DATA MANAGEMENT PRACTICES IN THE SOCIAL SCIENCES

Veerle Van den Eynden, Libby Bishop,
Laurence Horton and Louise Corti

10 AUGUST 2010

T +44 (0)1206 872001
E datasharing@data-archive.ac.uk
www.data-archive.ac.uk

UK DATA ARCHIVE
UNIVERSITY OF ESSEX
WIVENHOE PARK
COLCHESTER
ESSEX, CO4 3SQ

WE ARE SUPPORTED BY THE **UNIVERSITY OF ESSEX**, **THE ECONOMIC AND SOCIAL RESEARCH COUNCIL**, AND THE **JOINT INFORMATION SYSTEMS COMMITTEE**

Contents

1. Summary

Evidence has been gathered on existing data management practices and approaches to data management planning amongst researchers funded by the ESRC in its various investments. Data management was evaluated in the interdisciplinary Rural Economy and Land Use programme (Relu) and in the longitudinal qualitative Timescapes programme – two investments where special emphasis has been placed on data sharing which requires good data management practices. With ESRC research centres representing large and long-term research investments, data management practices in a selection of such centres was evaluated by interviews with directors and researchers. Data for individual research awards were compiled from ESDS information.

The communality is that all these research investments are bound by the ESRC data policy, which means that data need to be made available to the wider research community for re-use when research projects end.

The Relu programme approach is to support researchers to plan their own data management and to implement their own good data management practices through a programme-specific data policy that mandates data archiving and a dedicated support service funded by the research councils. The support service provides best practice guidance and tools for researchers to use. Data are being archived at existing research council infrastructures. Crucial is also the strong emphasis the programme director places on data sharing.

This programme piloted data management plans for the ESRC. Valuable lessons have been learnt about the usefulness of such plans. Researchers need clear information on how to plan data management in a meaningful way and often need additional support to develop good management procedures. Especially where research data may be confidential or sensitive, researchers need guidance on suitable informed consent procedures and anonymisation guidelines. Planning data management does not guarantee its implementation, and research funders need to consider how to ensure that good data management intentions are indeed implemented and revisited.

Timescapes provides an example of a centralised approach to data archiving and data management procedures at programme level. The programme built its own archive and has provided guidelines and tools for informed consent procedures that take data archiving into account, as well as anonymisation, transcription and documentation guidelines. Even with this central approach, engaging researchers into designing suitable data management and archiving solutions has been crucial to ensure researcher participation and workable solutions. Ultimately researchers have to implement the management procedures, which most of them have done.

The nature of the qualitative longitudinal research brings with it highly problematic data in terms of their management and archiving - data are sensitive, confidential, difficult to anonymise and require at times strict access control systems. Even with dedicated project funding, strong leadership and support services, significant challenges arose in creating archive-ready data.

In ESRC research centres and programmes the director may coordinate data management as part of research management. In practice most aspects of data management are the responsibility of individual researchers. Although data management is not formally planned, certain aspects are as part of ethical review procedures. Overall data management and data archiving is not costed in or planned much during a centre's planning stage.

Researchers have indicated that they want easy, practical and trustworthy solutions they can embed into research activities, rather than a range of guidelines or suggestions from which to choose. Centres often need solutions for easy file sharing, either for cross-institutional collaborations, or for remote working.

For individual grant holders the onus is on researchers themselves to look after data. Support is provided by ESDS via online guidance and a helpdesk answering queries. The main aspects for which researchers seek guidance is dealing with confidential research data, gaining consent for data to be archived, copyright of data and the costing of data management in grant applications.

When data are offered to ESDS for archiving, the main limiting factors are found to be lack of consent for data archiving and uncertainty over how to enable the archiving of confidential data.

2. Introduction

The 'Data Management Planning for ESRC Research Data-Rich Investments' project (DMP-ESRC)[1], funded by the Joint Information Systems Committee (JISC) under the Managing Research Data Programme, aims to:

- evaluate existing data management practices amongst researchers in the social sciences community
- help develop and implement effective data management planning procedures and tools in the research lifecycle
- expand individual and institutional data managing and sharing capacity by providing best practice guidance, support and training.

The project is coordinated by the Research Data Management team at the UK Data Archive.

Presented in this report is evidence gathered on existing data management practices and approaches to data management planning amongst researchers funded by the ESRC in its various investments – ESRC research centres, research programmes and individual research awards. This evidence is based either on experiences of data support services coordinated by the UK Data Archive, or on information gathered from researchers in research centres and programmes funded by the ESRC.

The UK Data Archive has managed the Economic and Social Data Service (ESDS) since 2003. In addition to its key activities of preserving and disseminating qualitative and quantitative social and economic data and providing user support and training for secondary use, ESDS also supports ESRC-funded researchers as they prepare and archive their research data. Since 2004, a pro-active data management and sharing service has been pioneered at the Archive by hosting the Data Support Service for the interdisciplinary Rural Economy and Land Use Programme. In this programme project-level data management planning is implemented and detailed guidance on good data management practices provided to researchers. The Archive also has a strong link with the qualitative longitudinal Timescapes programme, where data archiving is one of the core activities. Emphasis has been placed on developing an archive infrastructure as well as providing guidelines to researchers for managing and archiving often problematic qualitative research data.

Information presented in this report is based on data management experiences for:

- the Rural Economy and Land Use programme (Relu)[2]
- the qualitative longitudinal study Timescapes[3]
- eight past and present ESRC Research Centres and Programmes
- individual researchers (ESRC award holders), based on information gathered by the Economic and Social Data Service.

2.1. Data policy context

The Economic and Social Research Council (ESRC) is at the forefront of data sharing in the UK and in its data policy[4] requires researchers to offer all research data resulting from research grants to a designated data centre, the UK Data Archive. The Archive supports ESRC applicants and award holders in enabling data sharing for both quantitative and qualitative data, through the Economic and Social Data Service (ESDS). ESDS also ensures preservation and dissemination of archived research data in order to make them available to the research, learning and teaching communities.

2.2. Data management

Data management in research encompasses all aspects of looking after, handling, organising and enhancing research data. Managing data well enhances the scientific process, ensures high quality data and also increases the longevity of data and opportunities for data to be shared and re-used.

---

[1] www.data-archive.ac.uk/create-manage/projects/JISC-DMP

[2] www.relu.ac.uk

[3] www.timescapes.leeds.ac.uk/

[4] http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-12051.pdf

Key data management areas for the social sciences domain that were considered for this report are:

- data management planning in research design
- ethics, consent and confidentiality
- data copyright and rights management
- contextualising, describing and documenting data
- data formats and software
- data storage, back-up and security
- roles and responsibilities of data management

3. ESRC investments evaluated

3.1. Relu

The Rural Economy and Land Use (Relu) programme (2004-2011) is an interdisciplinary research programme, funded by the ESRC, NERC and BBSRC. Interdisciplinary teams of social and environmental scientists study contemporary challenges that face rural areas in Britain. Through three consecutive funding rounds, 29 large collaborative research projects have been funded. Most projects are cross-institutional, involving two to six partner institutions and apply a variety of qualitative and quantitative methods, as well as modelling and simulations.

