University of Nottingham – ADMIRe Project – Combined Focus Group Findings

Dr Ian Chowcat and David Kay, Sero Consulting – November 2012

Table of Contents

1 – DATA.......... 2

2 - PRACTICE RELATING TO RESEARCH DATA.......... 6

3 – DATA MANAGEMENT FUNCTIONS.......... 8

4 - POTENTIAL INTERVENTIONS.......... 12

5 - OMISSIONS FROM DISCUSSIONS

6 - QUESTIONNAIRE RESPONSES – RANKED FOR ALL FACULTIES.......... 15

Scope

This document compiles the statements in the four faculty focus group reports under a common series of headings. The statements are not edited in to a vanilla form so faculty nuances remain. Statements are occasionally repeated across categories where value is added.

This compilation was undertaken prior to receiving participant feedback on the individual focus group reports – requested for 7th December.

It is intended that this document should feed directly in to the next version of the Requirements Catalogue and should also inform high level project scoping.

Key to faculties

Comments are colour coded according to faculty

- Black – Arts & Social Sciences
- Blue – Engineering
- Orange – Medicine & Health Sciences
- Plum - Science

1 – Data

This section set out the range of data, metadata and associated assets (such as physical artefacts) identified by focus group participants. The list of data types and supporting functions suggested by the consultants (see Sections 3 & 5) was introduced and assessed AFTER the open discussion of data and practices reported here.

1.1 - Data size

- 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).
- Large files and datasets of terabyte proportions are often generated. Where analysis has to be carried out remotely from data storage, there can be connectivity issues.
- No datasets above 5TB currently. Most are smaller but a few get to this size.

1.2 - Data Identifiers

- One department has a standard system for file IDs but there is no uniform approach across the faculty.
- 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.
- The issue of identifying datasets and linking them to projects should be addressed at a different level [than standardising lab books], with other issues raised by lab book usage left to individual departments to address in their own way.

1.3 – Digital Data Types

Quantitative datasets

- Could be large sets of survey or experimental data, e.g. in economics, through to small datasets conducted by graduate students in many disciplines (even philosophy in some instances). In the case of quantitative data the use of *spreadsheets* for recording, analysing and sharing is widespread, alongside use of analytical software such as SPSS.

Text-based data

- For example the JISC funded project to digitise and make available the 86 volumes Survey of English Place Names; such texts may be subject to computer-enabled mining, as in Politics, generating indices and other forms of analysis.

Multimedia data sets

- As in film and broadcast, or 3D images in sculpture or archaeology; these can generate very large files and require specialised players in some cases.

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.
- 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.
- An essential part of workflows, in some cases representing the only way to link data sets to research projects.
- 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.

Other personal research data

- Personal notes, broadly speaking, collected by individual researchers of all sorts, including those engaged solely in theoretical and conceptual studies. Typically this is formative material for later published outputs, not required to be made public either by researchers or funding councils. In some disciplines, such as philosophy, literary criticism, and the theoretical end of most departments in the two faculties, these are the only form of research data that projects will typically produce. Although analogous in some senses to lab books in scientific research such material does not form a systematic part of the workflow.

1.4 - Digitised Assets

- The need to digitise data that is held in paper form looms large. Both English and Geography reported large projects that involved digitisation. In the case of Economics data was being collected in developing countries and often has to be recorded initially on paper, although use of mobile devices for field data collection is being considered.

1.5 - Physical artefacts and ephemera

- Could comprise curated archival collections as gathered through ethnographic research.
- 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).

- Real world artefacts represent 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.
- May need metadata that references physical objects.

1.6 - Metadata

Descriptive metadata

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

Vocabularies

- Important in domains such as map data. In some disciplines the desire to standardise vocabularies is compromised by competing proposals.
- *Vocabularies* are used in some cases, involving some well-known taxonomies.
- Used in biosciences to some extent but not in all disciplines.

1.7 - Paradata (aka analytical or usage data)

- Regarded by some as useful to help demonstrate impact, which can be a difficult issue for the more theoretical disciplines. However it was seen as a very blunt instrument and could be misleading for some of the small, specialised niche datasets these disciplines generate.
- Of limited use, aside from generating data for behavioural analysis where this might be appropriate for the research.
- May be of use for evidencing impact. Where appropriate, Google Analytics data would be used by some.

1.8 - External datasets (public or commercial)

- Very commonly used as inputs across disciplines. The point was made that, while these include datasets as scientists would understand them, for arts and social science researchers the key research data is found in the book and journal collections managed by the Library.
- Frequently used. They 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.

