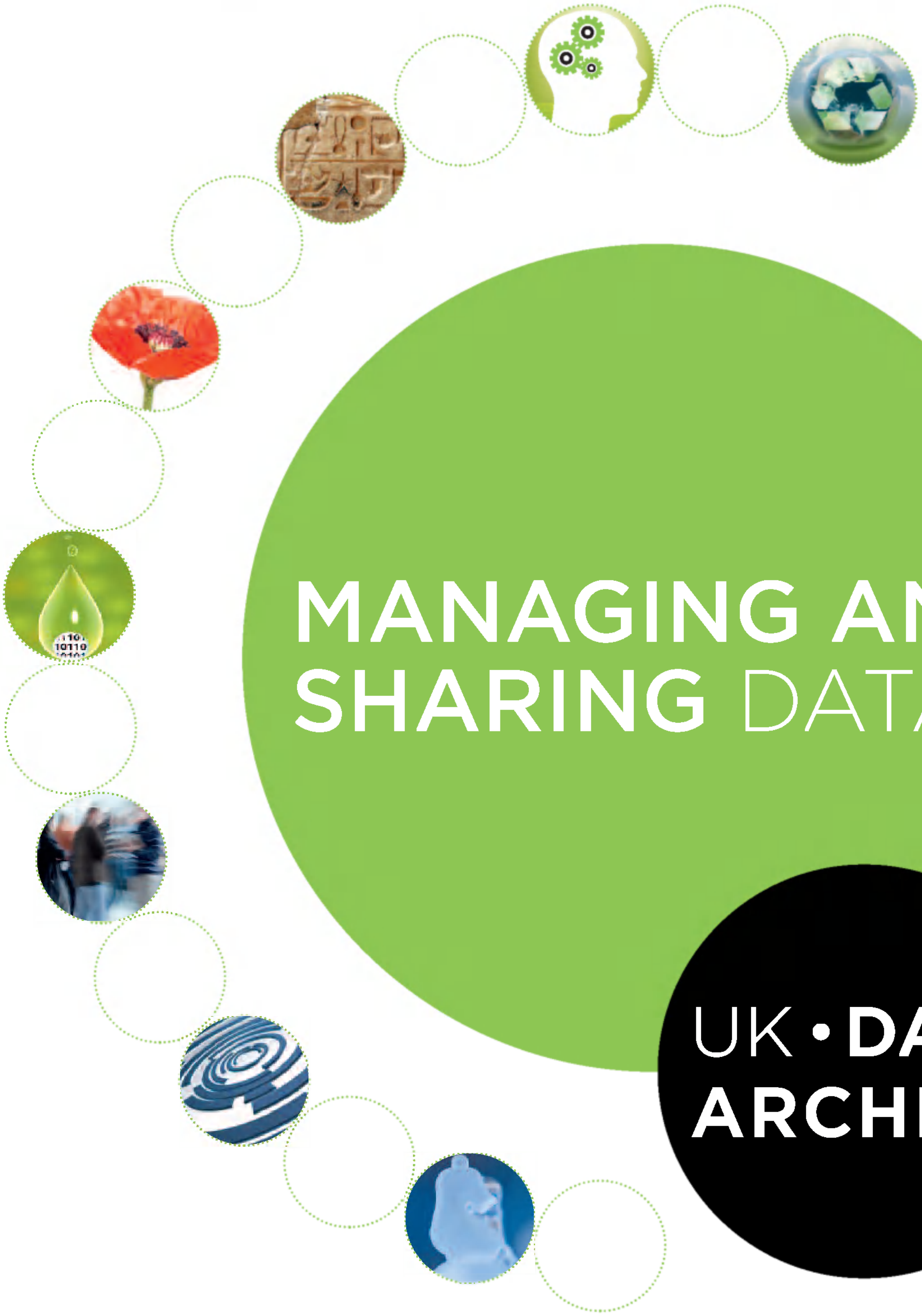


MANAGING AND SHARING DATA

UK • DATA ARCHIVE

BEST PRACTICE FOR RESEARCHERS

MAY 2011

First published 2009
Third edition, fully revised, 2011

Authors: Veerle Van den Eynden, Louise Corti, Matthew Woollard, Libby Bishop and Laurence Horton.

In memory of Dr. Alasdair Crockett who co-authored our first published Guide on Data Management in 2006.

All online literature references available 16 March 2011.

Published by:
UK Data Archive
University of Essex
Wivenhoe Park
Colchester
Essex
CO4 3SQ

ISBN: 1-904059-78-3

Designed and printed by
Print Essex at the University of Essex

50% recycled material

CONTENTS

ENDORSEMENTS

The scientific process is enhanced by managing and sharing research data. Good data management practice allows reliable verification of results and permits new and innovative research built on existing information. This is important if the full value of public investment in research is to be realised.

Since the first edition of this excellent guide, these principles have become even more widely endorsed and are increasingly supported by the mandates of research funders who are concerned to see the greatest possible return on investment both in terms of the quality of research outputs and the re-use of research data. In the USA, the National Science Foundation now requires grant applications to include a data management plan. In the UK: a joint Research Councils UK (RCUK) statement on research data is in preparation; the Engineering and Physical Sciences Research Council (EPSRC) is preparing a policy framework for management and access to research data; the Medical Research Council (MRC) is developing new, more comprehensive, guidelines to govern management and sharing of research data; and the Economic and Social Research Council (ESRC) has revised its longer standing research data policy and guidelines.

Given this rapidly changing environment, the Joint Information Systems Committee (JISC) considers it a priority to support researchers in responding to these requirements and to promote good data management and sharing for the benefit of UK Higher Education and Research. JISC funds the Digital Curation Centre, which provides internationally recognised expertise in this area, as well as support and guidance for UK Higher Education. Furthermore, through the Managing Research Data (MRD) programme, launched in October 2009, JISC has helped higher education institutions plan their data management practice, pilot the development of essential data management infrastructure, improve methods for citing data and linking to publications; and funded projects which are developing training materials in research data management for postgraduate students.

The UK Data Archive has been an important and active partner and stakeholder in these initiatives. Indeed, many of the revisions and additions that have occasioned the new edition of this guide were developed through the UK Data Archive's work in the JISC MRD Project 'Data Management Planning for ESRC Research Data-Rich Investments'. As a result, 'Managing and Sharing Data: best practice for researchers' has been made even more targeted and practical. I am convinced that researchers will find it an invaluable publication.

Simon Hodson, Joint Information Systems Committee (JISC)

Data are the main asset of economic and social research – the basis for research and also the ultimate product of research. As such, the importance of research data quality and provenance is paramount, particularly when data sharing and re-use is becoming increasingly important within and across disciplines. As a leading UK agency in funding economic and social research, the ESRC has been strongly promoting the culture of sharing the results and data of its funded research. ESRC considers that effective data management is an essential precondition for generating high quality data, making them suitable for secondary scientific research. It is therefore expected that research data generated by ESRC-funded research must be well-managed to enable data to be exploited to the maximum potential for further research. This guide can help researchers do so.

Jeremy Neathey, Director of Training and Resources, Economic and Social Research Council (ESRC)

The Rural Economy and Land Use (Relu) Programme, which undertakes interdisciplinary research between social and natural sciences, has brought together research communities with different cultures and practices of data management and sharing. It has been at the forefront of cross-disciplinary data management and sharing by developing a proactive data management policy and the first cross-council Data Support Service. It adopted a systematic approach from the start to data management, drawing on best practice from its constituent Research Councils. The Data Support Service helped to inculcate researchers from across different research communities into good data management practices and planning and its close engagement with researchers laid an important basis for this best practice guide. It also orchestrates the linked archiving of interdisciplinary datasets across data archives, accessed through a knowledge portal that for the first time for the Research Councils brings together data and other research outputs and publications. Relu shows that a combination of a programme level strategy, well-established data sharing infrastructures and active data support for researchers, result in increased availability of data to the research community.

Philip Lowe, Director, Rural Economy and Land Use (Relu) Programme

Also endorsed by:

Archaeology Data Service (ADS)
Biotechnology and Biological Sciences Research Council (BBSRC)
British Library (BL)
Digital Curation Centre (DCC)
History Data Service (HDS)
LSE Research Laboratory (RLAB)
National Environment Research Council (NERC)
Research Information Network (RIN)
University of Essex

FOREWORD

THE DATA MANAGEMENT AND SHARING ENVIRONMENT HAS EVOLVED SINCE THE PREVIOUS EDITION OF THIS GUIDE. RESEARCH FUNDERS PLACE THE SHARING OF RESEARCH DATA EVER HIGHER AMONGST THEIR PRIORITIES, REFLECTED IN THEIR DATA SHARING POLICIES AND THE DEMAND FOR DATA MANAGEMENT PLANS IN RESEARCH APPLICATIONS.

Initiatives by higher education institutions and supporting agencies follow suit and focus on developing data sharing infrastructures; supporting researchers to manage and share data through tools, practical guidance and training; and enabling data citation and linking data with publications to increase visibility and accessibility of data and the research itself.

Whilst good data management is fundamental for high quality research data and therefore research excellence, it is crucial for facilitating data sharing and ensuring the sustainability and accessibility of data in the long-term and therefore their re-use for future science.

If research data are well organised, documented, preserved and accessible, and their accuracy and validity is controlled at all times, the result is high quality data, efficient research, findings based on solid evidence and the saving of time and resources. Researchers themselves benefit greatly from good data management. It should be planned before research starts and may not necessarily incur much additional time or costs if it is engrained in standard research practice,

The responsibility for data management lies primarily with researchers, but institutions and organisations can provide a supporting framework of guidance, tools and infrastructure and support staff can help with many facets of data management. Establishing the roles and responsibilities of all parties involved is key to successful data management and sharing.

The information provided in this guide is designed to help researchers and data managers, across a wide range of research disciplines and research environments, produce highest quality research data with the greatest potential for long-term use. Expertise for producing this guidance comes from the Data Support Service of the interdisciplinary Rural Economy and Land Use (Relu) Programme, the Economic and Social Data Service (ESDS) and the Data Management Planning for ESRC Research Data-rich Investments project (DMP-ESRC) project. All these initiatives involve close liaising with numerous researchers spanning the natural and social sciences and humanities.

The UK Data Archive thanks data management experts from the National Environmental Research Council (NERC), the NERC Environmental Bioinformatics Centre (NEBC), the Environmental Information Data Centre (EIDC) of the Centre for Ecology and Hydrology (CEH), the British Library (BL), the Research Information Network (RIN), the Archaeology Data Service (ADS), the History Data Service (HDS), the London School of Economics Research Laboratory, the Wellcome Trust, the Digital Curation Centre, Naomi Korn Copyright Consultancy and the Commission for Rural Communities for reviewing the guidance and providing valuable comments and case studies.

This third edition has been funded by the Joint Information Systems Committee (JISC), the Rural Economy and Land Use Programme and the UK Data Archive.

This printed guide is complemented by detailed and practical online information, available from the UK Data Archive web site.

The UK Data Archive also provides training and workshops on data management and sharing, including advice via datasharing@data-archive.ac.uk.

www.data-archive.ac.uk/create-manage

SHARING YOUR DATA – WHY AND HOW

DATA SHARING DRIVES SCIENCE FORWARD

DATA CREATED FROM RESEARCH ARE VALUABLE RESOURCES THAT CAN BE USED AND RE-USED FOR FUTURE SCIENTIFIC AND EDUCATIONAL PURPOSES. SHARING DATA FACILITATES NEW SCIENTIFIC INQUIRY, AVOIDS DUPLICATE DATA COLLECTION AND PROVIDES RICH REAL-LIFE RESOURCES FOR EDUCATION AND TRAINING.

WHY SHARE RESEARCH DATA

Research data are a valuable resource, usually requiring much time and money to be produced. Many data have a significant value beyond usage for the original research.

Sharing research data:

- encourages scientific enquiry and debate
- promotes innovation and potential new data uses
- leads to new collaborations between data users and data creators
- maximises transparency and accountability
- enables scrutiny of research findings
- encourages the improvement and validation of research methods
- reduces the cost of duplicating data collection
- increases the impact and visibility of research
- promotes the research that created the data and its outcomes
- can provide a direct credit to the researcher as a research output in its own right
- provides important resources for education and training

The ease with which digital data can be stored, disseminated and made easily accessible online to users means that many institutions are keen to share research data to increase the impact and visibility of their research.

RESEARCH FUNDERS

Public funders of research increasingly follow guidance from the Organisation for Economic Co-operation and Development (OECD) that publicly funded research data should as far as possible be openly available to the scientific community.[1] Many funders have adopted research data sharing policies and mandate or encourage researchers to share data and outputs. Data sharing policies tend to allow researchers exclusive data use for a reasonable time period to publish the results of the data.