A programme-specific data policy[5] was developed for Relu, based on ESRC and NERC data policies, with the intent of advancing current practices in data management and sharing. The Relu data policy takes the view that:

- publicly-funded research data are a valuable, long term resource with usefulness both within and beyond the Relu programme
- all research data generated by funded Relu projects must be well managed by researchers throughout research
- all data must be made available for archiving at established data centres upon research projects finishing
- Relu funds support for data management throughout the lifetime of the Relu programme
- post-programme data management will be the responsibility of the research councils via existing data service providers
- a crossdisciplinary data support service (Relu-DSS), coordinated between the UK Data Archive and the Centre for Ecology and Hydrology (CEH) at Lancaster, through its Environmental Information Data Centre (EIDC) provides researchers with advice and support

In order for researchers to focus on their data management responsibilities throughout their research, award holders prepare a data management plan at the start of a project and implement it to ensure that data are well managed for the duration of the project. During research projects, advice and guidance on data management is provided to researchers by Relu-DSS through online guidance, training workshops and project visits to discuss data management and sharing issues with research teams.

3.2. Timescapes

Timescapes is a longitudinal qualitative study funded by the ESRC from 2007 to 2012. It is exploring how personal and family relationships develop and change over time. A consortium of five universities is conducting seven empirical projects that span the life course. In-depth interviews, oral narratives, photographs and other visual documents are being collected for the Timescapes archive. The archive is designed as a multi-media resource, giving equal consideration to textual, audio and visual data. The archive will also hold research outputs and an extensive array of documentation to enable the personal accounts of participants to be placed in historical, geographical and cultural contexts. Over 400 participants will be contributing data to the effort. The Timescapes study is distinctive in its explicit objective to simultaneously conduct and synchronise primary research, preservation and data reuse.

3.3. ESRC centres and programmes

The selected investments represent different (inter)disciplinary scopes within the wider social sciences domain. Most centres represent large cross-institutional collaborations, each receiving ESRC funding of £7

[5] www.data-archive.ac.uk/Relu/RELU%20Data%20Policy.pdf

to 12 million over 10 years and employing 20 to 50 research staff alongside centre-based administrative and support staff.

The ESRC Centre on Migration, Policy and Society[6] (COMPAS), 2003-2013, is based at the University of Oxford in the School of Anthropology and highly interdisciplinary in nature, embracing ten disciplines in its research interests.

The ESRC Centre for Economic and Social Aspects of Genomics[7] (Cesagen), 2002-2012, is part of the ESRC Genomics Network and based at Lancaster University and Cardiff University. It is a multidisciplinary centre addressing the social, economic and policy aspects of developments in genomics. The centre draws on social science and humanities research through ethnography and qualitative interviews, working in the natural and medical sciences alongside life scientists, clinicians, policy actors and key stakeholders in genomics.

The ESRC Centre for Social and Economic Research on Innovation in Genomics[8] (Innogen), 2002-2012, also part of the ESRC Genomics Network, studies the evolution of genomics and life sciences and their social and economic implications and is based at the University of Edinburgh and the Open University. Law, economics and social sciences researchers engage in research projects in the UK, Africa, China and India.

The ESRC Centre for Competition Policy[9] (CCP), 2004-2014, explores competition policy from the perspective of economics, law, business and political science and is based at the University of East Anglia.

The Centre for Research on Socio-Cultural Change[10] (CRESC), 2004-2014, is based at the University of Manchester and the Open University. Its mission is to analyse socio-cultural change in the context of socio-technical innovation, economic insecurity, and cultural diversity, with the intention to recognise different definitions and approaches to culture in its interface with processes of social change. CRESC research covers quantitative reuse of secondary data (e.g. longitudinal survey analysis) and qualitative research (ethnography, interviewing, audio and visual data).

The Third Sector Research Centre[11] (TSRC) c.2008 - 2013 is a collaboration across the Universities of Birmingham, Southampton, Kent and Middlesex. TSRC will bring together experts from a range of disciplines to develop a research programme that will lead to improved understanding of the key patterns, processes, and impacts of developments in the sector. This will strengthen the evidence base for policy towards the sector. TSRC will collaborate with, and offer a wide variety of services to support those working in and supporting the voluntary sector. In addition, TSRC co-ordinates the work of three Capacity Building Clusters which will support and enhance research capacity within the sector.

The New Dynamics of Ageing[12] (NDA) programme is a cross-council collaboration seeking to improve the quality of life of older people. It runs from 2005 until 2012 and is funded by the Economic and Social Research Council (ESRC), Engineering and Physical Sciences Research Council (EPSRC), Biotechnology and Biological Sciences Research Council (BBSRC), Medical Research Council (MRC) and the Arts and Humanities Research Council (AHRC). The emphasis is strongly on multidisciplinary and crossdisciplinary research. Funding is £22 million, with 43% contributed by the ESRC. The programme currently consists of 35 projects across 62 UK higher education institutes. NDA projects cover 47 disciplines, the main ones being psychology, sociology, health sciences and primary care.

The Teaching and Learning Research Programme[13] (TLRP) was a significant programmatic investment funded by ESRC from 2000 to 2009. It was made up of numerous investments including 4 research networks, 52 research projects, 5 associated projects, 2 career development associates, 5 research training fellowships, 25 ‘Meetings of Minds’ fellowships, over 20 thematic initiatives and 2 director’s fellowships. The majority of projects applied qualitative or mixed mode research methodologies, focusing on the roles and practices of teachers, learners, agencies and institutions across the entire lifecycle of learning. The specific Technology Enhanced Learning (TEL) projects began in 2007 and this phase continues to 2012. The UK Data Archive is a co-PI for one of these TEL projects.

3.4. Individual ESRC award holders

ESRC funds over 400 new research awards each year across the wider social sciences domain. Half of those awards plan to create new research data.

[6] www.compas.ox.ac.uk
[7] www.genomicsnetwork.ac.uk/cesagen/
[8] www.genomicsnetwork.ac.uk/innogen/
[9] www.uea.ac.uk/ccp
[10] www.cresc.ac.uk/
[11] www.tsrc.ac.uk/
[12] www.newdynamics.group.shef.ac.uk/
[13] www.tlrp.org/proj/index.html

4. Methodology

Data management practices and data management planning approaches were evaluated in a variety of ways.

For Relu, data management planning was evaluated by assessing data management plans from 36 projects (29 large projects, 4 pilot projects that created and archived data and 3 fellowships) for the quality of information provided by researchers in a plan. Data management evidence was compiled by the service manager from the interaction between the Relu Data Support Service and researchers for the period 2005-2010 and from information provided by researchers within the programme.

Timescapes evidence on data management was compiled by the Timescapes research archivist, the programme director and the technical officer, based on their experiences and on discussions with researchers.

Interviews with directors, selected researchers and administrative staff at six ESRC centres and in two programmes provided data management evidence for large ESRC investments. One programme has finished and was reviewed retrospectively, one centre is in its first 5-year funding cycle, the remainder are in their second 5-year funding phases.

Data management information gathered for individual ESRC award holders is based on:

- the ESDS query database
- data-related grant application information provided to ESDS by ESRC for all awarded grants
- data archiving difficulties encountered when data are offered to the UK Data Archive and reviewed by its Acquisitions Review Committee.

ESRC award holders are required under the ESRC Data Policy to offer their research data to the ESDS for archiving at the UK Data Archive when research projects finish. In the research grant application form, a section considers the plans for archiving data resulting from the grant award. This is information on whether projects plan to create new qualitative or quantitative data, with a brief description of the type of data, plans and costs to prepare data for archiving and any expected difficulties in archiving data. Such data-related information in grant applications was reviewed for the period Nov 2008 – July 2010, in particular to assess anticipated difficulties regarding data archiving.

