Towards a generic research data management infrastructure

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

[1]Research Computer Services, University of Manchester
[2]Manchester eResearch Centre, University of Manchester
[3]John Rylands University Library, University of Manchester

1 Introduction

Until recent years, a focused and centralized strategy for the annotation, storage and curation of research data is something that has not been widely considered within academic communities. The majority of research data sits, fragmented, on a variety of disk structures (Desktops, network & external hard drives) and is usually managed locally, with little interest paid to policies governing how it is backed up, disseminated and organized for short or long term reuse.

Recognition of how current practices and infrastructure present a barrier to research, has resulted in several recent academic programmes which have focused on developing comprehensive frameworks for the management and curation of research data[1-3]. Many of these frameworks (such as the Archer suite of e-Research tools[1]), however, are large and complex, and have an overreliance on new and novel technologies making them unwieldy and difficult to support.

The paper discusses the development of a simpler framework for the management of research data through its full lifecycle, allowing users to annotate and structure their research in a secure and backed up environment. The infrastructure is being developed as a pilot system and is expected to work with data from approximately a dozen researchers and manage several Terabytes of data. The technical work is a strand of the MaDAM (Manchester Data Management) project at The University of Manchester which is funded by the JISC Managing Research Data Programme.[1]

2 System Requirements

The requirements for the data management system were based on interviews and feedback from a variety of groups within Manchester University's Medical and Life Science Departments (Poschen et al., 2010). These groups were chosen on a variety of criteria, which included the quantity and diversity of their research data, whether they are representative of typical practice within their field, and the level of time and commitment they were willing to pledge to the project. The resulting project groups include members from standard and electron microscopy (EM) and from a neuropsychiatry group involved mainly with MRI analysis

The below represents a summary of users' high level requirements for a data management infrastructure:

- Provide a centralized and backed up storage area of research data
- Provide read/write access to data amongst defined groups in a secure way
- Organize data in a way that removes redundancy (ie the need to store the same data in multiple places)
- Allow the structured annotation of research data to provide context to the data
- Allow the search and dissemination of data
- Provide a system simple enough for users to pick up and utilize without structured training.

These requirements may sound fairly basic, however, none of them are satisfactorily addressed in the researchers' current infrastructure. Initial studies indicate that, subject to successful embedding in everyday research practices (Goff et al, 2010), infrastructure that could manage the above in a sound and comprehensive way would bring significant benefits to researchers and University as a whole. Moreover, the above requirements are very generic, and thus it was established fairly early on that infrastructure that delivered the above functionality in a flexible way could easily be rolled out to research groups beyond those that took part in the pilot study. The finalized technical requirements, therefore, do not refer to any

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

piece of functionality that is specific to EM or MRI research protocols, but describe a more generic data management infrastructure.

As time with end users is limited, the project team has adopted an iterative, prototyping-based development approach, based upon continued and focused engagement with users (Poschen et al., 2010). We anticipate that 2 or 3 prototypes will be developed before the requirements are evolved and distilled into a relatively stable and definitive form. At present, the prototype is being developed using Oracle's Application Express toolkit which allows rapid development of the web based systems on top of an Oracle database, together with a modified Apache server. Once the design has been finalized, the final solution may be deployed via this technology or another, more open source route.

At the time of writing, the team is in the process of developing the second prototype based upon feedback from a user evaluation workshop involving the first prototype.

3 Software Architecture & Functionality

The components of the data management infrastructure are shown in figure 1. It is essentially a 2 tier system comprising of an upstream web based / RDBMS solution which will look after short to medium term management of research data and a long term archiving solution based on Fedora Commons digital repository which has already successfully utilized in Manchester's eScholar project.

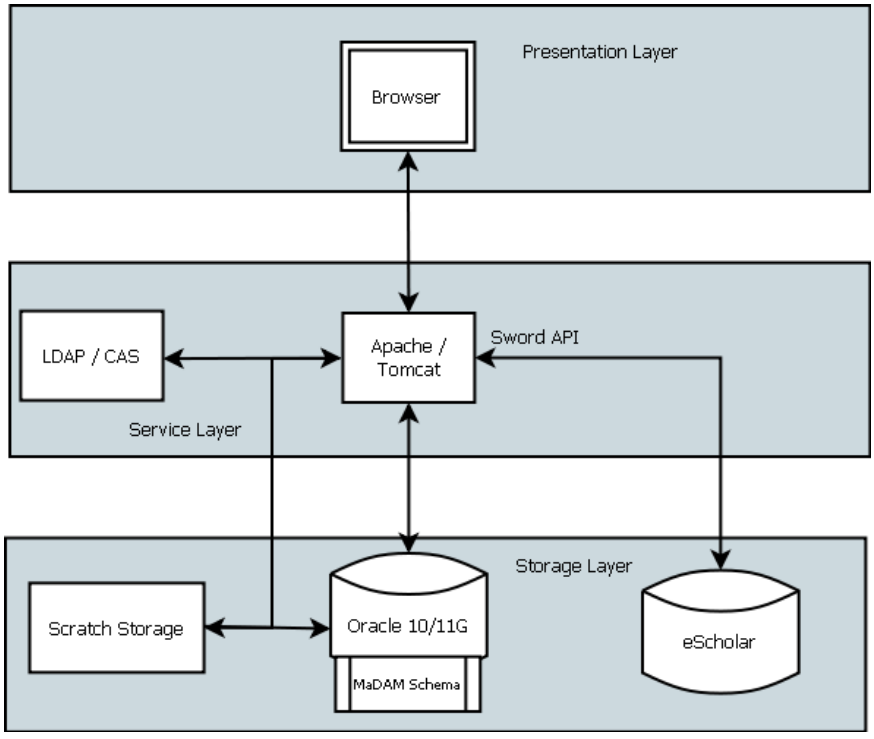

Fig 1. Architecture for the MaDAM data infrastructure.

