University of Nottingham – ADMIRe Project - Research Data Use Cases: Science Faculty

Dr Ian Chowcat & David Kay (Sero Consulting), Dr Tom Parsons (University of Nottingham) November 2012

1 - Introduction

The following scenarios for Research Data Management (RDM) in the science faculty were derived from a focus group held on 7 November 2012. Departments represented were Psychology, Computer Science, Chemistry, Pharmacy, Biology and biosciences, plus the MRI Unit.

Whilst feedback has been organised similarly across all participating faculty groups, care has been taken to remain faithful to the language used and approaches described by the practitioners.

Short questionnaires on data characteristics and researcher requirements were distributed for consideration and voluntary completion and the results are attached and incorporated in the narrative below.

2 - Data types and typical ways of working

Research groups – will often cross department and institution boundaries, involving commercial companies who have commissioned the research. Groups can have up to 250 members. The involvement of commercial concerns may mean confidentiality restrictions even within the University Department (e.g. Chemistry).

Data size – no data sets above 5TB currently. Most are smaller but a few get to this size.

Data on individuals - generally needs to be anonymised. This is normally done at the point of collection, with a single locally held database allowing trace-back to individuals if needed (e.g. in case a medical condition is identified in course of the research).

Lab books - an essential part of workflows, in some cases representing the only way to link data sets to research projects. However practice with lab books is very varied in terms of formats, whether electronic or hard copy, formal or informal, and what is deposited at the end of the research. Electronic lab books have been tried unsuccessfully – they were regarded as inflexible, and unsuitable in some experimentation contexts. While the lab books and associated practices raise some problems for research data management, the option of enforcing the use of electronic formats, or standard practice across departments, was not regarded as the solution. The issue of identifying datasets and linking them to projects should be addressed at a different level, with other issues raised by lab book usage left to individual departments to address in their own way.

Sharing outside the University – often needed, but an appropriate way to swap large data sets internationally has yet to be established. Some swap physical hard drives.

Others use Dropbox, which is liked for its ease of use and versioning, but raises privacy issues especially as the University has overseas campuses. Moreover Dropbox is not suitable for files that run into terabytes.

Public datasets – used by such as Computer Science. There are issues of managing these such as keeping up with the latest version, charges incurred for downloading. Sometimes there is a need to retain a snapshot of the public data as used for a particular research output.

Disposal – generally prefer to keep all data. There are marginal exceptions such as some videos of psychology experiments that are deleted once analysed, but generally researchers want to retain all data. Even if the data itself becomes outdated it may be needed decades later, such as in a patent dispute. Versioning is therefore important.

Real world artefacts – may need metadata that references physical objects.

Vocabularies – used in biosciences to some extent but not in all disciplines

3 – Data Management Requirements

Ingest – how the data gets into the system is an important issue and flexibility regarding 'push' and 'pull' and human involvement is required. Often it is better for the instrument to send data to storage rather the storage system harvesting it.

Storage – a key requirement with preference for data to be flowed onto central server on the university network. Needed for access by users (and by supervisors in the case of PhD students), and in the long term for access once individual researchers have left the University.

Search – the datasets themselves are typically not suitable for indexing or searching, but discovery by metadata is an important issue. A better system is needed as currently often the data is identified by project reference, thus making it hard to identify, especially retrospectively. Participants were against having a centrally allocated digital identifier, but the ability to search at file name level would be useful, and a central registry of funded project codes could help. Use of project codes in general in metadata is to be encouraged, along with project names / acronyms, which can help link related files.

Notification – can be useful for data sets that grow over time.

Annotation of data – most regarded this as marginally useful, although potentially of greater use for Psychology. Analysis of data is what is important to most.

Exposure to search engines – of very marginal importance. However, projects involved in creating data of international value (such as in plant science) may have visible web publication as a high priority.

Harvesting by open protocols – not regarded as relevant

Presentation in a user-friendly form – of limited importance

Authorisation – very important that access was properly controlled, given concerns of confidentiality of both research and research subjects.

4 - Potential Interventions

A number of interventions and support actions were identified that the University could undertake centrally:

- Handle long-term data storage, preservation and management, providing server infrastructure and backup. Departments might still opt to keep local copies of data for quick and reliable access.
- Issue advisory good practice guidance on data management, but not seek to mandate or enforce practice.
- Make available a way of securely and legally sharing large data sets with external collaborators.
- Provide a central system for exposing data which departments want to share but for which open public databases hosted by others don't exist (e.g. biosciences manage an open image database).
- Seek more uniform practice on individual research pages, and police for IP issues especially on pages put up by PhD students. These pages are largely for papers and presentations not undigested data.
- Provide a central registry of grant codes to aid discovery.
- Provide paradata on use of assets such as data sets, which can be useful for REF impact measurement.
- Provide guidance on licensing issues as currently data is typically not licensed when it is shared.

5 - Omissions

There was no mention of

- The role of Research Council repositories
- Research project data plans
- Datasets that cannot be hosted by the university for reasons of above-campus collaborative arrangements or required proximity to equipment

Questionnaire responses – Science Faculty (5 responses)

Your requirements

Operations	RELEVANCE >	High	Med	Low	Zero
Ingest	Getting the data into the system	4	1		
Storage	Storing for long term retention	5			
Replication	Replicating the data to other instances and for safety	5			
Search	Selective retrieval of data	2	2	1	
Index	Indexing based on full text or facets to optimize retrieval		2	2	
Notification	Notifying other instances or users of changes		3	1	1
Annotation	User generated annotation of records, such as notes and ratings	1		3	
Exposure	Tagging to be indexed by search engine spiders / robots			2	3
Harvesting	Open to harvesting via OAI-PMH				4
Presentation	Presentation useful to humans, such as listings and visualizations			3	1
Authorisation	Control of access based on appropriate granularity	4			

Your data

Data Set	RELEVANCE >	High	Med	Low	Zero
Metadata	Description of assets, such as Title, Author	4			
Paradata	Use of assets, such as Activity, Actor, Context, Date, Volume	1		2	1
Identity	Allocation of a unique digital identity to each asset (URI, DOI)	2	2		
Files	The digital objects themselves or related assets	3	1		
Stuff	Real world artefacts that need to be referenced	3			1
Vocabularies	Standardised terms used in metadata and paradata		4		
Licensing	Explicit licensing as open data (e.g. Creative Commons)		1	2	
Copyright	Necessary statements	1		2	
Links	Links to internal and external systems (ePrints, CRIS, RC)		2	1	