ESDS provides online guidance and support to award holders and applicants on data management issues[14], as well as a helpdesk for queries. All incoming queries and their resolution are logged by ESDS in a queries database. Queries on data management topics were reviewed for the 20-month period Nov 2008 – July 2010.

When data are offered by ESRC award holders to the Archive, data offers are reviewed by the Acquisitions Review Committee (ARC) against specific criteria to decide whether or not a data collection can be preserved at the Archive. ARC decisions provide some information on data management issues that may prevent data archiving, e.g. ethical, legal and copyright issues, as well as lack of sufficient documentation to enable secondary use of data. ARC data were reviewed for the period Jan-Dec 2008.

[14] www.data/archive.ac.uk/create-manage

5. Current data management practices

Data management practices vary widely across the various investments evaluated. This may depend on the emphasis a programme or centre places upon data management and data sharing, or on practices applied by individual researchers.

For each type of investment evaluated, data management practices are organised according to the relevant topical areas:

- data management planning
- ethics, consent and confidentiality when managing and sharing research data
- data copyright and rights management
- contextualising, describing and documenting data
- data formats and software
- data storage, back-up and security
- roles and responsibilities of data management

5.1. Data management in the Relu programme

5.1.1. Data management planning

The Relu programme implemented data management planning at the start of the programme in 2004. At the start of each research project funded by Relu, award holders are required to prepare a data management plan, which is reviewed and signed off by the Relu Data Support Service.

In a data management plan[15] researchers describe:

- the need for access to existing data sources and any access limitations that may exist
- datasets planned to be produced by the research project
- planned quality assurance and back-up procedures for data
- plans for management and archiving of collected data
- expected difficulties in making data available for re-use (through data archiving) and measures to overcome such difficulties
- who holds copyright and intellectual property rights of the data
- data management roles and responsibilities within the research team

The Relu-Data Support Service reviews all submitted data management plans by:

- verifying that datasets planned to be produced correspond with the planned research activities as described in the research proposal
- ensuring that all relevant data management aspects have been considered, with meaningful information provided in the plan
- where difficulties are anticipated to make data available, consider whether solutions have been suggested
- ensure that a team member with data management responsibility is in place at each participating institution

After the review, the plan is either signed off, with data management advice given where needed, or researchers are asked to revise the plan where information may be insufficient.

The pro-active engagement of the Relu Data Support Service with researchers enables a review of the data management planning approach adopted by Relu.

5.1.1.1. Quality of data management plans

The information is presented according to the questions asked in a plan.

Datasets planned to be produced by the research project

Most plans contain sufficiently detailed lists of the various datasets planned to be produced. In a few cases information was vague and award holders were asked to provide better or more detailed information. For

[15] www.data-archive.ac.uk/Relu/DMP.doc

each dataset, the format or software in which data will be created or stored is specified and storage details are provided. Dependent on projects, storage may be solely on an institutional server or on a combination of server, PCs, institutional virtual environments and back-ups on movable media (e.g. CD, DVD).

During research projects, research activities may change and actual datasets produced at the end of a project can be different from those initially planned.

Planned quality assurance for data

All plans include good information on how data quality will be ensured. Measures include:

- Institutional quality assurance procedures, ISO standards
- Standard data collection protocols
- Standardised data recording (data entry sheets, validation rules in databases)
- Instrument calibration
- Recording metadata, labelling data
- Documenting methods and procedures
- Training researchers
- Pilot studies
- Double data entry
- Validation check, cross-checking
- Random checks
- Peer review of data
- Data record forms
- File naming standards

Planned data back-up procedures for data

Overall the information provided within this section is excellent. Most data management plans describe institutional data storage and back-up procedures that are in place. Most projects store data on institutional servers, which guarantees regular back-up and transfers the responsibility to institutional IT staff.

Some projects mention additional back-ups researchers plan to carry out (e.g. onto disks or hard drives, or by sharing copies of data between partner institutions) or state that the principal investigator will hold a master copy of all data, besides data held on partner servers.

Three data management plans failed to incorporate information for partner institutions, only listing procedures at the host institute.

Only four projects have specific data management staff allocated to the project, which have a role in overseeing data storage and back-up procedures (besides other responsibilities).

Expected difficulties in making data available for re-use and measures to overcome difficulties

Only 14 plans provide excellent information on this topic; in 10 plans the information is sufficient, whereas in 12 plans the information is vague or contains a simple statement that ‘no difficulties in making data available for secondary use are anticipated’. In six project where no problems to make data available for archiving were foreseen, researchers did not consider obtaining consent for data obtained through interviews or surveys to be shared, or collected data under unnecessarily strict confidentiality agreements. Data obtained through interviews / surveys could therefore not be archived due to confidentiality restrictions. Researchers may thus underestimate potential difficulties to archive and share data, especially for confidential, commercial or sensitive data.

Almost half the plans (17) state that data confidentiality, the inclusion of personal data in research data, and copyright of third party sources may limit the archiving of some research data, with overall valid reasons given. Confidentiality restrictions may be in place due to commercial confidentiality (e.g. business information for farms) or where interviewees are easily identifiable (e.g. elite interviews with public body stakeholders and policy makers). Copyright limitations exist mostly where research projects use licensed data sources within GIS systems, to create derived data or to model research scenarios. Use of Ordnance Survey data in GIS typically limits sharing even many derived data.

Only six plans then provide information on how such difficulties may be overcome by the researchers, e.g. by anonymising data, aggregating data, obtaining consent to share data, or discussing data archiving with owners of licensed data.

Data copyright and IPR

Copyright / IPR of the data is generally with the researchers. At times there is joint copyright through use of third party data.

Data management responsibilities within the research team

Most projects allocate data management responsibilities to various researchers within the research team – typically one person per partner institution or one person per work package.

A few projects allocate only one person with data management responsibility for the entire project. For cross-institutional projects, it is not clear how that is manageable.

Four projects have a dedicated data manager, database manager or project manager with overall data management responsibility.

5.1.1.2. Data Management Plan reviews by Relu-DSS

The Relu Data Support Service reviewed all data management plans and either provided additional data management advice and suggestions, or - if information was insufficient - asked for a revised plan to be submitted. The latter was requested for 6 projects. The main advice that needed to be given was for researchers to consider the potential restrictions of data confidentiality and copyright on the ability to archive and share research data and the importance of gaining consent and copyright clearance where needed for data to be archived.

From 2006 onwards, thanks to strong emphasis placed by the Relu director on data archiving and with additional research council funding, Relu-DSS could take a more pro-active and targeted approach in working directly with each Relu research team, providing information on various data management topics at various stages of the research cycle (before the research proposal application, at the start of a research project and in the course of the research project) and providing training to researchers. This improved the quality of data management plans and increased the attention paid by researchers to data management issues.

5.1.1.3. Researchers' views on data management planning

Various Relu principal investigators of finished research projects were asked during telephone interviews for their critical views on the data management planning approach adopted by Relu and whether they thought that data management plans served a purpose, improved research quality or helped researchers; or whether the plans were perceived as additional administrative procedures. They were also asked to which extent the plan was shared within the research team, whether data issues were discussed at team meetings or whether researchers took autonomous decisions on how to look after their own data.