In the UK, funding bodies such as the Economic and Social Research Council (ESRC), the Natural Environment Research Council (NERC) and the British Academy mandate researchers to offer all research data generated during research grants to designated data centres – the UK Data Archive and NERC data centres. The Biotechnology and Biological Sciences Research Council (BBSRC), the Medical Research Council (MRC) and the Wellcome Trust have similar data policies in place which encourage researchers to share their research data in a timely manner, with as few restrictions as possible. Research programmes funded by multiple agencies, such as the cross-disciplinary Rural Economy and Land Use Programme, may also mandate data sharing.[2]

Research councils equally fund data infrastructures and data support services to facilitate data sharing within their subject domain, e.g. NERC data centres, UK Data Archive, Economic and Social Data Service and MRC Data Support Service.

In addition, BBSRC, ESRC, MRC, NERC and the Wellcome Trust require data managing and sharing plans as part of grant applications for projects generating new research data. This ensures that researchers plan how to look after data during and after research to optimise data sharing.

JOURNALS

Journals increasingly require data that form the basis for publications to be shared or deposited within an accessible database or repository. Similarly, initiatives such as DataCite, a registry assigning unique digital object identifiers (DOIs) to research data, help scientists to make their data citable, traceable and findable, so that research data, as well as publications based on those data, form part of a researcher's scientific output.

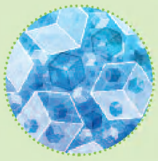
CASE STUDY

JOURNALS AND DATA SHARING

The Publishing Network for Geoscientific and Environmental Data (PANGAEA) is an open access repository for various journals.[3] By giving each deposited dataset a DOI, a deposited dataset acquires a unique and persistent identifier, and the underlying data can be directly connected to the corresponding article. For example, PANGAEA and the publisher Elsevier have reciprocal linking between research data deposited with PANGAEA and corresponding articles in Elsevier journals.

'*Nature* journals' have a policy that requires authors to make data and materials available to readers, as a condition of publication, preferably via public repositories.[4] Appropriate discipline-specific repositories are suggested. Specifications regarding data standards, compliance or formats may also be provided.

For example, for research on small molecule crystal structures, authors should submit the data and materials to the Cambridge Structural Database (CSD) as a Crystallographic Information File, a standard file structure for the archiving and distribution of crystallographic information. After publication of a manuscript, deposited structures are included in the CSD, from where bona fide researchers can retrieve them for free. CSD has similar deposition agreements with many other journals.



HOW TO SHARE YOUR DATA

There are various ways to share research data, including:

- depositing them with a specialist data centre, data archive or data bank
- submitting them to a journal to support a publication
- depositing them in an institutional repository
- making them available online via a project or institutional website
- making them available informally between researchers on a peer-to-peer basis

Each of these ways of sharing data has advantages and disadvantages: data centres may not be able to accept all data submitted to them; institutional repositories may not be able to afford long-term maintenance of data or support for more complex research data; and websites are often ephemeral with little sustainability.

Approaches to data sharing may vary according to research environments and disciplines, due to the varying nature of data types and their characteristics.

The advantages of depositing data with a specialist data centre include:

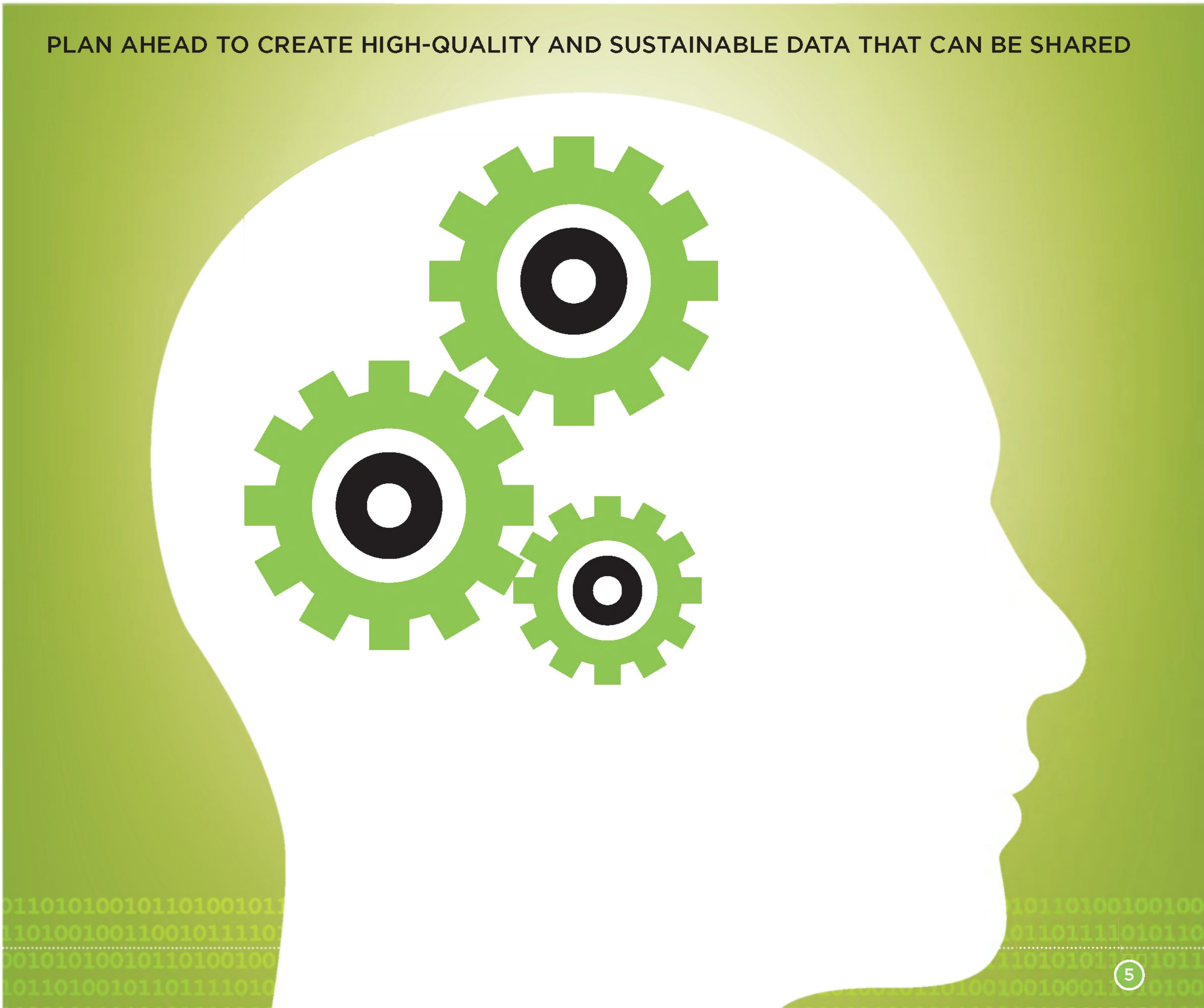
- assurance that data meet set quality standards
- long-term preservation of data in standardised accessible data formats, converting formats when needed due to software upgrades or changes
- safe-keeping of data in a secure environment with the ability to control access where required
- regular data back-ups
- online resource discovery of data through data catalogues
- access to data in popular formats
- licensing arrangements to acknowledge data rights
- standardised citation mechanism to acknowledge data ownership
- promotion of data to many users
- monitoring of the secondary usage of data
- management of access to data and user queries on behalf of the data owner

Data centres, like any traditional archive, usually apply certain criteria to evaluate and select data for preservation.

EXAMPLES OF RESEARCH DATA CENTRES

Antarctic Environmental Data Centre
Archaeology Data Service
Biomedical Informatics Research Network Data Repository
British Atmospheric Data Centre
British Library Sound Archive
British Oceanographic Data Centre
Cambridge Crystallographic Data Centre
Economic and Social Data Service
Environmental Information Data Centre
European Bioinformatics Institute
Geospatial Repository for Academic Deposit and Extraction
History Data Service
Infrared Space Observatory
National Biodiversity Network
National Geoscience Data Centre
NERC Earth Observation Data Centre
NERC Environmental Bioinformatics Centre
Petrological Database of the Ocean Floor
Publishing Network for Geoscientific and Environmental Data (PANGAEA)
Scran
The Oxford Text Archive
UK Data Archive
UK Solar System Data Centre
Visual Arts Data Service

DATA MANAGEMENT PLANNING



PLAN AHEAD TO CREATE HIGH-QUALITY AND SUSTAINABLE DATA THAT CAN BE SHARED

A DATA MANAGEMENT AND SHARING PLAN HELPS RESEARCHERS CONSIDER, WHEN RESEARCH IS BEING DESIGNED AND PLANNED, HOW DATA WILL BE MANAGED DURING THE RESEARCH PROCESS AND SHARED AFTERWARDS WITH THE WIDER RESEARCH COMMUNITY.

In the last few years many UK and international research funders have introduced a requirement within their data policies for data management and sharing plans to be part of research grant applications; in the UK this is the case for BBSRC[5], ESRC[6], MRC[7], NERC[8] and the Wellcome Trust[9].

Whilst each funder specifies particular requirements for the content of a plan, common areas are:

- which data will be generated during research
- metadata, standards and quality assurance measures
- plans for sharing data
- ethical and legal issues or restrictions on data sharing
- copyright and intellectual property rights of data
- data storage and back-up measures
- data management roles and responsibilities
- costing or resources needed

Best practice advice can be found on all these topics in this Guide.

It is crucial when developing a data management plan for researchers to critically assess what they can do to share their research data, what might limit or prohibit data sharing and whether any steps can be taken to remove such limitations.

A data management plan should not be thought of as a simple administrative task for which standardised text can be pasted in from model templates, with little intention to implement the planned data management measures early on, or without considering what is really needed to enable data sharing.

Key issues for data management planning in research:

- know your legal, ethical and other obligations regarding research data, towards research participants, colleagues, research funders and institutions
- implement good practices in a consistent manner
- assign roles and responsibilities to relevant parties in the research
- design data management according to the needs and purpose of research
- incorporate data management measures as an integral part of your research cycle
- implement and review data management throughout research as part of research progression and review

A big limitation for data sharing is time constraints on researchers, especially towards the end of research, when publications and continued research funding place high pressure on a researcher's time.

At that moment in the research cycle, the cost of implementing late data management and sharing measures can be prohibitively high. Implementing data management measures during the planning and development stages of research will avoid later panic and frustration. Many aspects of data management can be embedded in everyday aspects of research co-ordination and management and in research procedures.

Good data management does not end with planning. It is critical that measures are put into practice in such a way that issues are addressed when needed before mere inconveniences become insurmountable obstacles. Researchers who have developed data management and sharing plans found it beneficial to have thought about and discussed data issues within the research team.[10]

CASE STUDY

DATA MANAGEMENT PLANS

The Rural Economy and Land Use (Relu) Programme has been at the forefront of implementing data management planning for research projects since 2004.

Drawing on best practice in data management and sharing across three research councils (ESRC, NERC and BBSRC), Relu requires that all funded projects develop and implement a Data Management Plan to ensure that data are well managed throughout the duration of a research project.[11] In a data management plan researchers describe:

- the need for access to existing data sources
- data to be produced by the research project
- quality assurance and back-up procedures
- plans for management and archiving of collected data
- expected difficulties in making data available for secondary research and measures to overcome such difficulties
- who holds copyright and Intellectual Property Rights of the data
- who has data management responsibility roles within the research team

Example plans of past projects are available.[12]

ROLES AND RESPONSIBILITIES

Data management is not always simply the responsibility of the researcher; various parties are involved in the research process and may play a role in ensuring good quality data and reducing any limitations on data sharing. It is crucial that roles and responsibilities are assigned and not simply presumed. For collaborative research, assigning roles and responsibilities across partners is important.

People involved in data management and sharing may include:

- project director designing research
- research staff collecting, processing and analysing data
- external contractors involved in data collection, data entry, processing or analysis
- support staff managing and administering research and research funding
- institutional IT services staff providing data storage and back-up services
- external data centres or web services archives who facilitate data sharing

COSTING DATA MANAGEMENT

To cost research data management in advance of research starting, e.g. for inclusion in a data management plan or in preparation for a funding application, two approaches can be taken.

- Either all data-related activities and resources for the entire data cycle – from data creation, through processing, analyses and storage to sharing and preservation – can be priced, to calculate the total cost of data generation, data sharing and preservation.

- Or one can cost the additional expenses – above standard research procedures and practices – that are needed to make research data shareable beyond the primary research team. This can be calculated by first listing all data management activities and steps required to make data shareable (e.g. based on a data management checklist), then pricing each activity in terms of people's time or physical resources needed such as hardware or software.

