University of Nottingham – ADMIRe Project - Research Data Use Cases: Medicine & Health Sciences 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 in the Medicine and Health Sciences faculty were derived from a focus group held on 20 November 2012. Representatives attended from the Schools of Clinical Sciences, Veterinary Medicine and Science, Community Health Sciences, plus the Advanced Data Analysis Centre.

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

All research needs to comply with the Department of Health's Research Governance Framework.

Research groups – research is typically carried out in research groups and in many cases the project outlives individual researchers. This makes sharing within research groups important but in practice individually designed filing structures can make extracting previous research problematic. However in some settings (e.g. psychiatry) research may be more individual, involving such as patient questionnaires.

Data size – large files and datasets are often generated (of terabyte proportions). Where analysis has to be carried out remotely from data storage, there can be connectivity issues.

Data on individuals – a regular occurrence, generally needing to be anonymised with separate storage of personal details.

Lab books - used as standard. Most departments use paper lab books but at least one department also demands digital transcription of entries. The quality of lab books varies greatly. Some research groups enforce standards but not all.

Sharing – it is common for data to be deposited in external repositories (national, international) and some journals require this for publication. Often shared data is linked to personal data that is retained locally. However some data is not shared (e.g. MRI imaging), and some may be just retained by the researcher (e.g. survey data generated in PhD research).

Note that, unlike other faculties, Dropbox is not used for sharing as it is regarded as insufficiently secure.

Disposal – data is never discarded.

Real world artefacts – a complex and important interface in medical research. Some groups have databases linked to physical samples and tissue banks. Samples might be bar-coded and traceable back to research through lab books. There was some interest in better linkage of physical assets and data to which it related. However database ownership is often split between the University and the NHS, which introduces complications. Note that storage of human tissue is covered by specific legislation.

Vocabularies – used in some cases, involving some well-known taxonomies.

Metadata – might be stored in lab books, although it would be preferable for it to be stored with the data. In other cases the metadata is stored with the raw data but is not necessarily carried across into analysed sets.

Paradata – may be of use for evidencing impact. Where appropriate, Google Analytics data would be used by some.

Identity – one department has a standard system for file IDs but there is no uniform approach across the faculty.

Licensing – all data needs to contain a reference to the ethical approval covering it. Access may be restricted to specified research groups or named individuals. If ethical approval is given by an external body as well as by the University Ethics Committee then a copy should be stored with the data. Making these connections is a key requirement.

There was some interest in making software or scripts developed during research more widely available under an open software licence.

Copyright – some research will involve a commercial copyright.

3 – Data Management Requirements

Storage – fast-access to large disk space for live data, which gets frequently backed up, as well as a long-term archiving and preservation service.

Search - a university-wide data archiving system is needed. As well as keeping data secure, this should allow easy discovery of past research - particularly helpful to discover unpublished research (often due to obtaining negative results).

However ethical concerns prevent too much information being contained in metadata that anyone can access: in many cases no more than a grant number and project title can safely be given. However this is not a problem as researchers are only likely to search for data with which they already have established such a relationship.

Notification – needed by senior staff but not by more junior researchers

Annotation – could be useful to note re-use of data by other researchers.

Transformation – bulk operations are frequently done, such as aggregation, anonymisation and format transformations. However the complexity of specifying parameters probably prohibit this being done centrally using standardised code.

Exposure to search engines – not needed.

Presentation – there was a desire to link staff e-profiles to project websites where they exist. However research details appearing on individual researcher pages must be controlled by the individual as there are many sensitive subjects, e.g. research involving animal trials, or matters of public controversy where individual safety could be put at risk.

4 - Potential Interventions

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

- Provide adequate short-term disk space for large files that need to be analysed, and frequently back it up
- Provide an archiving and data preservation solution, including ethics documentation and links to physical assets
- Enable better searching of research data across departments, utilising meta-data (within the constraints of ethical approvals)
- Provide a better way to link staff e-profiles to project websites, under the control of individual members of staff.
- Provide advice, guidance and training on using technology to manage physical assets
- Provide guidance on licensing issues

5 - Omissions

There was no mention of

- The role of Research Council repositories
- Research project data plans

Questionnaire responses – Medicine & Health Sciences Faculty (6 responses)

Your requirements

Operations	RELEVANCE >	High	Med	Low	Zero
Ingest	Getting the data into the system	3	2	1	
Storage	Storing for long term retention	5	1		
Replication	Replicating the data to other instances and for safety	5	1		
Search	Selective retrieval of data	4	1	1	
Index	Indexing based on full text or facets to optimize retrieval	1	3	2	
Notification	Notifying other instances or users of changes	1	2	3	
Annotation	User generated annotation of records, such as notes and ratings		4	2	
Exposure	Tagging to be indexed by search engine spiders / robots		2	1	3
Harvesting	Open to harvesting via OAI-PMH		1	2	3
Presentation	Presentation useful to humans, such as listings and visualizations		1	4	
Authorisation	Control of access based on appropriate granularity	6			

Your data

Data Set	RELEVANCE >	High	Med	Low	Zero
Metadata	Description of assets, such as Title, Author	4	2		
Paradata	Use of assets, such as Activity, Actor, Context, Date, Volume	2	1	3	
Identity	Allocation of a unique digital identity to each asset (URI, DOI)	1	4	1	
Files	The digital objects themselves or related assets	5	1		
Stuff	Real world artefacts that need to be referenced	3	2	1	
Vocabularies	Standardised terms used in metadata and paradata	3	2	1	
Licensing	Explicit licensing as open data (e.g. Creative Commons)	1	1	4	
Copyright	Necessary statements	1	1	3	1
Links	Links to internal and external systems (ePrints, CRIS, RC)	2	2	2	