Researchers perceived the benefit of completing a data management plan that it made them think about and discuss data issues within the project team. Researchers found face-to-face meetings with the Relu-DSS to discuss project-specific data issues essential, as this allowed them to ask specific questions pertinent to their research, without needing to read extensive online guidance. Also training workshops on relevant data management topics were considered essential. Researchers indicated that more practical advice was needed on how to plan data management and complete a plan, with clear examples and instructions.

5.1.2. Data management practices

Whilst data management planning was coordinated for the programme through the use of standardised forms and procedures, no standardised data management procedures were implemented at programme level, due to the very diverse nature of the Relu research, the methods used and the data created. Instead, the DSS provided good practice guidance for researchers on all relevant data management topics via:

- online guidance[16] - later developed into an extensive web resource on the UK Data Archive website[17]
- published data management guides: Guidance on Data Management (2006) and Managing and Sharing Data – a best practice guide for researchers (1st ed. 2008; 2nd ed. 2009)[18]
- workshops on a variety of data management topics, with research projects required by the programme director to ensure that researchers attended such workshops.

Data management practices and knowledge are typically not uniform across teams of researchers. Most Relu projects involve various collaborating partner institutions and large teams of researchers. The exchange of data management knowledge or sharing of information on data management practices between the award holder and the rest of the research team or between collaborating institutions is often lacking. Data management issues are typically not discussed much during team meetings and may not be considered as

[16] relu.data-archive.ac.uk/reluadvice.asp

[17] www.data-archive.ac.uk/create-manage

[18] www.data-archive.ac.uk/media/2894/managingsharing.pdf

priorities within the research procedures. Various projects encountered situations where the award holder may have ideas about how best to manage research data, but not share this information with junior researchers, or where partner institutions have inconsistent procedures in place.

Data management intentions listed in the data management plan may not be followed through by researchers. At no time were award holders or research teams encouraged by the research council to re-visit the data management plan or asked to report back on data management practices. Relu-DSS re-visited the plan with each research team during project visits (typically midway through research) and discussed any issues that could affect the data sharing potential of data, such as data confidentiality, copyright of data, data preparations and descriptions. There was no particular evaluation of whether planned data storage, back-up and quality assurance procedures were indeed put in place by researchers.

Through meetings and discussions with research teams at various stages of research and at the time of archiving research data at the UK data Archive, Relu-DSS found that key data management issues that need particular attention in order to ensure that data can be shared and have re-use potential beyond the primary research are:

5.1.2.1. Consent, confidentiality and ethics

Many researchers were found to be unsure how to enable the archiving and sharing of research data deemed confidential when data are obtained from research with human participants. Researchers may hesitate to consider sharing confidential research data and often draw up confidentiality agreements with participants to encourage their participation. Such agreements usually do not take data sharing into account. Confidential research data can, however, be shared ethically, if researchers openly discuss data sharing with research participants, and gain their consent for research data to be shared, with the proviso that where necessary data are anonymised or access to archived data can be regulated in order to protect participants' identities. Many of the early projects did not discuss data sharing with participants, resulting in data that could not be archived.

Relu-DSS placed particular emphasis on this data management aspect from 2007 onwards, by providing all Relu researchers with detailed information and guidance, by encouraging researchers to use UK data Archive template consent forms and by organising various training workshops for Relu researchers on this topic.

5.1.2.2. Data copyright

Researchers are not always clear about who owns copyright over data, especially where data are created from existing data sources. The use of licensed input data and integration with newly developed data may at times prohibit the archiving of datasets, subject to license holders agreeing to data sharing.

5.1.2.3. Describing, contextualising and documenting data

Researchers were at times unsure how best to document and describe data with sufficient detail so they can be understood, interpreted and therefore re-used by a wide research audience. Various projects used no uniform transcription style or template when transcribing interviews. Data labelling was at times unclear or incomplete and spreadsheet data of varying organizational quality.

5.1.3. Conclusion and lessons learnt

- Data management plans need to contain clear, practical and detailed information for researchers, pointing out issues that are relevant.
- Support services need to give clear and specific advice that researchers understand (not speak in 'data archiving' jargon).
- A pro-active and individual approach by data support services in providing topical data management information to researchers helps to ensure that researchers are well informed about relevant issues, that data management plans are well developed and that data sharing limitations are avoided.
- Researchers value opportunities to be able to discuss data management issues face-to-face with experts.
- Researchers may under-estimate difficulties to archive and share data, especially for confidential, commercial or sensitive data.
- Researchers may fail to gain consent for data sharing due to being unaware of the importance of discussing data sharing during consent procedures, thereby jeopardising the ability to share research data when research ends.
- Preparing a data management plan does not mean that data management is put into practice.

Research funders need to consider how they can ensure that plans are regularly reviewed by researchers and that data management practices are implemented, e.g. by requiring researchers to report on data management practices applied.

- Strategies are needed to ensure that all researchers within a research team are equally informed about and aware of data management issues.
- Researchers may struggle to find time to implement data management practices, especially in the latter stage of research.
- Mandating data archiving has a role to play, as very few researchers would voluntarily make data available for archiving or implement relevant data management practices. The data policy alongside strong encouragement by the programme director meant researchers had to implement certain data management strategies.
- Relu shows that a combination of a data policy, well-established data sharing infrastructures (UK Data Archive and EIDC), data support for researchers and active support of programme directors, results in increased availability of research data to the research community.

5.2. Data management at Timescapes

5.2.1. Data management planning

From its inception, creating archive-ready data was a primary goal for Timescapes. While no formal data management plans were used, steps were taken consistent with such planning to assure the creation of archive-ready data.

The design of the programme was to develop a scaled up study comprising a network of projects that would be integrated in terms of their methodological focus and their common approach and commitment to archiving. To that end the programme built in common research and fieldwork questions, common base data, and a commitment to use centralised protocols for the management and deposit of data in the archive[19] – protocols for transcription, consent, anonymisation and so on — and high standards for the production of sound recordings that would be digitised 'at source' i.e., in the field. There was a commitment in the bid to using a standardised analysis package (NIVO) and assigning key words to interviews to enhance searching. Timescapes set up a legal sub-contract agreement with the projects in the first year which set out their commitment to archiving and general principles for data sharing etc.

Some aspects of data management planning received more emphasis than others. Priorities were shaped by two realities. First, the primary goal in the first two years was to establish the archive infrastructure—all other objectives had to wait. Second, there was patchwork funding for the projects so they ran over different lengths of time which limited their longitudinal integration.

One of the problems faced was a highly practical one – the projects began in Feb 2007 but data protocols were not available until much later, through late staff appointments (the senior archivist in July 2007 and the technical officer not until 2008) and also part-time role of the archivist which has undoubtedly been a factor in available resources to get things done. In hindsight, the director and archivist agree that resources should have been used differently to ensure an equivalent full-time person – such as a part time archiving officer – to help support the senior archivist.

As a result, decisions were taken to focus on key aspects of data management that determined future usability of the data (e.g., consent for reuse) and handle other issues (e.g., inconsistent formats) at later dates.