The UK Data Archive has developed a simple tool that can be used for the latter option of costing data management.[13]

CASE STUDY

CREATING A DATA MANAGEMENT PLAN

In April 2010, the Digital Curation Centre (DCC) launched DMP Online, a web-based tool designed to help researchers and other data stakeholders develop data management plans according to the requirements of major research funders.[14]

Using the tool researchers can create, store and update multiple versions of a data management plan at the grant application stage and during the research cycle. Plans can be customised and exported in various formats. Funder- and institution-specific best practice guidance is available.

The tool combines the DCC's comprehensive 'Checklist for a Data Management Plan' with an analysis of research funder requirements. The DCC is working with partner organisations to include domain- and subject-specific guidance in the tool.

ROLES AND RESPONSIBILITIES

SomnIA: Sleep in Ageing, a project co-ordinated by the University of Surrey as part of the New Dynamics of Ageing Programme, employed a research officer with dedicated data management responsibilities.[15]

This multidisciplinary project with research teams based at four UK universities, created a wide range of data from specialised actigraphy and light sensor measurements, self-completion surveys, randomised control clinical trials and qualitative interviews.

Besides undertaking research on the project, the research officer co-ordinated the management of research data created by each work package via SharePoint Workspace 2010. This allowed controlled permission-levels of access to data and documentation and provided encryption and automated version control.

The officer was also responsible for operating a daily back-up of all data and documentation to an off-site server. Individual researchers themselves were responsible for other data management tasks.

DOCUMENTING YOUR DATA

MAKE DATA CLEAR TO UNDERSTAND AND EASY TO USE

A CRUCIAL PART OF MAKING DATA USER-FRIENDLY, SHAREABLE AND WITH LONG-LASTING USABILITY IS TO ENSURE THEY CAN BE UNDERSTOOD AND INTERPRETED BY ANY USER. THIS REQUIRES CLEAR AND DETAILED DATA DESCRIPTION, ANNOTATION AND CONTEXTUAL INFORMATION.

DATA DOCUMENTATION

Data documentation explains how data were created or digitised, what data mean, what their content and structure are and any data manipulations that may have taken place. Documenting data should be considered best practice when creating, organising and managing data and is important for data preservation. Whenever data are used sufficient contextual information is required to make sense of that data.

Good data documentation includes information on:

- the context of data collection: project history, aim, objectives and hypotheses
- data collection methods: sampling, data collection process, instruments used, hardware and software used, scale and resolution, temporal and geographic coverage and secondary data sources used
- dataset structure of data files, study cases, relationships between files
- data validation, checking, proofing, cleaning and quality assurance procedures carried out
- changes made to data over time since their original creation and identification of different versions of data files
- information on access and use conditions or data confidentiality

At the data-level, documentation may include:

- names, labels and descriptions for variables, records and their values
- explanation or definition of codes and classification schemes used
- definitions of specialist terminology or acronyms used
- codes of, and reasons for, missing values
- derived data created after collection, with code, algorithm or command file
- weighting and grossing variables created
- data listing of annotations for cases, individuals or items

Data-level descriptions can be embedded within a data file itself. Many data analysis software packages have facilities for data annotation and description, as variable attributes (labels, codes, data type, missing values), data type definitions, table relationships, etc.

Other documentation may be contained in publications, final reports, working papers and lab books or created as a data collection user guide.

CASE STUDY

DOCUMENTING DATA IN NVIVO

Researchers using qualitative data analysis packages, such as NVivo 9, to analyse data can use a range of the software's features to describe and document data. Such descriptions both help during analysis and result in essential documentation when data are shared, as they can be exported from the project file alongside data at the end of research.

Researchers can create classifications for persons (e.g. interviewees), data sources (e.g. interviews) and coding. Classifications can contain attributes such as the demographic characteristics of interviewees, pseudonyms used, and the date, time and place of interview. If researchers create generic classifications beforehand, attributes can be standardised across all sources or persons throughout the project. Existing template and pre-populated classification sheets can be imported into NVivo.

Documentation files like the methodology description, project plan, interview guidelines and consent form templates can be imported into the NVivo project file and stored in a 'documentation' folder in the Memos folder or linked from NVivo 9 externally. Additional documentation about analyses or data manipulations can be created in NVivo as memos.

A date- and time-stamped project event log can record all project events carried out during the NVivo project cycle.

Additional descriptions can be added to all objects created in, or imported to, the project file such as the project file itself, data, documents, memos, nodes and classifications.

All textual documentation compiled during the NVivo project cycle can later be exported as textual files; classifications and event logs can be exported as spreadsheets to document preserved data collections. The structure of the project objects can be exported in groups or individually. Summary information about the project as a whole or groups of objects can be exported via project summary extract reports as a text, MS Excel or XML file.

CASE STUDY

DATA DOCUMENTATION

Online documentation for a data collection in the UK Data Archive catalogue can include project instructions, questionnaires, technical reports, and user guides.

FORMAT	NAME	SIZE IN KB	DESCRIPTION
PDF	6713dataset_documentation.pdf	1403	Dataset Documentation (variable list, derived variables, variables used in report tables)
PDF	6713project_instructions.pdf	1998	Project instructions (interviewer, nurse and coding and editing instructions)
PDF	6713questionnaires.pdf	2010	Questionnaires (CAPI and self-completion questionnaires and showcards)
PDF	6713technical_report.pdf	6056	Technical Report
PDF	6713userguide.pdf	256	User Guide
PDF	UKDA_Study_6713_Information.htm	19	Study information and citation

METADATA

In the context of data management, metadata are a subset of core standardised and structured data documentation that explains the origin, purpose, time reference, geographic location, creator, access conditions and terms of use of a data collection. Metadata are typically used:

- for resource discovery, providing searchable information that helps users to easily find existing data
- as a bibliographic record for citation

Metadata for online data catalogues or discovery portals are often structured to international standards or schemes such as Dublin Core, ISO 19115 for geographic information, Data Documentation Initiative (DDI), Metadata Encoding and Transmission Standard (METS) and General International Standard Archival Description (ISAD(G)).

The use of standardised records in eXtensible Mark-up Language (XML) brings key data documentation together into a single document, creating rich and structured content about the data. Metadata can be viewed with web browsers, can be used for extract and analysis engines and can enable field-specific searching. Disparate catalogues can be shared and interactive browsing tools can be applied. In addition, metadata can be harvested for data sharing through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Researchers typically create metadata records for their data by completing a data centre's data deposit form or metadata editor, or by using a metadata creation tool, like Go-Geo! GeoDoc[16] or the UK Location Metadata Editor[17]. Providing detailed and meaningful dataset titles, descriptions, keywords and other information enables data centres to create rich resource-discovery metadata for archived data collections.

Data centres accompany each dataset with a bibliographic citation that users are required to cite in research outputs to reference and acknowledge accurately the data source used. A citation gives credit to the data source and distributor and identifies data sources for validation.

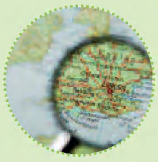
CASE STUDY

CREATING METADATA

Go-Geo! GeoDoc is an online metadata creation tool that researchers can use to create metadata records for spatial data.

The metadata records are compliant with the UK GEMINI standard, the European INSPIRE directive and the international ISO 19115 geospatial metadata standard.[18]

Researchers can create metadata records, export them in a variety of formats - UK AGMAP 2, UK GEMINI 2, ISO 19115, INSPIRE, Data Federal Geographic Data Committee, Dublin Core and Data Documentation Initiative (DDI) - or publish them in the Go-Geo! portal, a discovery portal for geospatial data.



FORMATTING YOUR DATA

CREATE WELL ORGANISED AND LONGER-LASTING DATA

USING STANDARD AND INTERCHANGEABLE OR OPEN LOSSLESS DATA FORMATS ENSURES LONG-TERM USABILITY OF DATA. HIGH QUALITY DATA ARE WELL ORGANISED, STRUCTURED, NAMED AND VERSIONED AND THE AUTHENTICITY OF MASTER FILES IDENTIFIED.

FILE FORMATS

The format and software in which research data are created and digitised usually depend on how researchers plan to analyse data, the hardware used, the availability of software, or can be determined by discipline-specific standards and customs.

All digital information is designed to be interpreted by computer programs to make it understandable and is – by nature – software dependent. All digital data may thus be endangered by the obsolescence of the hardware and software environment on which access to data depends.

Despite the backward compatibility of many software packages to import data created in previous software versions and the interoperability between competing popular software programs, the safest option to guarantee long-term data access is to convert data to standard formats that most software are capable of interpreting, and that are suitable for data interchange and transformation.

FILE FORMATS CURRENTLY RECOMMENDED BY THE UK DATA ARCHIVE FOR LONG-TERM PRESERVATION OF RESEARCH DATA

TYPE OF DATA	RECOMMENDED FILE FORMATS FOR SHARING, RE-USE AND PRESERVATION
Quantitative tabular data with extensive metadata a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information some structured text or mark-up file containing metadata information, e.g. DDI XML file
Quantitative tabular data with minimal metadata a matrix of data with or without column headings or variable names, but no other metadata or labelling	comma-separated values (CSV) file (.csv) tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate
Geospatial data vector and raster data	ESRI Shapefile (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data
Qualitative data textual	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml) Rich Text Format (.rtf) plain text data, ASCII (.txt)
Digital image data	TIFF version 6 uncompressed (.tif)
Digital audio data	Free Lossless Audio Codec (FLAC) (.flac)
Digital video data	MPEG-4 (.mp4) motion JPEG 2000 (.jp2)
Documentation	Rich Text Format (.rtf) PDF/A or PDF (.pdf) OpenDocument Text (.odt)

Note that other data centres or digital archives may recommend different formats.

This typically means using open standard formats – such as OpenDocument Format (ODF), ASCII, tab-delimited format, comma-separated values, XML – as opposed to proprietary ones. Some proprietary formats, such as MS Rich Text Format and MS Excel, are widely used and are likely to be accessible for a reasonable, but not unlimited, time.

Thus, whilst researchers should use the most suitable data formats and software according to planned analyses, once data analysis is completed and data are prepared for storing, researchers should consider converting their research data to standard, interchangeable and longer-lasting formats. Similarly for back-ups of data, standard formats should be considered.

For long-term digital preservation, data archives hold data in such standard formats. At the same time, data may be offered to users by conversion to current common and user-friendly data formats. Data may be migrated forward when needed.

CASE STUDY

PRESERVING AND SHARING MODELS

Various initiatives aim to preserve and share modelling software and code.

In biology, the BioModels Database of the European Bioinformatics Institute (EBI) is a repository for peer-reviewed, published, computational models of biological processes and molecular functions.[19] All models are annotated and linked to relevant data resources. Researchers are encouraged to deposit models written in an open source format, the Systems Biology Markup Language (SBML), and models are curated for long-term preservation.

FILE FORMATTING

The Wessex Archaeology Metric Archive Project has brought together metric animal bone data from a range of archaeological sites in England into a single database format.[20]

The dataset contains a selection of measurements commonly taken during Wessex Archaeology zoo-archaeological analysis of animal bone fragments found during field investigations. It was created by the researchers in MS Excel and MS Access formats and deposited with the Archaeology Data Service (ADS) in the same formats.

ADS has preserved the dataset in Oracle and in comma-separated values format (CSV) and disseminates the data via both as an Oracle/Cold Fusion live interface and as downloadable CSV files.

DATA CONVERSIONS

When researchers offer data to data archives for preservation, researchers themselves should convert data to a preferred data preservation format, as the person who knows the data is in the best position to ensure data integrity during conversions. Advice should be sought on up-to-date formats from the intended place of deposit.

When data are converted from one format to another – through export or by using data translation software – certain changes may occur to the data. After conversions, data should be checked for errors or changes that may be caused by the export process:

- for data held in statistical packages, spreadsheets or databases, some data or internal metadata such as missing value definitions, decimal numbers, formulae or variable labels may be lost during conversions to another format, or data may be truncated
- for textual data, editing such as highlighting, bold text or headers/footers may be lost

CASE STUDY

FILE CONVERSIONS

The JISC-funded Data Management for Bio-Imaging project at the John Innes Centre developed BioformatsConverter software to batch convert bio images from a variety of proprietary microscopy image formats to the Open Microscopy Environment format, OME-TIFF.[21] OME-TIFF, an open file format that enables data sharing across platforms, maintains the original image metadata in the file in XML format.

ORGANISING FILES AND FOLDERS

Well-organised file names and folder structures make it easier to find and keep track of data files. Develop a system that works for your project and use it consistently.

Good file names can provide useful cues to the content and status of a file, can uniquely identify a file and can help in classifying files. File names can contain project acronyms, researchers' initials, file type information, a version number, file status information and date.

Best practice is to:

- create meaningful but brief names
- use file names to classify broad types of files
- avoid using spaces and special characters
- avoid very long file names

Whilst computers add basic information and properties to a file, such as file type, date and time of creation and modification, this is not reliable data management. It is better to record and represent such essential information in file names or through the folder structure.

