, Manchester, January 2011. Available under:
http://www.merc.ac.uk/sites/default/files/MaDAM-BenefitsCaseStudy-FINAL_draft.doc

[4] McDerby, Mary (2011): MaDAM Data Management Plan Templates. MaDAM Project Document, Manchester, June 2011. Available under:
http://www.merc.ac.uk/sites/default/files/MaDAM-DataManagementPlan.doc