Functionally, the data management infrastructure is a cross between simple LIMS (Laboratory Information Management System) and a CMS (Content Management System). It will provide a web interface (see Fig 2) for end users, allowing them to define and create hierarchical entities (in much the same way as a window explorer) and then annotate these entities and upload data next to them. Example entities may be "Experiment", "Sample" and "Publication". Data management tools will also allow users to define templates to tag information next to the entities, enabling them to add context to the data and to search and disseminate the data in future. Users will be able to assemble discrete groups on a project basis allowing read / write access to data. The infrastructure will be built so it can robustly and securely deal with the high file throughput of large files (the scratch storage area will be used as an overflow for large database uploads

which will be processed via background services). Users may move and tag data around the system to remove problems with redundancy and to allow flexible relationships between entities (for example, an experiment may then belong to many projects). Where feasible, users will also be able to browse and preview image (and other) data.

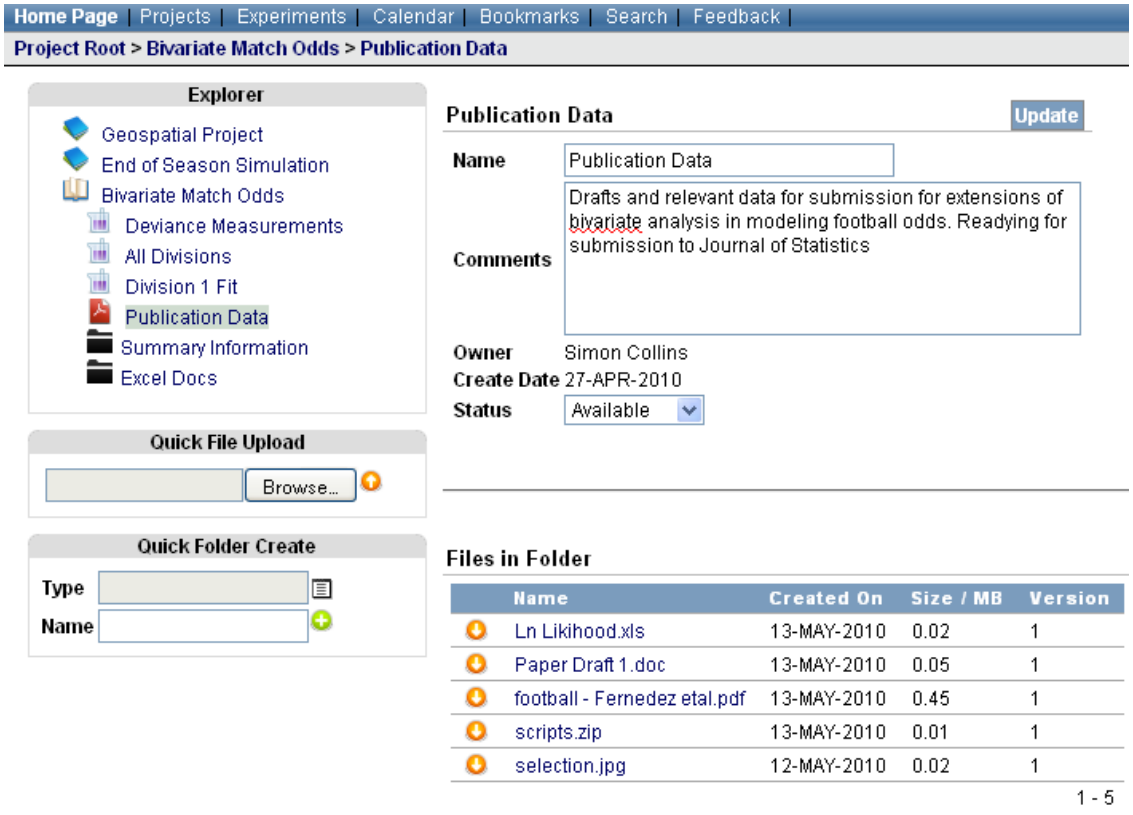

Fig 2.Project Management within the MaDAM prototype

A subset of the data in the database will be archived in the University's eScholar repository, through the SWORD API[4]. This API will be bidirectional so that archived data may be loaded back into the upstream system. Policies determined as part of this project will govern exactly which data gets absorbed into the eScholar repository. The infrastructure will support the definition of natural subsets for data archival, such as all data associated with a certain project, or all data associated with an accepted publication (it may then be possible, for example, to tie the scholarly depositions in the University eScholar system back to the raw data in the upstream MaDAM system).

4. References

[1]Steve Androulakis, ARCHER – e-Research Tools for Research Data Management. *The International Journal of Digital Curation* Issue 1, Volume 4 (2009)

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

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

[2]https://www.escidoc.org/JSPWiki/en/Overview

[3]http://islandora.ca/

[4]http://www.swordapp.org/