University of Nottingham – ADMIRe Project - Research Data Use Cases: Arts & Social Sciences Faculties

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 arts and social sciences faculties were derived from a focus group held on 15 November 2012. Departments represented were Philosophy, English, Education, Economics, Geography and Politics, along with the Digital Arts and Humanities 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

The discussion identified a spectrum of research data types in arts and social science departments each with its own dynamics:

(i) *Quantitative datasets* which 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);

(ii) *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;

(iii) *Multimedia data sets* as in film and broadcast, or in the generation of 3D images in sculpture or archaeology. These can generate very large files and require specialised players in some cases;

(iv) *Qualitative research data in multiple media*, including physical artefacts and ephemera, perhaps gathered through ethnographic research, which could comprise curated *archival collections*;

(v) *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 and is 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.

Public data sets are very commonly used as inputs across disciplines. In social sciences some researchers, including many economists for example, would often utilise publically available quantitative data sets. However, for other researchers in arts and some approaches in the social sciences the key research data is found in the book and journal collections managed by the Library.

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.

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.

Data on individuals is held in some cases and needs to be anonymised in the main datasets. 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, and there can be a need to store personal data to enable repeat visits to these individuals. In such cases a typical approach is to store ID numbers in the main anonymised data sets which link to a separate database of personal details. In the case of ethnographic surveys it can be 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 (e.g. photos, personal letters etc).

Paradata (also known as analytical or usage data) was 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.

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.

Disposal of data is thought to be 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.

3 – Data Management Requirements

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.

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.

Search – there was a range of searching needs. Those who work with large quantitative datasets tend to use analytical packages to conduct their searches, and often there is no other practical way of searching the data. To search qualitative data tagging is often used, while others rely on filtering in spreadsheets.

Annotation of data – some researchers, such as economists, annotate their large quantitative datasets. Sometimes this is done within the dataset itself to identify variable labels. More often, when a dataset is made publicly available or passed from one researcher to another, it comes with a text document describing the dataset (number of observations, sample design, etc.) and a list of variable names, definitions and descriptions. For others, annotation 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.

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.

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 the use being made of it.

IP issues – 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.

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 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;
- Develop systems for exposing data in a variety of formats;
- 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;
- Clarify the law on digitising assets when the rights holders no longer exist.

5 - Omissions

There was no mention of

- The role of Research Council repositories, although it was commented that the closure of the Arts and Humanities Data Service has meant the loss of a national repository for institutional collections in those disciplines.
- Research project data plans

Questionnaire responses – Arts and Social Sciences Faculties (5 completed)

Your requirements

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

Your data

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