Think carefully how best to structure files in folders, in order to make it easy to locate and organise files and versions. When working in collaboration the need for an orderly structure is even higher.

Consider the best hierarchy for files, deciding whether a deep or shallow hierarchy is preferable. Files can be organised in folders according to types of: data – databases, text, images, models, sound; research activities – interviews, surveys, focus groups; or material – data, documentation, publications.

QUALITY ASSURANCE

Quality control of data is an integral part of all research and takes place at various stages: during data collection, data entry or digitisation, and data checking. It is important to assign clear roles and responsibilities for data quality assurance at all stages of research and to develop suitable procedures before data gathering starts.

During data collection, researchers must ensure that the data recorded reflect the actual facts, responses, observations and events.

Quality control measures during data collection may include:

- calibration of instruments to check the precision, bias and/or scale of measurement
- taking multiple measurements, observations or samples
- checking the truth of the record with an expert
- using standardised methods and protocols for capturing observations, alongside recording forms with clear instructions
- computer-assisted interview software to: standardise interviews, verify response consistency, route and customise questions so that only appropriate questions are asked, confirm responses against previous answers where appropriate and detect inadmissible responses

The quality of data collection methods used strongly influences data quality and documenting in detail how data are collected provides evidence of such quality.

When data are digitised, transcribed, entered in a database or spreadsheet, or coded, quality is ensured and error avoided by using standardised and consistent procedures with clear instructions. These may include:

- setting up validation rules or input masks in data entry software
- using data entry screens
- using controlled vocabularies, code lists and choice lists to minimise manual data entry
- detailed labelling of variable and record names to avoid confusion
- designing a purpose-built database structure to organise data and data files

During data checking, data are edited, cleaned, verified, cross-checked and validated.

Checking typically involves both automated and manual procedures. These may include:

- double-checking coding of observations or responses and out-of-range values
- checking data completeness
- verifying random samples of the digital data against the original data
- double entry of data
- statistical analyses such as frequencies, means, ranges or clustering to detect errors and anomalous values
- peer review

Researchers can add significant value to their data by including additional variables or parameters that widen the possible applications. Including standard parameters or generic derived variables in data files may substantially increase the potential re-use value of data and provide new avenues for research. For example, geo-referencing data may allow other researchers to add value to data more easily and apply the data in geographical information systems. Equally, sharing field notes from an interviewing project can help enrich the research context.

CASE STUDY

ADDING VALUE TO DATA

The Commission for Rural Communities (CRC) often use existing survey data to undertake rural and urban analysis of national scale data in order to analyse policies related to deprivation.

In order to undertake this type of spatial analysis, original postcodes need to be accessed and retrospectively recoded according to the type of rural or urban settlements they fall into. This can be done with the use of products such as the National Statistics Postcode Directory, which contains a classification of rural and urban settlements in England.

The task of applying these geographical markers to datasets can often be a long and sometimes unfruitful process – sometimes the CRC have to go through this process to just find that the data do not have a representative rural sample frame.

If rural and urban settlement markers, such as the Rural/Urban Definition for England and Wales were included in datasets, this would be of great benefit to those undertaking rural and urban analysis.[22]

VERSION CONTROL AND AUTHENTICITY

A version is where a file is closely related to another file in terms of its content. It is important to ensure that different versions of files, related files held in different locations, and information that is cross-referenced between files are all subject to version control. It can be difficult to locate a correct version or to know how versions differ after some time has elapsed.[23]

A suitable version control strategy depends on whether files are used by single or multiple users, in one or multiple locations and whether or not versions across users or locations need to be synchronised or not.

It is important to keep track of master versions of files, for example the latest iteration, especially where data files are shared between people or locations, e.g. on both a PC and a laptop. Checks and procedures may also need to be put in place to make sure that if the information in one file is altered, the related information in other files is also updated.

Best practice is to:

- decide how many versions of a file to keep, which versions to keep, for how long and how to organise versions
- identify milestone versions to keep
- uniquely identify files using a systematic naming convention
- record version and status of a file, e.g. draft, interim, final, internal
- record what changes are made to a file when a new version is created
- record relationships between items where needed, e.g. relationship between code and the data file it is run against; between data file and related documentation or metadata; or between multiple files
- track the location of files if they are stored in a variety of locations
- regularly synchronise files in different locations, e.g. using MS SyncToy software
- maintain single master files in a suitable file format to avoid version control problems associated with multiple working versions of files being developed in parallel
- identify a single location for the storage of milestone and master versions

The version of a file can be identified via:

- date recorded in file name or within file
- version numbering in file name (v1, v2, v3 or 00.01, 01.00)
- version description in file name or within file (draft, final)
- file history, version control table or notes included within a file, where versions, dates, authors and details of changes to the file are recorded

Version control can also be maintained through:

- version control facilities within software used
- using versioning software, e.g. Subversion (SVN)
- using file sharing services such as Dropbox, Google Docs or Amazon S3
- controlling rights to file editing
- manual merging of entries or edits by multiple users

Best practice to ensure authenticity is to:

- keep a single master file of data
- assign responsibility for master files to a single project team member
- regulate write access to master versions of data files
- record all changes to master files
- maintain old master files in case later ones contain errors
- archive copies of master files at regular intervals
- develop a formal procedure for the destruction of master files

Because digital information can be copied or altered so easily, it is important to be able to demonstrate the authenticity of data and to be able to prevent unauthorised access to data that may potentially lead to unauthorised changes.

TRANSCRIPTION

Good quality and consistent transcription that matches the analytic and methodological aims of the research is part of good data management planning. Attention needs to be given to transcribing conventions needed for the research, transcription instructions or guidelines and a template to ensure uniformity across a collection.

Transcription is a translation between forms of data, most commonly to convert audio recordings to text in qualitative research. Whilst transcription is often part of the analysis process, it also enhances the sharing and re-use potential of qualitative research data. Full transcription is recommended for data sharing.

If transcription is outsourced to an external transcriber, attention should be paid to:

- data security when transmitting recordings and transcripts between researcher and transcriber
- data security procedures for the transcriber to follow
- a non-disclosure agreement for the transcriber
- transcriber instructions or guidelines, indicating required transcription style, layout and editing

Best practice is to:

- consider the compatibility of transcription formats with import features of qualitative data analysis software, e.g. loss of headers and formatting, before developing a template or guidelines
- develop a transcription template to use, especially if multiple transcribers carry out work
- ensure consistency between transcripts
- anonymise data during transcription, or mark sensitive information for later anonymisation

Transcripts should:

- have a unique identifier that labels an interview either through a name or number
- have a uniform layout throughout a research project or data collection
- use speaker tags to indicate turn-taking or question/answer sequence in conversations
- carry line breaks between turn-takes
- be page numbered
- have a document cover sheet or header with brief interview or event details such as date, place, interviewer name, interviewee details

Transcription of statistical tables from historical sources into spreadsheets requires the digital data to be as close to the original as possible, with attention to consistency in transcribing and avoiding the use of formatting in data files.

VERSION CONTROL TABLE FOR A DATA FILE	
Title:	Vision screening tests in Essex nurseries
File Name:	VisionScreenResults_00_05
Description:	Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007
Created By:	Chris Wilkinson
Maintained By:	Sally Watsley
Created:	04/07/ 2007
Last Modified:	25/11/ 2007
Based on:	VisionScreenDatabaseDesign_02_00

VERSION	RESPONSIBLE	NOTES	LAST AMENDED
00_05	Sally Watsley	Version 00_03 and 00_04 compared and merged by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from SK	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

CASE STUDY

MODEL TRANSCRIPT RECOMMENDED BY THE UK DATA ARCHIVE

This model transcript shows the suggested layout with the inclusion of contextual and identifying information for the interview.

Study Title: Immigration Stories
Depositor: K. Clark
Interviewer: Ina Jones

Interview number: 12
Interview ID: Yolande
Date of interview: 12 June 1999

Information about interviewee:
Date of birth: 4 April 1947
Gender: female
Geographic region: Essex

Marital status: married
Occupation: catering
Ethnicity: Chinese

Y: I came here in late 1968.

I: You came here in late 1968? Many years already.

Y: 31 years already. 31 years already.

I: (laugh) It is really a long time. Why did you choose to come to England at that time?

Y: I met my husband and after we got married in Hong Kong, I applied to come to England.

I: You met your husband in Hong Kong?

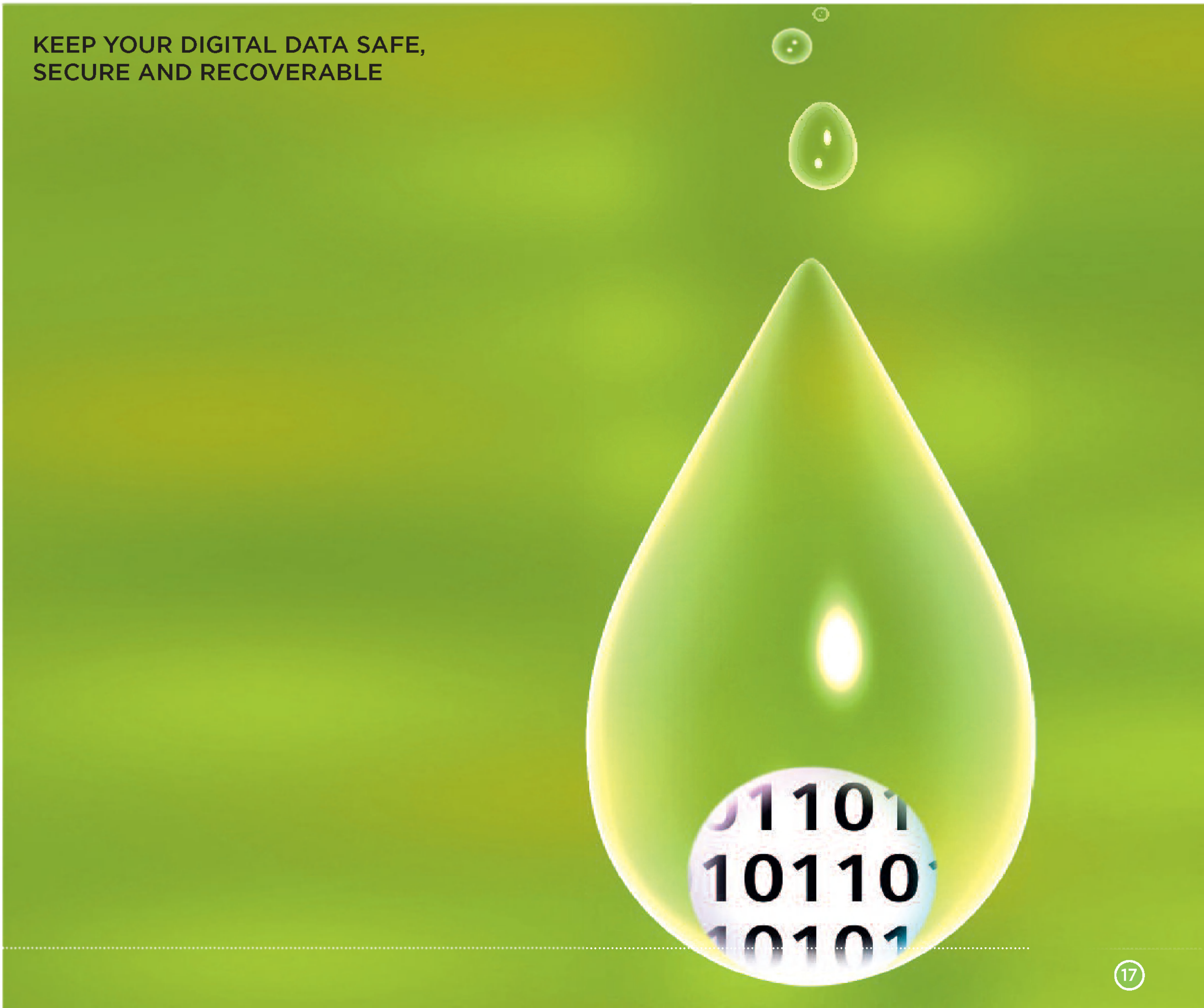
Y: Yes.

I: He was working here [in England] already?

Y: After he worked here for a few years — in the past, it was quite common for them to go back to Hong Kong to get a wife. Someone introduced us and we both fancied each other. At that time, it was alright to me to get married like that as I wanted to leave Hong Kong. It was like a gamble. It was really like a gamble.
...