5.2.2. Ethics, consent and confidentiality

In Timescapes, differences emerged between archiving staff and researchers around the use of written informed consent. The legal clarity of written consent was very attractive to the archive, and less so for researchers. For example, for some participants, the language of the form was too complicated or intimidating and a more informal, verbal approach worked better. Also researchers wanted to gain optimal levels of research participation (and some projects had very difficult recruiting profiles to fill) and build rapport with participants before discussing archiving. In latter stages of research, they argued, participants would have a much clearer understanding of what would go into to the archive and would be more able to grant authentic informed consent. The longitudinal method meant there would be repeat contacts with participants and numerous opportunities to seek consent later in the process, though there was always the risk of participants dropping out without giving consent for archiving.

The process of producing a model consent form acceptable to all the projects took a great deal of time, including drafting the form, holding consultations with representatives from each of the seven projects, incorporating their feedback into the form and keeping the wording of the Timescapes form aligned with changes and updates in UK Data Archive policies. The effort has yielded a standardised consent form[20] that covers areas of consent for participation, research outputs and data sharing and archiving.

The outcome has been positive, though not ideal: most but not all the teams have agreed to use the form or at least use it as a model for versions customised to specific audiences such as young people. Rather than written consent, some projects gain verbal consent and record and transcribe the consent agreement. And some projects defer asking for consent for archiving until later contacts with respondents. Given that frequent contacts are built into the research design, this is acceptable and probably very little data will be lost to archiving here. However, it is likely that in some cases, researchers' reluctance (and not participant views) may contribute to delays and extended negotiation over consent. *To date, over 95% of participants (over*

[19] www.timescapes.leeds.ac.uk/the-archive/

[20] www.timescapes.leeds.ac.uk/assets/files/timescapes/Timescapes%20CONSENT%20FORM%20Final%201%20May%2008.doc

150) asked have agreed to have their data archived. It may seem that a simple solution would be to mandate use of the form but this is neither easy nor optimal. It is politically difficult to impose this kind of mandate as the programme provides considerable autonomy – intellectual and financial – to the seven projects. Moreover, even if it were possible, it would be antithetical to the collegial culture of research and of this research team in particular.

A set of anonymisation guidelines[21] was drafted for projects to follow with instructions about which content to anonymise and formats for doing so. The system had to be easy to read, easy to teach to transcribers (some projects works with pools of transcribers and have little or no control over their quality),. This negotiation also took a very long time and in some cases, teams had to start using earlier versions of the guidelines that later changed (the earlier version attempted to use TEI tags for mark-up, but this was found to be too difficult for many of the transcribers). These “mid-stream” revisions created some difficulties: researchers found the changes hard to follow, some used the older (more labour-intensive) guidelines and felt they wasted valuable project resources. In sum, there is a far higher degree of diversity in anonymisation than is ideal. However, given other priorities, it was decided revisions could be made at a later time if resources permitted.

5.2.3. Data copyright and IPR

Most of the research materials have been created by the research team, and thus researchers have copyright. Exceptions are items such as participant-produced video, family photos, etc. In the deposit form that researchers sign, they are asked to have clear copyright to any materials they deposit. In addition, the consent form asks participants to transfer copyright to the researcher.

5.2.4. Documentation and metadata

Timescapes developed uniform documentation and metadata procedures to ensure coherence in data archiving. In building a schema for Timescapes, UK Data Archive standards were a starting point. The study description and catalogue record created at the UK Data Archive for each dataset follow the international standard for social science data, the Data Documentation Initiative (DDI) and are supported employing eXtensible Mark-up Language (XML). The study metadata is also mapped to the Dublin Core standard, and is compliant with the Open Archives Initiative (OAI) and Z39.50 for metadata harvesting and sharing. The Timescapes metadata schema[22] has been produced by the project's technical officer in consultation with Leeds Library staff members.

A set of baseline metadata elements has been created that will be used to support the creation of data sub-sets and to enable advanced searching. Additional descriptive metadata also includes very detailed information about the subject (personal characteristics, employment, education, living arrangements and so on). An XML schema is in development that will be used to capture this metadata. Importantly, this metadata is being collected to comply with variable definitions and coded responses used in the British Household Panel Study (BHPS). This attention to comparability assures that the extensive individual level quantitative data will be available for comparing cases within Timescapes (peoples' experiences over time).

It was initially intended to capture both the descriptive baseline metadata and the detailed metadata using the form capabilities of DigiTool during the deposit process. However due the complex nature of the metadata and the large number of elements it was not possible to create a usable form that we could reasonably expect a researcher to use. This led to the use of the Microsoft Infopath software to create the form-based input system that will produce an alternative form and interface to capture this metadata. The form will be used initially by the central Timescapes team with the potential to deploy the form on a web-based system so that the project researchers may use the form independently.

Additionally software is being developed to manage the process of capturing metadata using the InfoPath form and to collect the information that is required to support access control to the objects in DigiTool. This information describes whether data can be made public or is accessible only to defined groups of authorised individuals, and if there are any constraints on when the data can be made available (e.g., embargo).

A central focus of activity has been efforts to gain the engagement of researchers in collecting and providing metadata. One activity has been to provide a transcription template with a cover page of selected metadata fields. Researchers are also collecting much of the metadata described above in spreadsheets and databases that can be imported into DigiTool. Although there is some redundancy in metadata collection, these systems allow the researchers to work in the applications with which they are most comfortable. Additionally, the research team was actively involved in defining the expanded set of baseline descriptive metadata (e.g., age, gender, marital status, etc.) that will be collected about each participant.

---

[21] www.timescapes.leeds.ac.uk/assets/files/timescapes/Timescapes%20Anonymisation%20Guidelines%2018Aug08%20in%20use.doc

[22] www.timescapes.leeds.ac.uk/assets/files/timescapes/mdspec.xsd

The extent of documentation will be determined by each project. All researchers have been strongly encouraged to include diverse documentation, including field notes, researcher diaries, analytic memos, CAQDAS files, etc. In some cases, special conditions (exemplars or researcher-edited notes), are being prepared so that even if full documents are not available, examples will give subsequent researchers insights into the data collection procedures followed. One very rich source of material is a collection of papers about doing fieldwork for Timescapes (including one about the experience of preparing data for the archive): *WP2 — Conducting Qualitative Longitudinal Research: Fieldwork Experiences — January 2010*[23].

5.2.5. Data formats and transcription

The archive team has done a light touch monitoring of formats used. Most of the data arrives for ingest in expected formats (.doc, .tif). Advice is provided early in the project about preferred audio and video formats, attempting to ensure that preservation versions are created. In some cases this led to difficulties as teams sought cheaper, proprietary audio equipment when the archiving team recommended more expensive equipment capable of preservation-quality recordings for audio and video. Inevitably, there have been compromises on quality in some cases.

Early in the project, the archive team defined a set of transcription guidelines[24] accompanied by an exemplar which provided a consistent and standard format for transcriptions. This model specified a level of conformity required for digital preservation while not being unnecessarily prescriptive. The model also attempted to meet diverse requirements: easily explained to transcribers, usable for direct analysis by researchers, readable in hard copy, compatible with qualitative software applications, and suitable (or easily adapted) for web publication and preservation. There are several sections of the exemplar: a cover page that contains project, interviewee and interview details; a header that contains collection and interviewee identifier information and appears on every page; and the body of the transcription. If projects did not capture metadata in other places, the archive team used these transcription cover pages to capture essential metadata.

5.2.6. Storage, back-up and security

Projects manage local storage and back-up at their own institutions, as well as their own version control of data files. The issue of changes in longitudinal data versions has been noted, but no specific plan is in place as yet.

