University of Nottingham – ADMIRe Project - Research Data Use Cases: Engineering 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 Engineering faculty were derived from a focus group held on 20 November 2012. Departments and research divisions represented were Architecture and Built Environment, Electrical and Electronic Engineering (Applied Optics), Manufacturing and the Geospatial Institute.

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 – many groups, some inter-disciplinary, some distributed. It is common to work with industry partners, often involving commercial confidentiality agreements.

Data size – varies greatly. Some work with quite small files, others generate large volumes on a regular basis, though not typically 'big data' (small numbers of terabytes).

Lab books - often used but not uniformly across the faculty. Where they are used, e.g. in applied optics research, they can be a crucial part of the data trail, referencing such as filenames. Currently these tend to be paper books but it would be desirable to use online formats such as wikis. Extensive use can be made of past lab books.

Sharing - whether or not lab books are used a general issue was identified around 'passing the baton' when researchers change on continuing projects. Often the record keeping practice of a previous researcher can be hard to understand – a standard format for folder and file structure would help, with notes to document processes. Sometimes past research actions can only be reconstructed through email trails.

There is often a need to collaborate on shared documents and it would be desirable to find a better way of collaborating than via emails. Some individuals use Dropbox to share files but pay for this themselves - it is easy to use, has a reasonable version control, and is good for sharing large files. Sharing via FTP also takes place.

For shared document editing a system with proper version control is desirable – Dropbox is not the tool for this, and Word's version controlling is regarded as inadequate.

Generally data is not currently made publicly available, although the applied optics research group have a public website which displays sample images and allows users to contact the group if they want more information; commercially sensitive data is accessed through a password protected wiki.

External data sets – used by a number: can be owned by external customers (e.g. Network Rail) or bought under licence (e.g. Ordnance Survey). There are a number of cases where departments somewhere in the University have bought external data which others also want to use, but there is no system for discovering this compounded typically by uncertainty whether the licence allows sharing even within the University.

Real world artefacts – physical materials and samples used in research need to be better managed. A cross-university approach would be helpful to access previous assessments of cost and risk, especially for departments where acquiring some materials is exceptional (e.g. hazardous chemicals in a setting where chemicals are not normally handled).

Vocabularies – important in domains such as map data. In some disciplines the desire to standardise vocabularies is compromised by competing proposals.

Metadata – some is instrument-generated, some hand-coded by individual researchers. There are some external metadata standards but not in all research areas.

Paradata – regarded to be of limited use, aside from generating data for behavioural analysis where this might be appropriate for the research.

Licensing – it is important to store licence terms along with data, e.g. to cover attribution if reused.

Disposal – in general all iterations of data sets are retained. Some need to interrogate historic data, others keep even failures as they may be useful for future research and learning.

3 – Data Management Requirements

Ingest – any central service needs to leave maximum scope for local autonomy; for example, centralised file naming conventions would be unhelpful.

Storage – the storage space the university currently provides is inadequate: this is one of the drivers for individual use of Dropbox, and for use of local backup servers. Long-term digital preservation is a key issue that is currently approached haphazardly. There is a need both to ensure data stored on old media is migrated to up-to-date formats, and that the data remains accessible. This can mean storing a version of the software that can read the original data, or a link to where it can be obtained, along with the data itself.

Alternatively a good description of the file format used should be stored. Amazon's S3 (Simple Storage Service) was cited as a storage model.

Search – usually applies to the metadata not the data itself. It would be useful to be able to find what other departments have.

Notification and *Annotation* – of interest to some.

Exposure to search engines – mostly not needed, although the impact agenda may increase pressure. Beware of websites that are left to wither once the project ends.

Presentation – a better way of linking the "who I am" contained in staff e-profiles with the "what I do" of projects was thought to be needed. The e-profile template should contain scope for project links, which could be made to project websites or personal pages. Putting research data directly on e-profiles alongside publications was not favoured. There was also concern about how e-profiles were discoverable on Google if searches were made just on the researcher name.

Authorisation – in general, open access to research data was not favoured.

4 - Potential Interventions

A number of interventions and support actions were identified that the University could undertake centrally. Principally, the University should provide a walled garden environment that provides economies of scale and secure provision for what many researchers often currently have to do either on their own or using commercial provision. This should include

- Data storage adequate to the file sizes now being generated
- A secure provision for data sharing which works as seamlessly as Dropbox
- A collaborative document editing tool like Googledocs that also provides proper version control
- A better link between the e-profiles of individual researchers and project or personal websites, which are usually the appropriate places to expose public research data
- Facilities for discovering assets held by individual departments that could be of use to others
- Guidance on the range of licensing issues, both for data that is purchased and data generated by researchers that is made available to others (including later researchers)
- Advice and guidance on how to ensure data remains accessible to future researchers, through using recognised filing structures, proper documenting of procedures and formats, storage of software along with data, etc – but without imposing rigid requirements across the university
- Training on data management to researchers that can create the right culture and social norms across the university.

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 – Engineering Faculty (9 responses)

Your requirements

Operations	RELEVANCE >	High	Med	Low	Zero
Ingest	Getting the data into the system	8	1		
Storage	Storing for long term retention	9			
Replication	Replicating the data to other instances and for safety	6	3		
Search	Selective retrieval of data	7	2		
Index	Indexing based on full text or facets to optimize retrieval	1	5	2	1
Notification	Notifying other instances or users of changes	1	5	3	
Annotation	User generated annotation of records, such as notes and ratings	4	3	2	
Exposure	Tagging to be indexed by search engine spiders / robots		2	5	2
Harvesting	Open to harvesting via OAI-PMH		2	3	3
Presentation	Presentation useful to humans, such as listings and visualizations	4	3	1	1
Authorisation	Control of access based on appropriate granularity	6	3		

Your data

Data Set	RELEVANCE >	High	Med	Low	Zero
Metadata	Description of assets, such as Title, Author	4	5		
Paradata	Use of assets, such as Activity, Actor, Context, Date, Volume		4	3	2
Identity	Allocation of a unique digital identity to each asset (URI, DOI)	4	2	2	1
Files	The digital objects themselves or related assets	8	1		
Stuff	Real world artefacts that need to be referenced	1	4	2	2
Vocabularies	Standardised terms used in metadata and paradata	3	4	1	1
Licensing	Explicit licensing as open data (e.g. Creative Commons)		5	4	
Copyright	Necessary statements	1	6	2	
Links	Links to internal and external systems (ePrints, CRIS, RC)	1	4	3	1