STORING YOUR DATA

KEEP YOUR DIGITAL DATA SAFE, SECURE AND RECOVERABLE



LOOKING AFTER RESEARCH DATA FOR THE LONGER-TERM AND PROTECTING THEM FROM UNWANTED LOSS REQUIRES HAVING GOOD STRATEGIES IN PLACE FOR SECURELY STORING, BACKING-UP, TRANSMITTING, AND DISPOSING OF DATA. COLLABORATIVE RESEARCH BRINGS CHALLENGES FOR THE SHARED STORAGE OF, AND ACCESS TO, DATA.

MAKING BACK-UPS

Making back-ups of files is an essential element of data management. Regular back-ups protect against accidental or malicious data loss due to:

- hardware failure
- software or media faults
- virus infection or malicious hacking
- power failure
- human errors

Backing-up involves making copies of files which can be used to restore originals if there is loss of data. Choosing a precise back-up procedure depends on local circumstances, the perceived value of the data and the levels of risk considered appropriate. Where data contain personal information, care should be taken to only create the minimal number of copies needed, e.g. a master file and one back-up copy. Back-up files can be kept on a networked hard drive or stored offline on media such as recordable CD/DVD, removable hard drive or magnetic tape. Physical media can be removed to another location for safe-keeping.

For best back-up procedures, consider:

- whether to back-up particular files or the entire computer system (complete system image)
- the frequency of back-up needed, after each change to a data file or at regular intervals
- strategies for all systems where data are held, including portable computers and devices, non-network computers and home-based computers
- organising and clearly labelling all back-up files and media

CRITICAL FILES AND MASTER COPIES

Critical data files or frequently used ones may be backed-up daily using an automated back-up process and are best stored offline. Master copies of critical files should be in open, as opposed to proprietary, formats for long-term validity. Back-up files should be verified and validated regularly, either by fully restoring them to another location and comparing them with the originals or by checking back-up copies for completeness and integrity, for example by checking the MD5 checksum value, file size and date.

INSTITUTIONAL BACK-UP POLICY

Most institutions have a back-up policy for data held on a network space. Check with your institution to find out which strategy or policy is in place. If you are not happy with the robustness of the solution, you should keep independent back-up copies of critical files.

INCREMENTAL OR DIFFERENTIAL BACK-UPS

Incremental back-ups consist of first making a copy of all relevant files, often the complete contents of a PC, then making incremental back-ups of the files which have altered since the last back-up. Removable media (CD/DVD) are recommended for this procedure.

For differential back-ups, a complete back-up is made first, and then back-ups are made of files changed or created since the first full back-up and not just since the last partial back-up. Fixed media, such as hard drives, are recommended for this method.

Whichever method is used, it is best not to overwrite old back-ups with new.

DATA STORAGE

A data storage strategy is important because digital storage media are inherently unreliable, unless they are stored appropriately, and all file formats and physical storage media will ultimately become obsolete. The accessibility of any data depends on the quality of the storage medium and the availability of the relevant data-reading equipment for that particular medium. Media currently available for storing data files are optical media (CDs and DVDs) and magnetic media (hard drives and tapes).

Best practice is to:

- store data in non-proprietary or open standard formats for long-term software readability (see file formats table)
- copy or migrate data files to new media between two and five years after they were first created, since both optical and magnetic media are subject to physical degradation
- check the data integrity of stored data files at regular intervals
- use a storage strategy, even for a short-term project, with two different forms of storage, e.g. on hard drive and on CD
- create digital versions of paper documentation in PDF/A format for long-term preservation and storage
- organise and clearly label stored data so they are easy to locate and physically accessible
- ensure that areas and rooms for storage of digital or non-digital data are fit for the purpose, structurally sound, and free from the risk of flood and fire

The National Preservation Office has published guidelines on caring for CDs and DVDs, which are vulnerable to poor handling, changes in temperature, relative humidity, air quality and lighting conditions.[24]

CASE STUDY

DATA BACK-UP AND STORAGE

A research team carrying out coral reef research collects field data using handheld Personal Digital Assistants (PDAs). Digital data are transmitted daily to the institution's network drive, where they are held in password-protected files. All data files are identified by an individual version number and creation date. Version information (version numbers and notes detailing differences between versions) is stored in a spreadsheet, also on the network drive. The institution's network drive is fully backed-up onto Ultrium LTO2 data tapes. Incremental back-ups are made daily Monday to Thursday; full server back-ups are made from Friday to Sunday. Tapes are securely stored in a separate building. Upon completion of the research the data are deposited in the institution's digital repository.

DATA BACK-UP AND STORAGE

In February 2008 the British Library (BL) received the recorded output of the Survey of Anglo-Welsh Dialects (SAWD), carried out by University College, Swansea, between 1969 and 1995. This survey recorded the English spoken in Wales by interviewing and tape-recording elderly speakers on topics including the farm and farming, the house and housekeeping, nature, animals, social activities and the weather. The collection was deposited in the form of 503 digital audio files, which were accessioned as .wav files in the BL's Digital Library. Digital clones of all files are held at the Archive of Welsh English, alongside the original master recordings on 151 audio cassettes, from which the digital copies were created.

The BL's Digital Library is mirrored on four sites – at Boston Spa, St Pancras, Aberystwyth and a 'dark' archive which is provided by a third party. Each of these servers has inbuilt integrity checks. The BL makes available access copies for users, in the form of .mp3 audio files, in the British Library Reading Rooms via the Soundserver system. A small set of audio extracts from the SAWD recordings are also available online on the BL's Accents and Dialects web site, Sounds Familiar.

DATA SECURITY

Physical security, network security and security of computer systems and files all need to be considered to ensure security of data and prevent unauthorised access, changes to data, disclosure or destruction of data. Data security arrangements need to be proportionate to the nature of the data and the risks involved. Attention to security is also needed when data are to be destroyed.

Data security may be needed to protect intellectual property rights, commercial interests, or to keep personal or sensitive information safe.

Physical data security requires:

- controlling access to rooms and buildings where data, computers or media are held
- logging the removal of, and access to, media or hardcopy material in store rooms
- transporting sensitive data only under exceptional circumstances, even for repair purposes, e.g. giving a failed hard drive containing sensitive data to a computer manufacturer may cause a breach of security

Network security means:

- not storing confidential data such as those containing personal information on servers or computers connected to an external network, particularly servers that host internet services
- firewall protection and security-related upgrades and patches to operating systems to avoid viruses and malicious code

Security of computer systems and files may include:

- locking computer systems with a password and installing a firewall system
- protecting servers by power surge protection systems through line-interactive uninterruptible power supply (UPS) systems
- implementing password protection of, and controlled access to, data files, e.g. no access, read only, read and write or administrator-only permission
- controlling access to restricted materials with encryption
- imposing non-disclosure agreements for managers or users of confidential data
- not sending personal or confidential data via email or through File Transfer Protocol (FTP), but rather transmit as encrypted data
- destroying data in a consistent manner when needed

SECURITY OF PERSONAL DATA

Where the safeguarding of personal data is involved, data security is based on national legislation, the Data Protection Act 1998, which dictates that personal data should only be accessible to authorised persons. Personal data may also exist in non-digital format, for example as patient records, signed consent forms, or interview cover sheets. These should be protected in the same secure way as digital files.

Data that contain personal information should be treated with higher levels of security than data which do not. Security can be made easier by:

- anonymising or aggregating data
- separating data content according to security needs
- removing personal information, such as names and addresses, from data files and storing them separately
- encrypting data containing personal information before they are stored - encryption is certainly needed before transmission of such data.

How confidential data or data containing personal information are stored may need to be addressed during informed consent procedures. This ensures that the persons to whom the personal data belong are informed and give their consent as to how the data are stored or transmitted.

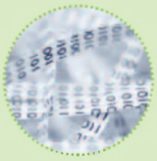
CASE STUDY

DATA DESTRUCTION

The German Institute for Standardisation (DIN) has standardised levels of destruction for paper and discs that have been adopted by the shredding industry.

For shredding confidential material, adopting DIN 3 means objects are cut into two millimetre strips or confetti like cross-cut particles of 4x40mm. The UK government requires a minimum standard of DIN 4 for its material, which ensures cross cut particles of at least 2x15mm.

The highest security level is known as DIN 6, this is used by the United States federal government for ultra secure shredding of top secret or classified material, cross cutting into 1x5mm particles.



DATA TRANSMISSION AND ENCRYPTION

Transmitting data between locations or within research teams can be challenging for data management infrastructure. To ensure sensitive or personal data is secure for transmitting it must be encrypted to an appropriate standard. Only data confirmed as anonymised or non-sensitive should be transmitted in unencrypted form. Encryption maintains the security of data during transmission.

Relying on email to transmit data, even internally, is a vulnerable point in protecting sensitive data. Be aware that anything sent by email persists in numerous exchange servers – the sender's, the receiver's and others in-between.

TRANSFERRING LARGE FILES

In an era of large-scale data collection, transferring large files can be problematic. Third party commercial file sharing services exist to facilitate the movement of files. However, services such as Google Docs or YouSendIt are not necessarily permanent or secure, and are often located overseas and therefore not covered by UK law. They may even be in potential violation of UK law, particularly in relation to the UK Data Protection Act (1998) which states data should not be transferred to other countries without adequate protection. Encrypting data before transfer offers protection.

A dropbox service can be a safe solution for transferring large data files, if it is managed and controlled by the responsible institution. For example, the UK Data Archive recommends data deposits from researchers are made via the University of Essex dropbox service, with data files containing sensitive or personal information encrypted before submission.

ENCRYPTING DATA

After testing a number of software applications for encrypting data to enable secure data transmission from government departments to the Archive, the UK Data Archive recommends the use of Pretty Good Privacy (PGP), an industry-standard encryption technology. Using this method, encrypted data can be transferred via portable media or electronically via file upload or email. Reliable open source encryption software exists, e.g. GnuPG.

Encryption requires the creation of a public and private key pair and a passphrase. The private PGP key and passphrase are used to digitally sign each encrypted file, and thus allow the recipient to validate the sender's identity. The recipient's public PGP key is installed by the sender in order to encrypt files so that only the authorised recipient can decrypt them.

DATA DISPOSAL

Deleting files and reformatting a hard drive will not prevent the possible recovery of data that have previously been on that hard drive. Having a strategy for reliably erasing data files is a critical component of managing data securely and is relevant at various stages in the data cycle. During research, copies of data files no longer needed can be destroyed. At the conclusion of research, data files which are not to be preserved need to be disposed of securely.

For hard drives, which are magnetic storage devices, simply deleting does not erase a file on most systems, but only removes a reference to the file. It takes little effort to restore files deleted in this way. Files need to be overwritten to ensure they are effectively scrambled. Software is available for the secure erasing of files from hard discs, meeting recognised standards of overwriting to adequately scramble sensitive files. Example software is BC Wipe, Wipe File, DeleteOnClick and Eraser for Windows platforms. Mac users can use the standard 'secure empty trash' option; an alternative is Permanent Eraser software.

Flash-based solid state discs, such as memory sticks, are constructed differently to hard drives and techniques for securely erasing files on hard drives can not be relied on to work for solid state discs as well. Physical destruction is advised as the only certain way to erase files.

The most reliable way to dispose of data is physical destruction. Risk-adverse approaches for all drives are to: encrypt devices when installing the operating software and before first use; and physically destroy the drive using a secure destruction facility approved by your institution when data need to be destroyed.

Shredders certified to an appropriate security level should be used for destroying paper and CD/DVD discs. Computer or external hard drives at the end of their life can be removed from their casings and disposed of securely through physical destruction.

FILE SHARING AND COLLABORATIVE ENVIRONMENTS

Collaborative research can be challenging when it comes to facilitating data sharing, transfer and storage, and providing access to data across various partners or institutions. Whilst various virtual research environments exist, at the time of writing these are all relatively undeveloped, require significant set up and maintenance costs, and are usually mono-institutional. Consequently, many researchers are not comfortable with their features and may still resort to data transfer via email and online file sharing services.

Cloud-based file sharing services and wikis may be suitable for sharing certain types of data, but they are not recommended for data that may be confidential, and users need to be aware that they do not control where data are ultimately stored.

The ideal solution would be a system that facilitates co-operation yet is able to be adopted by users with minimal training and delivers benefits that help with data management practices. Such solutions using open source repository software like Fedora and DSpace are currently being developed at the time of writing, but are not yet user-friendly or scalable. Institutional IT services should be involved in finding the best solution for a research project.

Virtual research environments often provide an encrypted shared workspace for data files and documents in group collaboration. Users can create workspaces, add and invite members to a workspace, and each member has a privately editable copy of the workspace. Users interact and collaborate in the common workspace which is a private virtual location. Changes are tracked and sent to all members and all copies of the workspace are synchronised via the network in a peer-to-peer manner. If a platform is adopted it should be encrypted to an appropriate standard and be version controlled.

CASE STUDY

VIRTUAL RESEARCH ENVIRONMENTS

A range of systems are commonly used across higher education institutions, which have advantages and disadvantages for file transfer and sharing. Examples are SharePoint and Sakai.