Projects manage local security. The archive team has provided procedures for secure data transfer (after some projects used email) such as CD by registered courier, File Transfer Protocol (FTP), and encrypted email. Such procedures are recommended for other file transfer (such as to transcribers) but no monitoring is done.

The Timescapes Archive is hosted at an institutional repository at Leeds, LUDOS. The LUDOS system is hosted on a fully redundant hardware architecture consisting of two primary servers with automatic failover between the servers. The storage is provided by a managed Storage Area Network that gives RAID Level 5 storage with replication across storage across the University campus. Backup and restore is provided by a fully automated tape based system.

Data preparation for ingest into the archive is conducted on local PC's using the Managed Storage System (mSTOR) provided by the University for the Storage of the data and the associated Submission Information Packages. mSTOR provides fully redundant and replicated storage across the campus network with full backup and restore on a nightly basis. Local version control, using SubVersion, is used to manage the day to day creation of the SIPs and to store intermediate files during the ingest process.

5.2.7. Rights management

Timescapes has a rights management framework under which materials are deposited, preserved and disseminated to users. The 'Depositor Licence' specifies the rights and responsibilities of the depositor and the Timescapes archive. The 'User Terms and Conditions' requires users of the data collection to respect confidentiality and not to disseminate any identifying information and has contractual force in law. In addition to these standard protections, depositors can impose bespoke conditions on uses of, and access to, data. For especially sensitive research data, additional restrictions may be imposed beyond the standard license. Data access authorisation may be required from the data depositor prior to release of the data or sensitive data may be placed under an embargo for a given period of time. This is decided on a case-by-case basis in dialogue between the Timescapes Archive team and the data depositor.

[23] www.timescapes.leeds.ac.uk/assets/files/WP2_final_Jan%202010.pdf

[24] www.timescapes.leeds.ac.uk/assets/files/timescapes/Transcription_guidelines_and_model_23July08_current.doc

This degree of access control is essential for ‘hard to anonymise’ and other sensitive data. Data vary in how amenable they are to these procedures. Visual data, for example, can be technically altered (e.g., digitally blurred faces), but at high cost to the integrity and quality of the data. Similarly, qualitative longitudinal data is challenging to anonymise, even textual transcripts, as the accumulation of personal, geographic, and other details over time becomes more disclosive than most cross-sectional interviews. The Timescapes Archive is using a system of four levels of access that will permit each file in the collection to be assigned a unique access group: public, registered users who sign a standard licence, users approved on a case-by-case basis, or embargoed.

5.2.8. Conclusion and lessons learnt

Although it is premature to do an extensive evaluation of the stakeholder model and the data management practices particular to it at this early stage, such an evaluation is being planned as one of the final programme outputs. However, some interim lessons and reflections are possible.

The total time burden of data issues clearly became too great on researchers – they were simultaneously being asked to: help with archive design, co-design template, conduct primary research and analysis, prepare data for archiving, and anticipate reuse. However, there did not seem to be another way to get researcher engagement; and proceeding without researcher participation might have produced a greater back-lash.

At the point of data preparation and deposit, the issue emerged of ‘exposure of researchers’, whereas until then protection of participants had been at the forefront. When data had to be sent to the archive, concerns about exposure arose:

- fear of revealing ‘other than textbook perfect’ ways of working.
- consequences fall far more heavily on early career researchers, who are usually the ones doing primary data collection and data preparation for archiving
- possible consequences for quality of data, e.g. in the form of sanitised accounts, self-censorship during interview (asking less probing questions)

Some of these concerns have been addressed by strengthening access control system, publishing and archiving parallel accounts from researchers and showcasing all the work and skill that goes into collecting and preparing data for archiving.

There is also a changing ethos in that datasets become part of research outputs and should be just as open to scrutiny as other kinds of outputs and should also be subject to the same strong ethics of building on existing scholarship and fieldwork without doing harm. This should be part of professional development and should not therefore become an excuse not to archive. There is still a long way to go to promote that ethos and Timescapes has been the test bed for it, but inevitably it can not happen all in one go.

Ethics issues and debates have been intense and involved many meetings, calls, etc. One example is the problem of internal confidentiality – concern about family members being able to get hold of data. It has been discussed and where data are very sensitive, they are being deposited only under condition of depositor approval for reuse.

Phasing of projects had huge implications. At the beginning, some projects were ready to do data management planning, whilst other projects had not even started hiring. As Timescapes concludes, it may be harder to get final information, such as detailed documentation, from the projects that finish early.

In sum, a preliminary assessment would have to acknowledge that even with dedicated project funding, strong leadership, and support services, significant challenges arose in creating archive-ready data. These challenges must be viewed in context: the data being collected by Timescapes represent hard-to-archive materials: sensitive, confidential, hard-to-anonymise, and so on. That said, the enthusiasm for data sharing from most of Timescapes’ participants is noteworthy and nearly all of the researchers strongly support the archive and are pleased their data are both safe and sharable.

5.3. Data management in the TLR programme

Despite the ESRC data policy to offer all data, there was no central steer from the TLRP director about sharing or archiving data; it was never a priority for the programme. There were the usual concerns about IPR and consent. The UK Data Archive did organise a TLRP Research Capacity Building network workshop on creating and managing data in 2003. Although the Programme did appreciate that data management was an area in which capacity building was needed, attendance by TLTP researchers was limited. Again, at the time, it was not seen to be a priority area for researchers who had their own established methods of undertaking team-based qualitative research.

In the TLRP, projects could use a Sakai Virtual Research Environment (VRE) to support collaboration activities. One project used the VRE effectively as a digital repository, uploading raw data, transcripts, and to support collaborative writing. Another project surveyed 10 research projects around the UK who were using the VRE to support their research and these were found to be using various VRE functions very differently. For example, some used their VRE as a lock-down repository with strict administrative control over uploading content, whereas another used it more for discussion space. Another project with confident web developers on the team developed a custom content management system, which was an early semantic web prototype. It has sophisticated fine grained access control so that the team and stakeholders could all access project materials. Local authorities, schools and researchers could see different versions of findings – which were accordingly anonymised - findings were aggregated in different ways. The evaluation of the project concluded that the best way of using VREs was to view them as a project lifetime tool and to provide them as a skeleton workspace and let the research groups configure them for their own needs. Longer term archiving or sharing of materials in the VRE was never considered.

To date 7 data collections from TLRP project have been archived at the UK Data Archive (3 still to be deposited) and 3 into UKDA-store, with a further 8 collections due to be self-archived into UKDA-store.

5.4. Data management in ESRC centres

Research centres, in the same way as individual grant holders, receive ESDS advice and guidance on creating and managing data. Various centres have had an ESDS visit when the centre starts, to discuss data requirements under the ESRC data policy.

The responsibility for data management in most of the research centres we studied lies generally with individual researchers and projects. Due to the nature of research, this responsibility generally lies with more junior researchers who typically collect, process and analyse research data. Also complying with the requirements of the ESRC data policy (i.e. making data available for archiving) is seen as a responsibility for individual researchers. Centres largely rely on researchers' own good data management practices. This is not necessarily true for other data-centric ESRC centres such as the Institute for Social and Economic Research and the Centre for Longitudinal Studies, whose centralised data management practices and set-up are highly sophisticated.