- 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.

2 - Practice relating to research data

2.1 – Data Governance

- All research needs to comply with the Department of Health's Research Governance Framework.
- While the lab books and associated practices raise 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.

2.2 - Handling Personal Data

- Data linked to individuals is held in some cases and needs to be anonymised. In some cases the data being analysed is old and the individuals no longer living, but there are instances of current experimentation, as in economics.
- In the case of ethnographic surveys it can be very difficult to anonymise data, because of the in-depth and personal nature of the research, and there is no clear understanding of how to make it available in such cases.
- Data on individuals is a regular occurrence, generally needing to be anonymised with separate storage of personal details.
- 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).

2.3 – Working in Research Groups

- Engineering has many groups, some inter-disciplinary, some distributed. It is common to work with industry partners, often involving commercial confidentiality agreements.
- Medical 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.
- Science 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).

2.4 - Sharing Files

- Sharing outside the university is often essential involving Dropbox, Mendeley and shared websites. However in other cases there was some resistance to sharing hard-won data too freely.
- 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.
- 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 and other data is not shared (e.g. MRI imaging). Note that, unlike other faculties, Dropbox is not used for sharing as it is regarded as insufficiently secure.
- Sharing outside the University is 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.

2.5 – Collaborative Authoring

- 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.

2.6 - Disposing of data

- Disposal is regarded as never needed, indeed the view was that the University was too quick to dispose of resources. In these some of these disciplines, datasets which are out of date become historical artefacts and so the object of potential research in their own right.
- 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.
- Data is never discarded.
- 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.

3 – Data Management Functions

This section contains responses to a set of functions proposed for consideration by the consultants, broadly based on the California Digital Library (CDL) model. Several of these aspects were raised in the discussion of practice reported in Section 2.

3.1 – Authentication & Authorisation

- A controversial issue - while some favoured open access to data, others wanted to insist on individual registration before data could be accessed to provide further information on use.
- In general, open access to research data was not favoured.
- It is very important that access was properly controlled, given concerns of confidentiality of both research and research subjects.

3.2 - Licensing, Copyright & Intellectual Property (IP)

- Some data has licence agreements but some is openly published without a licence. Problems often arise for researchers in the Arts because of out-dated IP law preventing old collections from being digitised as no one can give the necessary permissions.
- It is important to store licence terms along with data, e.g. to cover attribution if reused.
- 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.
- Some research will involve a commercial copyright.

3.3 - Identification

- One department has a standard system for file IDs but there is no uniform approach across the faculty.
- 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.
- The issue of identifying datasets and linking them to projects should be addressed at a different level [than standardising lab books], with other issues raised by lab book usage left to individual departments to address in their own way.
- A central registry of grant codes would aid naming and discovery.

3.4 - Ingest

- Data needs to be under the control of the researcher who knows the material. Although issues of confidentiality are less widespread than in Science they do exist; for example, one research project in Sociology works with the NHS.
- Any central service needs to leave maximum scope for local autonomy; for example, centralised file naming conventions would be unhelpful.
- 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.

3.5 - Storage

- A key issue about which there is currently much uncertainty. In many cases it seems that data is retained on researchers' own computers. In other cases there is vagueness about where the data is stored, who controls it, and an alarming story of software upgrades on central servers corrupting data sets. Nonetheless storage of data on central servers was generally favoured providing it was properly managed.
- Storage of archival materials, which can be a mix of digital and real-world artefacts, is a neglected area, currently left entirely to individual researchers, with collections often being lost when they move on as there is no university museum function. The Middletown Archive at Ball State University was cited as a positive model.
- 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 Elastic Compute Cloud was cited as a storage model.
- Fast-access to large disk space for live data, which gets frequently backed up, as well as a long-term archiving and preservation service.
- Storage is 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.

3.6 - 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.

3.7 - Discovery

- There is a range of searching needs. Some quantitative data is only searchable by analytical packages. However more qualitative data needs to be searchable. Tagging is often used, while others rely on filtering in spreadsheets.
- Search usually applies to the metadata not the data itself. It would be useful to be able to find what metadata other departments have.
- 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.
- 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.

3.8 - Exposure to global search engines

- Mostly not needed, although the impact agenda may increase pressure. Beware of websites that are left to wither once the project ends.
- 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 to purchase more, which are through a password protected wiki.
- Not needed
- Of 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.

3.9 - Presentation in a user-friendly form

- Of considerable importance for qualitative data, and often a key part of funded projects is to make data more accessible, usable and reusable.