SharePoint is used across UK universities, with varying degrees of satisfaction. Researchers finding it cumbersome to install and configure. Advantages are: enabling external access and collaboration, the use of document libraries, team sites, workflow management processes, and version control ability.

Open source system Sakai has been used by projects of the JISC virtual research environment programme and can be obtained under an educational community licence.[25] It has an established support network and features announcement facilities, a drop box for private file sharing, email archive, resources library, communications functions, scheduling tools and control of permissions and access.

ETHICS AND CONSENT

SHARE SENSITIVE AND CONFIDENTIAL RESEARCH DATA ETHICALLY

A COMBINATION OF GAINING CONSENT FOR DATA SHARING, ANONYMISING AND REGULATING ACCESS TO DATA WILL INCREASE THE POTENTIAL FOR MAKING PEOPLE-RELATED RESEARCH DATA MORE READILY AND WIDELY AVAILABLE.

When research involves obtaining data from people, researchers are expected to maintain high ethical standards such as those recommended by professional bodies, institutions and funding organisations, both during research and when sharing data.

Research data — even sensitive and confidential data — can be shared ethically and legally if researchers pay attention, from the beginning of research, to three important aspects:

- when gaining informed consent, include provision for data sharing
- where needed, protect people's identities by anonymising data
- consider controlling access to data

These measures should be considered jointly. The same measures form part of good research practice and data management, even if data sharing is not envisioned.

Data collected from and about people may hold personal, sensitive or confidential information. This does not mean that all data obtained by research with participants are personal or confidential.

LEGAL AND ETHICAL ISSUES

Strategies for dealing with confidentiality depend upon the nature of the research, but are essentially informed by a researcher's ethical and legal obligations. A duty of confidentiality towards informants may be explicit, but need not be.

DEFINITIONS

Personal data
Personal data are data which relate to a living individual who can be identified from those data or from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller and includes any expression of opinion about the individual and any indication of the intentions of the data controller. This includes any other person in respect of the individual (Data Protection Act 1998).

Confidential data
Confidential data are data given in confidence or data agreed to be kept confidential, i.e. secret, between two parties, that are not in the public domain such as information on business, income, health, medical details, and political opinion.

Ssensitive personal data
Sensitive personal data are defined in the Data Protection Act 1998 as data on a person's race, ethnic origin, political opinion, religious or similar beliefs, trade union membership, physical or mental health or condition, sexual life, commission or alleged commission of an offence, proceedings for an offence (alleged to have been) committed, disposal of such proceedings or the sentence of any court in such proceedings.

Legislation that may impact on the sharing of confidential data:[26]

- Data Protection Act 1998
- Freedom of Information Act 2000
- Human Rights Act 1998
- Statistics and Registration Services Act 2007
- Environmental Information Regulations 2004

SAMPLE CONSENT STATEMENT FOR QUANTITATIVE SURVEYS

Thank you very much for agreeing to participate in this survey.

The information provided by you in this questionnaire will be used for research purposes. It will not be used in any manner which would allow identification of your individual responses.

Anonymised research data will be archived at .... in order to make them available to other researchers in line with current data sharing practices.

INFORMED CONSENT AND DATA SHARING

Researchers are usually expected to obtain informed consent for people to participate in research and for use of the information collected. Where possible, consent should also take into account any future uses of data, such as the sharing, preservation and long-term use of research data. At a minimum, consent forms should not preclude data sharing, such as by promising to destroy data unnecessarily.

Researchers should:

- inform participants how research data will be stored, preserved and used in the long-term
- inform participants how confidentiality will be maintained, e.g. by anonymising data
- obtain informed consent, either written or verbal, for data sharing

To ensure that consent is informed, consent must be freely given with sufficient information provided on all aspects of participation and data use. There must be active communication between the parties. Consent must never be inferred from a non-response to a communication such as a letter. Without consent for data sharing, opportunities for sharing research data with other researchers can be jeopardised.

SAMPLE CONSENT FORM FOR INTERVIEWS

CONSENT FORM FOR [NAME OF PROJECT]

Please tick the appropriate boxes	**Yes**	**No**
Taking Part		
I have read and understood the project information sheet dated DD/MM/YYYY.	□	□
I have been given the opportunity to ask questions about the project.	□	□
I agree to take part in the project. Taking part in the project will include being interviewed and recorded (audio or video).[a]	□	□
I understand that my taking part is voluntary; I can withdraw from the study at any time and I do not have to give any reasons for why I no longer want to take part.	□	□
Use of the information I provide for this project only		
I understand my personal details such as phone number and address will not be revealed to people outside the project.	□	□
I understand that my words may be quoted in publications, reports, web pages, and other research outputs.	□	□
Please choose one ***of the following two options:***		
I would like my real name used in the above	□	
I would **not** like my real name to be used in the above.	□	
Use of the information I provide beyond this project		
I agree for the data I provide to be archived at the UK Data Archive.[b]	□	□
I understand that other genuine researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
I understand that other genuine researchers may use my words in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	□	□
So we can use the information you provide legally		
I agree to assign the copyright I hold in any materials related to this project to [name of researcher].	□	□

Name of participant [printed] Signature ______________________________ Date ______________

Researcher [printed] Signature ______________________________ Date ______________

Project contact details for further information: Names, phone, email addresses, etc.

Notes:
[a] Other forms of participation can be listed.
[b] More detail can be provided here so that decisions can be made separately about audio, video, transcripts, etc.

RESEARCH ETHICS COMMITTEES AND DATA SHARING

The role of Research Ethics Committees (RECs) is to help protect thesafety, rights and well-being of research participants and to promote ethically sound research. This involves ensuring that research complies with the Data Protection Act 1998 regarding the use of personal information collected in research.

In research with people, there can be a perceived tension between data sharing and data protection where research data contain personal, sensitive or confidential information. However, in many cases, data obtained from people can be shared while upholding both the letter and the spirit of data protection and research ethics principles.

RECs can play a role in this by advising researchers that:

- most research data obtained from participants can be successfully shared without breaching confidentiality
- it is important to distinguish between personal data collected and research data in general
- data protection laws do not apply to anonymised data
- personal data should not be disclosed, unless consent has been given for disclosure
- identifiable information may be excluded from data sharing
- many funders recommend or require data sharing or data management planning
- even personal sensitive data can be shared if suitable procedures, precautions and safeguards are followed, as is done at major data centres

For example, survey and qualitative data held at the UK Data Archive are typically anonymised, unless specific consent has been given for personal information to be included. They are not in the public domain and their use is regulated for specific purposes after user registration. Users sign a licence in which they agree to conditions such as not attempting to identify any individuals from the data and not sharing data with unregistered users. For confidential or sensitive data stricter access regulations may be imposed.

RECs can play a critical role by providing such information to researchers, at the consent planning stages, on how to share data ethically.

WRITTEN OR VERBAL CONSENT?

Whether informed consent is obtained in writing through a detailed consent form, by means of an informative statement, or verbally, depends on the nature of the research, the kind of data gathered, the data format and how the data will be used.

For detailed interviews or research where personal, sensitive or confidential data are gathered:

- the use of written consent forms is recommended to assure compliance with the Data Protection Act and with ethical guidelines of professional bodies and funders
- written consent typically includes an information sheet and consent form signed by the participant
- verbal consent agreements can be recorded together with audio or video recorded data

For surveys or informal interviews, where no personal data are gathered or personal identifiers are removed from the data:

- obtaining written consent may not be required • an information sheet should be provided to participants detailing the nature and scope of the study, the identity of the researcher(s) and what will happen to data collected (including any data sharing)

Sample consent forms are available from the UK Data Archive.[27]

ONE-OFF OR PROCESS CONSENT?

Discussing and obtaining all forms of consent can be a oneoff occurrence or an ongoing process.

- One-off consent is simple, practical, avoids repeated requests to participants, and meets the formal requirements of most Research Ethics Committees. However, it may place too much emphasis on 'ticking boxes'.

- Process consent is considered throughout the research project and assures active informed consent from participants. Consent for participation in research, for primary data use and for data sharing, can be considered at different stages of the research. This gives participants a clearer view of what their participation means and how their data are to be shared. It may, however, be repetitive and burdenesome.

CASE STUDY

SHARING CONFIDENTIAL DATA

The Biological Records Centre (BRC) is the national custodian of data on the distribution of wildlife in the British Isles.[28] Data are provided by volunteers, researchers and organisations. BRC disseminates data for environmental decision-making, education and research. Data whose publication could present a significant threat to a species or habitat (e.g. nesting location of birds of prey) will be treated as confidential. The BRC provides access to the data it holds via the National Biodiversity Network Gateway. Standard access controls are as follows:

- public access to view and download all records at a minimum 10 km^2 level of resolution, and at higher resolution if the data provider agrees
- registered users have access to view and download all except confidential records at the 1 km^2 level of resolution
- conservation organisations have access to view and download all except confidential records at full resolution with attributes
- conservation officers in statutory conservation agencies have access to view and download all records, including confidential records at full resolution with attributes
- records that have been signified as confidential by a data provider will not be made available to the conservation agencies without the consent of the data provider.

ANONYMISING DATA

Before data obtained from research with people can be published or shared with other researchers, they may need to be anonymised so that individuals, organisations and businesses cannot be identified from the data.[29]

Anonymisation may be needed for ethical reasons to protect people's identities, for legal reasons to not disclose personal data, or for commercial reasons. Personal data should not be disclosed from research information, unless a respondent has given specific consent to do so.

Anonymisation may not be required for example, in oral histories where it is customary to publish and share the names of people interviewed, for which they have given their consent.

It can be time consuming, and therefore costly, to anonymise research data, in particular qualitative textual data. This is especially the case if not planned early in the research or left until the end of a project.

Data may be anonymised by:

- removing direct identifiers, e.g. name or address
- aggregating or reducing the precision of information or a variable, e.g. replacing date of birth by age groups
- generalising the meaning of detailed text, e.g. replacing a doctor's detailed area of medical expertise with an area of medical speciality
- using pseudonyms
- restricting the upper or lower ranges of a variable to hide outliers, e.g. top-coding salaries

A person's identity can be disclosed from:

- direct identifiers, e.g. name, address, postcode information or telephone number
- indirect identifiers that, when linked with other publicly available information sources, could identify someone, e.g. information on workplace, occupation or exceptional values of characteristics like salary or age

Special attention may be needed for:

- relational data, where relations between variables in related datasets can disclose identities
- geo-referenced data, where identifying spatial references such as point co-ordinates also have a geo-spatial value

Removing spatial references prevents disclosure, but it means that all geographical information is lost. A better option may be to keep spatial references intact and to impose access regulations on the data instead. As an alternative, point co-ordinates may be replaced by larger, non-disclosing geographical areas or by meaningful alternative variables that typify the geographical position.

Consideration should be given to the level of anonymity required to meet the needs agreed during the informed consent process. Researchers should not presume the only way to maintain confidentiality is by keeping data hidden. Obtaining informed consent for data sharing or regulating access to data should also be considered alongside any anonymisation.

Managing anonymisation:

- plan anonymisation early in the research at the time of data collection
- retain original unedited versions of data for use within the research team and for preservation
- create an anonymisation log of all replacements, aggregations or removals made
- store the log separately from the anonymised data files
- identify replacements in text in a meaningful way, e.g. in transcribed interviews indicate replaced text with [brackets] or use XML markup tags <anon>.....</anon>

AUDIO-VISUAL DATA

Digital manipulation of audio and image files can be used to remove personal identifiers. However, techniques such as voice alteration and image blurring are labour-intensive and expensive and are likely to damage the research potential of the data. If confidentiality of audio-visual data is an issue, it is better to obtain the participant's consent to use and share the data unaltered, with additional access controls if necessary.

CASE STUDY

DATA SECURITY AND ANONYMISATION

UK Biobank aims to collect medical and genetic data from 500,000 middle-aged people across the UK in order to create a research resource to study the prevention and treatment of serious diseases. Stringent security, confidentiality and anonymisation measures are in place.[30] UK Biobank holds personal data on recruited patients, their medical records and blood, urine and genetic samples, with data made available to approved researchers. Data or samples provided to researchers never include personal identifying details.

All data and samples are stored anonymously by removing any identifying information. This identifying information is encrypted and stored separately in a restricted access database that is controlled by senior UK Biobank staff. Identifying data and samples are only linked using a code that has no external meaning. Only a few people within UK Biobank have access to the key to the code for re-linking participants' identifying information with data and samples. All staff sign confidentiality agreements as part of their employment contracts.

ACCESS CONTROL

Under certain circumstances, sensitive and confidential data can be safeguarded by regulating use of or restricting access to such data, while at the same time enabling data sharing for research and educational purposes.