Research centre directors raised the problem of trying to integrate different research approaches into coherent data management policies at the centre level, with centres consisting of a multitude of researchers and often various institutions. All centres are established on a basis of innovative interdisciplinary research. These disciplines also bring different methodological and epistemological traditions and attitudes towards collecting, managing, sharing and re-using data. Changes of research staff, support staff and even directors may mean that data-related procedures may not always be followed through consistently throughout the lifetime of a centre.

Centres that receive a second 5-year ESRC grant have their data archiving obligations deferred until the end of the 2nd grant. This means that 10 years may pass before data need to be archived, which may make it difficult to keep track of data from early in the centre's research.

All research centres obtain external funding from other sources, besides the core ESRC grant. Whilst the ESRC mandates data archiving, other funders may have different or no data requirements. For various projects it is not always clear where the ESRC data policy applies and where not. For one centre, the total funding was still administered by ESRC, despite significant external funding being contributed.

There is a belief that the ESRC needs to provide data-related support and advice early in the lifecycle of a centre and of research projects. One centre stated that during contract negotiations with the ESRC it would be useful to involve the UK Data Archive to advise on data management issues, as such early planning of data management would ensure better returns for ESRC funded reusable data.

Whilst some interviewees expressed scepticism over archiving and reuse of research data, there was no

doubt over the usefulness of good data management. There was, however, a widespread unawareness of what data management consists of, what may be required of researcher, and which resources may be available to support researchers. Centres would like to be able to embed data management as a routine simple and straightforward process and to change researchers' existing practices with as much incentive and little coercion as possible.

5.4.1. Data management planning

Although no centre uses data management plans as such, centres address data management planning aspects in different ways and at different stages of the research process. Data management planning is to some extend applied as part of the ethical review process at the start of research. During ethical review researchers typically address ethical aspects of data management as well as anonymisation strategies, and possibly data storage and security. Some researchers and directors indicate that participating in the current DMP-ESRC project was a useful stimulus to think about aspects of data management planning they had yet to consider.

Data management is mostly not costed into centre grant proposal. Especially where data need to be anonymised or require considerable preparation prior to archiving, this is seen by some centres as a significant cost that may require additional funding.

In two centres, the centre management group (director and seniors) takes on a role of overseeing data management aspects as part of their research coordination role (e.g. implement data guidelines, policies, deal with ethical issues). These centres state a keenness to embed good data management practices within the centre, especially amongst PhD students and young researchers.

One centre has a researcher with a part-time data management role with responsibility for making researchers aware of data issues such as ethics, data security, copyright and anonymisation; being a point of contact for data management support researchers may need; and ensuring that researchers complete dataset log records as a basis for a centre-wide inventory of research data. The inventory ensures that the centre's research data are recorded and that ownership of datasets is established.

In some centres administrative staff fulfils a data management role and may develop centre policies, act as a point of contact for researcher queries on data management policies or procedures, coordinate e.g. interview transcription, maintain an intranet resource for data-related guidelines or keep logs of centre publications and dissemination activities.

Two centres stated a keen wish to establish a good management system for shared access to data for the purpose of archiving at the end of the centre's life, but also to facilitate sharing and awareness of data within the centre.

5.4.2. Ethics, consent and confidentiality

All centres have their research activities go through ethical review, mostly through an institutional Research Ethics Committee (REC) at the host institution of a centre, project or principal investigator. Ethical review is mostly carried out at project-level. One centre has gained centre-level ethics approval from the institutional REC, with ethical review of individual projects undertaken by the centre's research management committee, with regular discussion with and reporting to the REC as new research activities are initiated. Ethics policy is under review in two centres in response to the ESRC's reviewed Framework for Research Ethics 2010.

Sometimes researchers have to follow additional ethical review procedures, e.g. via National Health Service (NHS) research ethics committees. Such committees apply medical ethics standards and may require the destruction of data at the conclusion of a project. Even though the ESRC requirement is for anonymised data to be offered for archiving, it can still be difficult to reconcile this when securing ethical approval.

Some researchers perceive that it may be difficult to get permission for data archiving through university RECs in light of data protection and freedom of information concerns (although there were no real cases to support this perception). However, it may mean that researchers opt not to mention data archiving or gaining informant's consent for data sharing during ethical review, to avoid delays in ethical approval.

The heterogeneous character of research disciplines and practices across centres means that usually a diversity of informed consent procedures and agreements may be used. No centre had a standard consent form for use across its research projects, although there may be standardised wording for consent forms, or researchers may adapt exemplar forms provided by other researchers within a centre or from the UK Data Archive.

Consent may be gained by using a consent form or verbally, either at the start of research or negotiated throughout the research. E.g. for elite interviews the interviewee's approval of notes or a transcript may act as granting consent.

It is not yet a widespread practice amongst researchers to gain informed consent for research data to be archived at the end of a project. Only researchers who have previously offered data to the UK Data Archive and have therefore been made aware of the need for consent to be able to archive data (especially for qualitative data) do consider this during consent procedures. Researchers interviewing elite public or private sector respondents felt that discussing data archiving could introduce complications that could prevent gaining access to such persons, or could compromise the quality of interviews.

Confidentiality of research data within a centre may range from data that require very strict confidentiality conditions to data that can be made publicly available. For certain data only anonymised data may be shared across the centre, with raw data strictly protected by researchers

For qualitative research data, especially in the case of elite interviews, anonymisation is seen as very costly and often difficult to apply. There is also a lack of awareness as to what the conditions for re-using archived data entail, with at times a perception that data held by the UK Data Archive are made publicly available and researchers unaware of the access restrictions the Archive can implement to protect sensitive data.

5.4.3. Data copyright

Overall there exists little certainty over who exactly owns copyright over research data, the individual researcher, the centre or the institution. This may be complicated when researchers leave a centre and expect to take data along.

Researchers using secondary data are at times uncertain over data ownership. There was uncertainty as to what could be done with data obtained under licence when that licence or the project funding came to an end. There may equally be uncertainty over intellectual property rights when existing data are collated and significantly enhanced.

Many centres obtain external funding, besides core ESRC funding for the centre. There may be questions over what counts as core or externally funded research and how data ownership may be defined.

Some researchers believe that data being publicly available, particularly on the internet, are free of copyright, without considering the implications this could hold for archiving or even research use itself.

5.4.4. Describing, contextualising and documenting data

In general researchers produce their own contextual information and documentation for data, mainly for publication requirements. As a consequence, documentation may vary from project to project.

Some centres have facilities for conducting in-house data entry and transcription or facilitate the transcription of interviews centrally.

At a minimum centres keep a log of outputs for reporting to ESRC, but not always a log of datasets held or information about datasets.

At an early stage one centre had considered a central data inventory of all projects. In practice this did not work, as creating records for data collections required time and resources that were not available and researchers from different disciplines were unsure what information to include in a record.

5.4.5. Data formats and software

Researchers generally use familiar and default file formats. NVivo is widely used for analysis of qualitative data. Interviews and notes are generally transcribed into MS Word documents. Spreadsheet data are organised in MS Excel format and MS Access databases. Quantitative survey data is usually organised in SPSS and Stata formats. Little awareness exists for audio and visual data of the types of existing formats or the implications of using proprietary formats or lossy file types for long-term validity of data, particularly in the case of audio where digital voice recorders were used.

One centre has a guidance document which includes advice on open source and preservation-friendly formats based on UK Data Archive advice.