- Need 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.
- 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.
- Presentation in a user-friendly form is of limited importance

3.10 - Annotation of data

- This was regarded as most useful as a way of generating new research data, which was only of use if there was resource available to analyse it.
- Of interest to some
- Could be useful to note re-use of data by other researchers.
- Most regarded annotation as marginally useful, although potentially of greater use for Psychology. Analysis of data is what is important to most.

3.11 - Notification about new data

- Of interest to some
- Needed by senior staff but not by more junior researchers
- Notification can be useful for data sets that grow over time.

4 - Potential Interventions

During the discussions, a number of interventions and support actions were identified that the University could undertake centrally. These are summarised as follows.

4.1 - Metadata

- Provide a central registry of grant codes to aid naming and discovery.

4.2 - Current Research Tools & Services – Live Research Data

- 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.
- Data storage adequate to the file sizes now being generated
- Provide adequate short-term disk space for large files that need to be analysed, and frequently back it up
- 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
- Make available a way of securely and legally sharing large data sets with external collaborators
- Provide paradata on use of assets such as data sets, which can be useful for REF impact measurement

4.3 - Storage & Preservation – Archived Research Data

- Handle long-term data storage, preservation and management, providing a server infrastructure and backup. Individual researchers might still opt to keep local copies of data for quick and reliable access
- Develop facilities for accessible storage of archival materials that mix both digital and physical assets
- Provide an archiving and data preservation solution, including ethics documentation and links to physical assets
- 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.

4.4 - Discovery

- Develop systems for exposing data in a variety of formats
- Facilities for discovering assets held by individual departments that could be of use to others
- Enable better searching of research data across departments, utilising meta-data (within the constraints of ethical approvals)

- 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
- Provide a better way to link staff e-profiles to project websites, under the control of individual members of staff
- Provide a central system for exposing data that departments want to share but where open public databases don't exist

4.5 - Policy & Guidance 1 – Access & Licensing

- Develop policy on making data openly available when anonymity can easily be compromised, as in the case of ethnographic data
- Develop policy on whether data can be accessed without registration being required
- Provide guidance on licensing issues
- 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)
- Provide guidance on licensing issues
- Provide guidance on licensing issues as currently data is typically not licensed when it is shared.
- Clarify the law on digitising assets when the rights holders no longer exist

4.6 - Policy & Guidance 2 – Research Data Practice

- 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.
- Provide advice, guidance and training on using technology to manage physical assets
- Issue advisory good practice guidance on data management, but not seek to mandate or enforce practice
- 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.

.

5 – Omissions from discussions

There was little or no reference in discussions to the following aspects.

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

6 - Questionnaire responses – Ranked for all faculties

6.1 - Faculty requirements

Operations	RELEVANCE >	A&SS (5)	Eng (9)	MHS (6)	Sci (5)	ALL (25)
Storage	Storing for long term retention	4	9	6	5	24
Replication	Replicating the data to other instances and for safety	4	9	6	5	24
Ingest	Getting the data into the system	4	9	5	5	23
Authorisation	Control of access based on appropriate granularity	2	9	6	4	21
Search	Selective retrieval of data	3	9	5	4	21
Index	Indexing based on full text or facets to optimize retrieval	1	6	4	2	13
Notification	Notifying other instances or users of changes	1	6	3	3	13
Annotation	User generated annotation of records, such as notes	1	7	4	1	13
Presentation	Presentation for humans, such as lists, visualizations	4	7	1	0	12
Exposure	Tagging to be indexed by search engines	1	2	2	0	5
Harvesting	Open to harvesting	0	2	1	0	3

6.2 – Faculty data

Data Set	High & Medium RELEVANCE >	A&SS (5)	Eng (9)	M&HS (6)	Sci (5)	ALL (25)
Metadata	Description of assets, such as Title, Author	4	9	6	4	23
Files	The digital objects themselves or related assets	2	9	6	4	21
Identity	Allocation of a unique digital identity to each asset	4	6	5	4	19
Vocabularies	Standardised terms used in metadata and paradata	2	7	5	4	18
Stuff	Real world artefacts that need to be referenced	1	5	5	3	14
Copyright	Necessary statements	3	7	2	1	13
Links	Links to internal and external systems (ePrints, CRIS, RC)	2	5	4	2	13
Paradata	Usage of assets, by such as Activity, Actor, Context	4	4	3	1	12
Licensing	Explicit licensing as open data (e.g. Creative Commons)	1	5	2	1	9