Data held at data centres and archives are not generally in the public domain. Their use is restricted to specific purposes after user registration. Users sign an End User Licence in which they agree to certain conditions, e.g. not to use data for commercial purposes or identify any potentially identifiable individuals.

Data centres may impose additional access regulations for confidential data such as:

- needing specific authorisation from the data owner to access data
- placing confidential data under embargo for a given period of time until confidentiality is no longer pertinent
- providing access to approved researchers only
- providing secure access to data through enabling remote analysis of confidential data but excluding the ability to download data

Mixed levels of access regulations may be put in place for some data collections, combining regulated access to confidential data with user access to non-confidential data.

Data centres typically liaise with the researchers who own the data in selecting the most suitable type of access for data. Access regulations should always be proportionate to the kind of data and confidentiality involved.

OPEN ACCESS

The digital revolution has caused a strong drive towards open access of information, with the internet making information sharing fast, easy, powerful and empowering.

Scholarly publishing has seen a strong move towards open access to increase the impact of research, with e-journals, open access journals and copyright policies enabling the deposit of outputs in open access repositories. The same open access movement also steers towards more open access to the underlying data and evidence on which research publications are based. A growing number of journals require for data that underpin research findings to be published in open access repositories when manuscripts are submitted.

CASE STUDY

ACCESS RESTRICTIONS VS. OPEN ACCESS

Working with data owners, the Secure Data Service provides researchers with secure access to data that are too detailed, sensitive or confidential to be made available under the standard licences operated by its sister service, the Economic and Social Data Service (ESDS).[31] The service's security philosophy is based upon training and trust, leading-edge technology, licensing and legal frameworks (including the 2007 Statistics Act), and strict security policies and penalties endorsed by both the ONS and the ESRC.

The technical model shares many similarities with the ONS Virtual Microdata Laboratory[32] and the NORC Secure Data Enclave.[33] It is based around a Citrix infrastructure which turns the end user's computer into a remote terminal. All data processing is carried out on a central secure server; no data travels over the network. Outputs for publication are only released subject to Statistical Disclosure Control checks by trained Service staff.

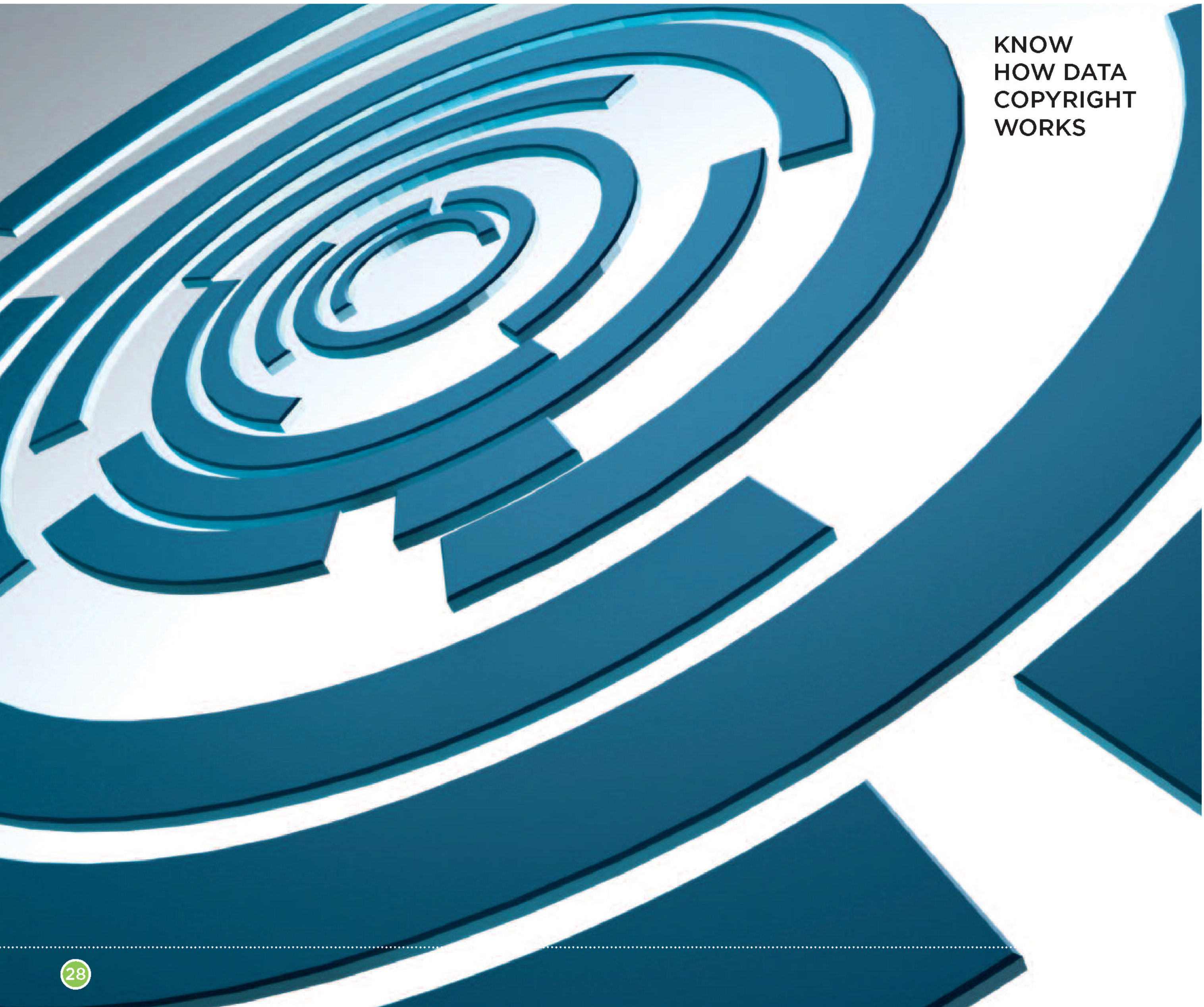
Secure Data Service data cannot be downloaded. Researchers analyse the data remotely from their home institution at their desktop or in a safe room. The Service provides a 'home away from home' research facility with familiar statistical software and MS Office tools to make remote collaboration and analysis secure and convenient.

The clearing-house mechanism established following the Convention on Biological Diversity to promote information sharing, has resulted in an exponential increase in openly accessible biodiversity and ecosystem data since 1992.

The Forest Spatial Information Catalogue is a web-based portal, developed by the Center for International Forestry Research (CIFOR), for public access to spatial data and maps.[34] The catalogue holds satellite images, aerial photographs, land usage and forest cover maps, maps of protected areas, agricultural and demographic atlases and forest boundaries. For example, forest cover maps for the entire world, produced by the World Conservation Monitoring Centre in 1997 can be downloaded freely as digital vector data.

The Global Biodiversity Information Framework (GBIF) strives to make the world's biodiversity data accessible everywhere in the world.[35] The framework holds millions of species occurrence records based on specimens and observations, scientific and common names and classifications of living organisms and map references for species records. Data are contributed by numerous international data providers. Geo-referenced records can be mapped to Google Earth.

COPYRIGHT



KNOW
HOW DATA
COPYRIGHT
WORKS

COPYRIGHT IS AN INTELLECTUAL PROPERTY RIGHT ASSIGNED AUTOMATICALLY TO THE CREATOR, THAT PREVENTS UNAUTHORISED COPYING AND PUBLISHING OF AN ORIGINAL WORK. COPYRIGHT APPLIES TO RESEARCH DATA AND PLAYS A ROLE WHEN CREATING, SHARING AND RE-USING DATA.

WHO OWNS COPYRIGHT?

Researchers creating data typically hold copyright in their data. Most research outputs – including spreadsheets, publications, reports and computer programs – fall under literary work and are therefore protected by copyright. Facts, however, cannot be copyrighted. The creator is automatically the first copyright owner, unless there is a contract that assigns copyright differently or there is written transfer of copyright signed by the copyright owner.

If information is structured in a database, the structure acquires a database right, alongside the copyright in the content of the database. A database may be protected by both copyright and database right. For database right to apply, the database must be the result of substantial intellectual investment in obtaining, verifying or presenting the content.

For copyright to apply, the work must be original and fixed in a material form, e.g. written or recorded. There is no copyright in ideas or unrecorded speech. If researchers collect data using interviews and make recordings or transcriptions of the interviewee's words, then the researcher holds the copyright of those recordings. In addition, each speaker is an author of his or her recorded words in the interview.[36]

In the case of collaborative research or derived data, copyright may be held jointly by various researchers or institutions. Copyright should be assigned correctly, especially if datasets have been created from a variety of sources; for example those which have been bought or 'lent' by other researchers. When digitising paper-based materials or images, or analogue recordings, attention needs to be paid to the copyright of the original material.

In academia, in theory the employer is the owner of the copyright in a work made during the employee's employment. Many academic institutions, however, assign copyright in research materials, data and publications to the researchers. Researchers should check how their institutions assign copyright.

RE-USE OF DATA AND COPYRIGHT

Secondary users of data must obtain copyright clearance from the rights holder before data can be reproduced. Data can be copied for non-commercial teaching or research purposes without infringing copyright, under the fair dealing concept, providing that the owner of the data is acknowledged. An acknowledgement should give credit to the data source used, the data distributor and the copyright holder.

When data are shared through a data centre, the researcher or data creator keeps the copyright over data and licences the centre to process and provide access to the data. A data centre cannot effectively hold data unless all the rights holders are identified and give their permission for the data to be archived and shared. Data centres typically specify how data should be acknowledged and cited, either within the metadata record for a dataset or in a data use licence.

CASE STUDY

COPYRIGHT OF MEDIA SOURCES

A researcher has collated articles about the Prime Minister from *The Guardian* over the past ten years, using the LexisNexis newspaper database to source articles. They are then transcribed/copied by the researcher into a database so that content analysis can be applied. The researcher offers a copy of the database together with the original transcribed text to a data centre.

Researchers cannot share either of these data sources as they do not have copyright in the original material. A data centre cannot accept these data as to do so would be breach of copyright. The rights holders, in this case *The Guardian* and LexisNexis, would need to provide consent for archiving.

USING THIRD PARTY DATA

The Stockholm Environmental Institute (SEI) has created an integrated spatial database, Social and Environmental Conditions in Rural Areas (SECRA).[37] This contains a wide range of socio-economic and environmental characteristics for all rural Census 2001 Super Output Areas (SOAs) for England.

Multiple third party data sources were used, such as Census 2001 data, Land Cover Map data and data from the Land Registry, Environment Agency, Automobile Association, Royal Mail and British Trust for Ornithology. Derived data have been calculated and mapped onto SOAs. The researchers would like to distribute the database for wider use.

Whilst the database contains no original third party data, only derived data, there is still joint copyright shared between the SEI and the various copyright holders of the third party data. The researchers have sought permission from all data owners to distribute the data and the copyright of all third party data is declared in the documentation. The database can therefore be distributed.

When research data are submitted to a journal to supplement a publication, researchers need to verify whether the publisher expects copyright transfer of the data.

Some researchers may have come across the concept of Creative Commons licences which allow creators to communicate the rights which they wish to keep and the rights which they wish to waive in order for other people to make re-use of their intellectual properties more straightforward. The Creative Commons licence is not usually suitable for data. Other licences with similar objectives are more appropriate, such as the Open Data Commons Licence or the Open Government Licence.[38]

CASE **STUDY**

COPYRIGHT OF INTERVIEWS WITH 'ELITES'

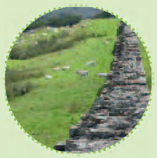
A researcher has interviewed five retired cabinet ministers about their careers, producing audio recordings and full transcripts. The researcher then analyses the data and offers the recordings and transcripts to a data centre for preserving. However the researcher did not get signed copyright transfers for further use of the interviewees' words.

In this case it would be problematic for a data centre to accept the data. Large extracts of the data cannot be quoted by secondary users. To do so would breach the interviewees' copyright over their recorded words. This is equally a problem for the primary researcher. The researcher should have asked for transfer of copyright or a licence to use the data obtained through interviews, as the possibility exists that the interviewee may at some point wish to assert the right over their words, e.g. when publishing memoirs.

COPYRIGHT OF LICENSED DATA

A researcher subscribes to access spatial AgCensus data from the data centre EDINA. These data are then integrated with data collected by the researcher. As part of the ESRC research award contract the data has to be offered for archiving at the UK Data Archive. Can such integrated data be offered?

The subscription agreement on accessing AgCensus data states that data may not be transferred to any other person or body without prior written permission from EDINA. Therefore, the UK Data Archive cannot accept the integrated data, unless the researcher obtains permission from EDINA. The researcher's partial data, with the AgCensus data removed, can be archived. Secondary users could then re-combine these data with the AgCensus data, if they were to obtain their own AgCensus subscription.