5.4.6. Data storage, back-up and security

Data storage, access regulations and security are generally left to researchers to organise. Researchers store data on university drive space, but may not be sure how this space is managed by their institution. Back-up of data held on servers is assumed to be carried out by the university, although many researchers also do ad-hoc back-up and storage using varying devices – CDs, email, flash drives, external hard-drives, and by storing original paper copies of measurements and questionnaires.

Data storage and sharing problems may exist when researchers are working remotely or in the case of cross-institutional collaborations. In centres that have no collaborative software it is difficult to gain access to

a shared drive away from the university. In one centre, the clustered nature of a centre's research interests means that there is strong exchange of data between individual researchers (e.g. for co-publishing). However this is not a formalised process of centralising or exchanging data at the centre level.

Some cross-institutional collaborative centres have adopted collaborative software platforms, such as Microsoft SharePoint and Microsoft Groove (now SharePoint WorkSpace). One centre uses SharePoint to store and share documentation and data, although in practice not all centre data are held there. Another centre is using MS Groove, which was found to be the best way to store quantitative data in a way that would allow the research team access across institutions and remotely. MS Groove stores anonymised data in one location (unanonymised data is stored offline) and researchers have access by invitation, with permissions granted by the principal investigator. The system is password protected with a user password (not a university password) and provides a form of version control.

University PCs are always password protected. Few data are encrypted. Staff outside of research teams typically has no access to data. This is due to concerns over the intellectual property rights (publications), possible misinterpretation and time and effort required to explain the dataset to users.

Keeping track of datasets may be made difficult by researcher mobility; for example, there have been cases in one centre where it has obtained datasets under licence or funded creation of datasets, but the data has been taken by researchers when they leave. However, this centre now has a system in place to manage the location and ownership of data.

Data may be transmitted between researchers and/or between home and institution via email, CD, flash drive; as well as through file sharing services (e.g. YouSendIt).

One centre asks each project to identify a 'data keeper' and to identify a back-up source. Another centre keeps master copies of files in a secure server area.

No researcher had clear understanding of requirements to, and strategy for, destroying data and documentation at the end of a project. Some researchers felt there may be institutional policies on retaining data, but could not recall specifics. Otherwise, the assumption was the researcher would retain data indefinitely. No researcher or centre had a destruction strategy in place.

5.4.7. Conclusion

- Many aspects of data management are not centralised or coordinated by a centre director, but are generally the responsibility of individual researchers. Some centres have centralised certain data management responsibilities under the auspices of a management committee or centre manager.
- Irrespective of whether or not researchers plan to share research data, good data management is considered to be an important issue, yet researchers are not always sure what may be needed.
- Researchers indicate that they want easy, practical and trustworthy solutions they can embed into research activities, rather than a range of guidelines or suggestions from which to choose.
- Current practice indicates that a gap exists between the start and conclusion of centres where data management and data policy obligations are not considered. At the start good intentions are put in place and data management may be planned as part of ethical review. At the end of a centre contract data are being prepared for archiving. In between, ongoing data management may be forgotten amongst the diversity of research activities, resulting in much extra time and effort needed to record and prepare research data for archiving. This is not made easier by the fact that the requirement to archive data may only be 10 years after the start of a centre's research.
- Data management needed to make data shareable is often not costed into grants by centres.
- A variety of sources of research funding brings with it diverse data sharing requirements
- Centres need solutions for easy file sharing, either for cross-institutional collaborations, or for remote working
- Some data management planning is done as part of ethical review.
- The perception that data sharing may be considered unfavourable by a REC and cause delays in approval means some researchers avoid considering data sharing.
- The lack of awareness about the role data archives can play in providing secure access and safeguards to archived data means that researchers believe that sensitive or confidential data cannot be archived and therefore do not consider data archiving during e.g. consent procedures.

5.5. Data management amongst individual award holders

The ESDS query database was analysed for a period of 20 months (November 2008 – July 2010). Out of 400 queries made to the acquisitions and data management services section, 174 related to data management issues. Amongst those the most prominent issues that researchers have questions about is: how to enable data archiving at the end of research when research data are deemed confidential or difficult to anonymise, especially in the case of qualitative research data (85 queries); how best to obtain consent for data archiving (37); which data formats to use for data archiving (34); how to cost data management into a research proposal (26), how to plan data management for a project (25) and how to regulate access to data (21). Other issues that researchers have queries about are copyright (19), transcription (10), how best to document data prior to archiving (9) and technical issues such as data storage and back-ups (4).

Most data management-related queries are thus related to research ethics, the confidentiality of data, gaining consent for data archiving, or how to anonymise data.

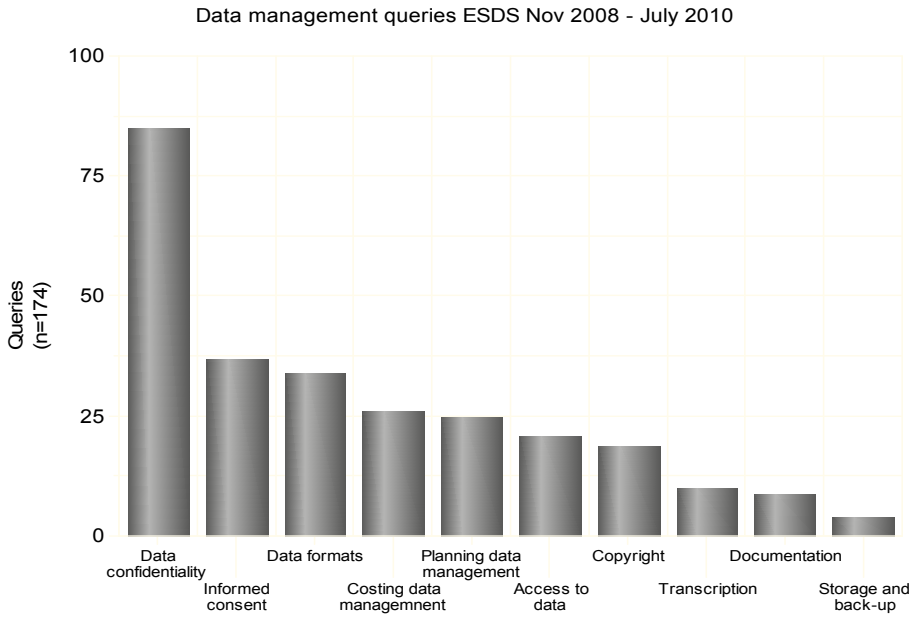

Fig. 1. Queries made by ESRC applicants or award holders to ESDS on various data management issues

In comparison, 800 new ESRC-funded research awards started within that same 20-month period. Of those, 362 project plan to create new research data – 118 qualitative data, 138 quantitative data and 106 both types of data. When asked whether any difficulties in archiving research data were anticipated, 49 % of award holders anticipate no problems in archiving data (where necessary gaining consent to archive or anonymising data), 19 % provide no information, 8 % state that data confidentiality or copyright issues would make it impossible to archive data and the remaining recognise that data confidentiality will need to be considered during consent procedures or by anonymising data.

Therefore, for each research award granted by ESRC, ESDS receives a query on data management issues.

When reviewing the offers of data form ESRC research grants to the UK Data Archive to analyse what may prohibit the archiving of data, of 297 offers of research data received by the UK data Archive in 2008, 11 % could not be accepted due to issues of data confidentiality and 1.5 % due to copyright issues.