STRATEGIES FOR CENTRES

PROVIDE A DATA MANAGEMENT FRAMEWORK FOR RESEARCHERS

RESEARCH CENTRES AND PROGRAMMES CAN SUPPORT RESEARCHERS THROUGH A CO-ORDINATED DATA MANAGEMENT FRAMEWORK OF SHARED BEST PRACTICES. THIS CAN INCLUDE LOCAL GUIDANCE, TEMPLATES AND POINTERS TO KEY POLICIES.

During 2010, the UK Data Archive worked closely with selected ESRC research centres and programmes to evaluate existing data management practices and develop data management planning strategies. The recommendations contained in this guide are based on those real-case experiences.

Research hubs may operate a centralised or devolved approach to data management, co-ordinating how data are handled and organised or giving researchers all authority and responsibility for handling their research data. Factors that may influence which approach is taken include the size of the research hub, whether it is a single or cross-institutional entity and how much methodological and discipline diversity the data exhibit.

Advantages of a centralised approach to data management include:

- researchers can share good practice and data management experiences, thereby building capacity for the centre
- a uniform approach to data management can be established as well as relevant central data policies
- data ownership can be identified and kept track of over time, especially useful when researchers move
- data can be stored at a central location
- assurance that all researchers and staff are aware of duties, responsibilities and funder requirements regarding research data, with easy access to relevant information

At the same time, researchers need to take responsibility for managing their own data and a devolved approach to data management may give more flexibility to adapt to discipline requirements or may be needed to adapt data management according to research methods.

To provide a data management framework, research hubs can develop:

- the assignment of data management responsibilities to named individuals
- standardised forms, e.g. for consent procedures, ethical review, data management plans
- standards and protocols, e.g. data quality control standards, data transcription standards, confidentiality agreements for data handlers
- file sharing and storage procedures
- a security policy for data storage and transmission
- a data retention and destruction policy
- data copyright and ownership statements for the centre and for individual researchers
- standard data format recommendations
- version control and file naming guidelines
- information on funder requirements or policies on managing and sharing data that apply to projects or the centre
- a research data sharing strategy, e.g. via an institutional repository, data centre, website

Centralised data management is especially beneficial for data formatting, storage and back-up. It also helps to govern data sharing policies, establish copyright and IPR over data and assign roles and responsibilities.

DATA MANAGEMENT RESOURCES LIBRARY

A research hub can centralise all relevant data management and sharing resources for researchers and staff in a single location: on an intranet site, website, wiki, shared network drive or within a virtual research environment.

This resources library may contain relevant data policy and guidance documents, templates, tools and exemplars developed by the centre, as well as external policy and guidance resources or links to such resources. Resources can be developed by the centre. Good practices used by particular researchers or projects can be used as exemplars for others, so that these good practices can be shared.

A resources library might contain both locally created and external documents.

Locally created documents:

- declaration on copyright of research data and outputs
- declaration of institutional IT data management and existing back-up procedures
- statement on data sharing
- statement on retention and destruction of data
- file naming convention guidance
- version control guidance
- data inventory for individual projects
- template consent forms and information sheets which take data sharing into account
- example ethical review forms
- data anonymisation guidelines
- transcription guidance

External documents:

- research funder data policies
- research ethics guidance of professional bodies
- codes of practice or professional standards relevant to research data
- JISC Freedom of Information and research data: Questions and answers (2010)[39]
- Data Protection Act 1998 guidance
- Environmental Information Regulations 2004 information
- UK Data Archive's Guidance on Managing and Sharing Data
- Data Handling Procedures in Government[40]

DATA INVENTORY

A research centre can develop a data inventory to keep track of all data that are being created or acquired by various researchers within the centre.

The inventory can record what the data mean, how they are created, where they were obtained, who owns them, who has access, use and editing rights and who is responsible for managing them.

It can at the same time be used as a data management tool to plan management when research starts and then keep track of management implementation during the research cycle via a regular update strategy. The inventory can record storage and back-up strategies, keep track of versions and record quality control procedures. Data management can be reviewed annually or alongside research review such as during project progress meetings.

An inventory facilitates data sharing as it records data ownership, permissions for data sharing and contains basic descriptive metadata. It can be combined with the recording of research outputs for research excellence tracking purposes.

CASE STUDY

DATA MANAGING AND SHARING STRATEGIES

ESRC centres and programmes have a contractual responsibility to manage and share their research data.

The Rural Economy and Land Use (Relu) Programme undertakes interdisciplinary research between social and natural sciences. When it started it developed a programme-specific data management policy and set up a cross council data support service.

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.

Award holders are responsible for managing data throughout their research and making them available for archiving at established data centres after research. Preparing a data management plan at the start of a project helps in this respect, with pro-active advice and training given by the data support service.

The research councils are responsible for the long-term preservation and dissemination of the resulting data.

The Third Sector Research Centre (TSRC) is an ESRC-funded centre primarily shared between the Universities of Birmingham and Southampton.

The TSRC employs a centre manager to provide an organisational lead in its operation, including data management aspects such as co-ordinating ethical review of research, consent procedures and data re-use.

The centre has a data management committee of deputy director, selected research staff and centre manager, to devise and implement administrative aspects of data management throughout the centre's research. Administration staff help implement data management by providing transcription management services and data organisation support to projects.

REFERENCES

All online literature references retrieved 8 April 2011.

[1] Organisation for Economic Co-operation and Development (2007) *OECD principles and guidelines for access to research data from public funding*. www.oecd.org/dataoecd/9/61/38500813.pdf

[2] Rural Economy and Land Use Programme (2004) *Relu data management policy*. relu.data-archive.ac.uk/relupolicy.asp

[3] Publishing Network for Geoscientific and Environmental Data (2005) *Policy for the data library PANGAEA*. www.pangaea.de/curator/files/pangaea-data-policy.pdf

[4] Nature Publishing Group (2009) *Nature journals' policy on availability of materials and data*. www.nature.com/authors/policies/availability.html

[5] Biotechnology and Biological Sciences Research Council (2010) *BBSRC data sharing policy*. www.bbsrc.ac.uk/organisation/policies/position/policy/data-sharing-policy.aspx

[6] Economic and Social Research Council (2010) *ESRC research data policy*. www.esrc.ac.uk/about-esrc/information/data-policy.aspx

[7] Medical Research Council (n.d.) *MRC policy on data sharing and preservation*. www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm

[8] Natural Environment Research Council (2011) *NERC data policy*. www.nerc.ac.uk/research/sites/data/policy.asp

[9] Wellcome Trust (2010) *Wellcome Trust policy on data management and sharing*. www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

[10] Van den Eynden, V., Bishop, L., Horton, L. and Corti, L. (2010) *Data management practices in the social sciences*. www.data-archive.ac.uk/media/203597/datamanagement_socialsciences.pdf

[11] Rural Economy and Land Use Programme (n.d.). *Project communication and data management plan*. relu.data-archive.ac.uk/DMP2010.doc

[12] Rural Economy and Land Use Programme Data Support Service (2011) Example data management plans. relu.data-archive.ac.uk/DMPexample.asp

[13] UK Data Archive (2011) *Activity-based data management costing tool for researchers*. www.data-archive.ac.uk/media/257647/ukda_jiscdmcosting.pdf

[14] Digital Curation Centre (2010) DMP Online. www.dcc.ac.uk/dmponline

[15] SomnIA (n.d.) www.somnia.surrey.ac.uk/

[16] EDINA (n.d.) Go-Geo! GeoDoc metadata creator tool. www.gogeo.ac.uk/cgi-bin/cauth.cgi?context=editor

[17] UK Location (2011) UK Location Metadata Editor. location.defra.gov.uk/resources/discovery-metadata-service/metadata-editor/

[18] Association for Geographic Information (2010) *UK GEMINI specification for discovery metadata for geospatial data resources,* v.2.1, Cabinet Office, London. www.agi.org.uk/storage/standards/uk-gemini/GEMINI2_1_published.pdf

[19] Li, C. et al. (2010) 'BioModels Database: a database of annotated published models', *BMC Systems Biology*, 4(92). www.ebi.ac.uk/biomodels-main/

[20] Grimm, J. (2008) Wessex Archaeology Metric Archive Project (WAMAP). ads.ahds.ac.uk/catalogue/resources.html?abmap_grimm_na_2008

[21] Avondo, J. (2010) BioformatsConverter. cmpdartsvr1.cmp.uea.ac.uk/wiki/BanghamLab/index.php/BioformatsConverter

[22] Defra, ODPM, NAW, ONS and Countryside Agency (2004) Rural and Urban Area Classification 2004. www.statistics.gov.uk/geography/rudn.asp

[23] The London School of Economics and Political Science Library (2008) *Versions Toolkit for authors, researchers and repository staff*. http://www2.lse.ac.uk/library/versions/VERSIONS_Toolkit_v1_final.pdf

[24] Finch, L. and Webster, J. (2008) *Caring for CDs and DVDs.* NPO Preservation Guidance, Preservation in Practice Series, London: National Preservation Office. www.bl.uk/blpac/pdf/cd.pdf

[25] JISC (2004) Virtual research environment programme. www.jisc.ac.uk/whatwedo/programmes/vre.aspx

[26] UK Data Archive (2009) *Research ethics and legislation relevant to data sharing*. www.data-archive.ac.uk/create-manage/consent-ethics/legal

[27] UK Data Archive (2009) Consent forms. www.data-archive.ac.uk/create-manage/consent-ethics/consent

[28] Biological Records Centre (n.d.) www.brc.ac.uk

[29] UK Data Archive (2009) *Anonymisation*. www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation

[30] UK Biobank (2007) *UK Biobank ethics and governance framework*. www.ukbiobank.ac.uk/docs/EGFlatestJan20082.pdf

[31] Secure Data Service (n.d.) securedata.data-archive.ac.uk/

[32] Office for National Statistics (n.d.). Virtual Microdata Laboratory. www.ons.gov.uk/about/who-we-are/our-services/vml

[33] National Opinion Research Center (n.d.).. Technical assistance on metadata documentation. www.norc.org/DataEnclave

[34] CIFOR (2005). Forest Spatial Information Catalogue. Available at gislab.cifor.cgiar.org/fsic/index.htm

[35] Global Biodiversity Information Framework (n.d.) GBIF Data Portal. data.gbif.org/welcome.htm

[36] Padfield, T (2010) *Copyright for archivists and records managers,* 4th ed., London: Facet Publishing.

[37] Social and Environmental Conditions in Rural Areas (SECRA) (n.d.) www.sei.se/relu/secra/

[38] Ball, A (2011) How to Licence Research data. Digital Curation Centre. www.dcc.ac.uk/resources/how-guides/license-research-data

[39] Charlesworth, A. and Rusbridge, C. (2010) *Freedom of Information and research data: questions and answers*. www.jisc.ac.uk/publications/programmerelated/2010/foiresearchdata.aspx

[40] Cabinet Office (2008) *Data handling procedures in Government: final report*. www.cabinetoffice.gov.uk/sites/default/files/resources/final-report.pdf

DATA MANAGEMENT CHECKLIST

☑ Are you using standardised and consistent procedures to collect, process, check, validate and verify data?

☑ Are your structured data self-explanatory in terms of variable names, codes and abbreviations used?

☑ Which descriptions and contextual documentation can explain what your data mean, how they were collected and the methods used to create them?

☑ How will you label and organise data, records and files?

☑ Will you apply consistency in how data are catalogued, transcribed and organised, e.g. standard templates or input forms?

☑ Which data formats will you use? Do formats and software enable sharing and long-term validity of data, such as non-proprietary software and software based on open standards?

☑ When converting data across formats, do you check that no data or internal metadata have been lost or changed?

☑ Are your digital and non-digital data, and any copies, held in a safe and secure location?

☑ Do you need to securely store personal or sensitive data?

☑ If data are collected with mobile devices, how will you transfer and store the data?

☑ If data are held in various places, how will you keep track of versions?

☑ Are your files backed up sufficiently and regularly and are back-ups stored safely?

☑ Do you know what the master version of your data files is?

☑ Do your data contain confidential or sensitive information? If so, have you discussed data sharing with the respondents from whom you collected the data?

☑ Are you gaining (written) consent from respondents to share data beyond your research?

☑ Do you need to anonymise data, e.g. to remove identifying information or personal data, during research or in preparation for sharing?

☑ Have you established who owns the copyright of your data? Might there be joint copyright?

☑ Who has access to which data during and after research? Are various access regulations needed?

☑ Who is responsible for which part of data management?

☑ Do you need extra resources to manage data, such as people, time or hardware?

UK Data Archive
University of Essex
Wivenhoe Park
Colchester, CO4 3SQ

Email: datasharing@data-archive.ac.uk
Telephone: +44 (0)1206 872974

www.data-archive.ac.uk/create-manage

UK • DATA
ARCHIVE