MANAGING AND SHARING YOUR RESEARCH DATA

VEERLE VAN DEN EYNDEN
UK DATA ARCHIVE
UNIVERSITY OF ESSEX

Stirling University
19 March 2014
UK DATA ARCHIVE

• Curator of the largest collection of digital data in the social sciences and humanities in the UK
• Experience of supporting researchers and data creators of social science data and related disciplines
• We do data sharing for the ESRC Data Policy (since 1995)
• Our best practice approaches to making data shareable are based on:
  • challenges faced by researchers to share data
  • archiving research data – quantitative and qualitative

www.data-archive.ac.uk
OUR MANAGING AND SHARING DATA RESOURCES

- Online best practice guidance: ukdataservice.ac.uk/manage-data.aspx
- Managing and Sharing Research Data – a Guide to Good Practice: www.uk.sagepub.com/books/9781446267264 (SAGE Publications Ltd)
- Helpdesk for all queries: ukdataservice.ac.uk/help/get-in-touch.aspx
- Training programme
RESOURCES AT STIRLING

Policies:

- Record Retention Guidance (includes guidance for records other than research data): http://www.rec-man.stir.ac.uk/rec-ret/guidance.php
- Freedom of Information Act guidance: http://www.foi.stir.ac.uk/index.php
- Ethics: http://www.goodresearchpractice.stir.ac.uk/ethics/index.php
- Data protection: http://www.goodresearchpractice.stir.ac.uk/managing/protecting/data-protection.php

STORRE Stirling Online Research Repository https://dspace.stir.ac.uk//

Research data guidance: http://www.stir.ac.uk/is/researchers/data/

- planning for data
- keeping data safe
- sharing and reusing data
- archiving data
- support, advice, training
OTHER RESOURCES

Research Data Management Training MANTRA (Edinburgh) – online learning units
http://datalib.edina.ac.uk/mantra/

Digital Curation Centre:
• Data management planning http://www.dcc.ac.uk/resources/data-management-plans
• Data management training / courses http://www.dcc.ac.uk/training/data-management-courses-and-training
OVERVIEW FOR TODAY

- Policy landscape (funders, publishers) and data sharing
- Publishing and citing research data
- Storing and transferring data, incl. encryption and security
- Formatting and organising data
- Documenting data

Lunch
- Legal, ethical issues in handling and sharing data, incl. informed consent and anonymisation
- Rights related to research data
- Writing and implementing a data management plan
PUBLICLY FUNDED RESEARCH DATA ARE A PUBLIC GOOD, PRODUCED IN THE PUBLIC INTEREST, THAT SHOULD BE MADE OPENLY AVAILABLE WITH AS FEW RESTRICTIONS AS POSSIBLE IN A TIMELY AND RESPONSIBLE MANNER THAT DOES NOT HARM INTELLECTUAL PROPERTY

Topics:

• Research funder policies
• Journal publisher policies
Publicly funded research data are a public good, produced in the public interest, that should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.

- in accordance with relevant standards and community best practice
- metadata to make research data discoverable
- legal, ethical, commercial constraints on release of research data
- recognition for collecting & analysing data; limited privileged use
- acknowledge sources of data, intellectual contributions, terms & conditions
- use public funds to support the management and sharing of publicly-funded research data

Research Councils UK Common Principles on Data Policy (2011)
RESEARCH FUNDER POLICIES (RCUK)

• peer reviewed research papers published in journals that are compliant with Research Council policy on Open Access
• include statement on how the underlying research materials such as data, samples or models can be accessed
• for publications submitted for publication from 1 April 2013

RESEARCH FUNDER DATA POLICIES (RCUK)

Research Councils:

• data sharing policy mandating or encouraging data sharing
• data management / sharing planning required
• award holders responsible for managing & sharing data, except EPSRC
• fund data sharing support services and infrastructure
e.g. UK Data Service (ESRC)
  NERC data centres (NERC)
  MRC Data Support Service (MRC)
  Atlas Petabyte Storage (STFC)
  Archaeology Data Service (AHRC)
<table>
<thead>
<tr>
<th>Funder</th>
<th>Plan required?</th>
<th>Required at application</th>
<th>Data topics in DMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts and Humanities Research Council (AHRC)</td>
<td>Yes</td>
<td>Technical plan</td>
<td>Standards, preservation, continued access and use</td>
</tr>
<tr>
<td>Biotechnology and Biological Sciences Research Council (BBSRC)</td>
<td>Yes</td>
<td>Data management and sharing plan</td>
<td>Type, format, standards, sharing methods, restrictions, sharing timeframe</td>
</tr>
<tr>
<td>Cancer Research UK (CRUK)</td>
<td>Yes</td>
<td>Data sharing plan</td>
<td>Volume, format, standards, metadata, documentation, sharing method, timescale, preservation, restrictions</td>
</tr>
<tr>
<td>Department for International Development (DFID)</td>
<td>Yes</td>
<td>Access and data management plan</td>
<td>Repositories, limits, timescale, responsibilities, resources, access strategy</td>
</tr>
<tr>
<td>Engineering and Physical Sciences Research Council (EPSRC)</td>
<td>No</td>
<td>Policy framework at institutional level (from 2015)</td>
<td></td>
</tr>
<tr>
<td>Economic and Social Research Council (ESRC)</td>
<td>Yes</td>
<td>Data management plan</td>
<td>Volume, type, quality, archiving plans, difficulties sharing, consent sharing, IPR, responsibilities</td>
</tr>
<tr>
<td>Medical Research Council (MRC)</td>
<td>Yes</td>
<td>Data management plan</td>
<td>Collection methods, documentation, standards, preservation, curation, security, confidentiality, sharing and access, timescale, responsibilities</td>
</tr>
<tr>
<td>Natural Environment Research Council (NERC)</td>
<td>Yes</td>
<td>Outline data management plan</td>
<td>Data management procedures, created data</td>
</tr>
<tr>
<td>Science and Technology Facilities Council (STFC)</td>
<td>Yes</td>
<td>Data management plan</td>
<td>Type, preservation, metadata, value, sharing, timescale, resources needed</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>Yes</td>
<td>Data management and sharing plan</td>
<td>What data, when share, where share, how access, limits, how preserve, what resources</td>
</tr>
</tbody>
</table>

UK DATA ARCHIVE
EPSRC POLICY FRAMEWORK ON RESEARCH DATA

Research organisations receiving EPSRC funding responsible:

• publish metadata online, with DOI (digital object identifier)
• maintain data securely for 10 years
• institutional roadmap for compliance May 2012
• institutional policy implemented May 2015
• papers to include statements on access to supporting data
• expectations follow RCUK principles

• influenced by Freedom of Information Act
• may in future influence policies of other RCs

EPSRC Policy Framework on Research Data
European open access policies: Horizon 2020, European Research Council (ERC)

- communication & recommendation on access to / preservation of scientific information (July 2012) (publications & research data)
- pilot on open access to research data, primarily data underlying (open access) scientific publications for Horizon 2020
- FAQ open access to publications & data in Horizon 2020
- data management guidelines for Horizon 2020 (~ policies)

generally based on OECD Principles and Guidelines for Access to Research Data from Public Funding
Many science journals (Science, Nature, PLOS ONE etc.) have data policies relating to data sharing

“PLOS ONE will not consider a study if the conclusions depend solely on the analysis of proprietary data” … “the paper must include an analysis of public data that validates the conclusions so others can reproduce the analysis.”

- data underpinning publication accessible
  - upon request from author
  - as supplement with publication
  - in public repository
  - in mandated repository (e.g. PANGAEA – Elsevier)
- citation via unique DOIs
  - e.g. BioMed Central open data statement
DATA SHARING – A SHARED RESPONSIBILITY

• Funders: policies, mandates and some infrastructure funding

• Funded researchers: create, manage and use data

• Departments/centres: provide local support and some infrastructure

• Institutions: provide a supporting framework
  • grant-application and funding support
  • research integrity framework
  • IT and data storage facilities
  • Data management guidance and training
Topics:
- Where publish /deposit research data
- Persistent identifiers for citation
WHERE SHARE / PUBLISH RESEARCH DATA

- Institutional repository
- Discipline specific repositories and data centres dedicated to archiving, preserving and disseminating discipline specific digital data
- Generic repositories: Dryad, figshare, Dataverse,....
- Registries of data repositories (to find suitable repository):
  - databib.org/
  - www.datacite.org/repolist/
  - re3data
- Data journal
- Data submissions to journals
  E.g.: Nature, BMJ

Source: Databib 2012
UK DATA CENTRES

- Antarctic Environmental Data Centre
- Archaeology Data Service
- Biomedical Informatics Research Network Data Repository
- British Atmospheric Data Centre
- British Library National Sound Archive
- British Oceanographic Data Centre
- Cambridge Crystallographic Data Centre
- Endangered Languages Archive
- Environmental Information Data Centre
- European Bioinformatics Institute
- Geospatial Repository for Academic Deposit and Extraction
- National Biodiversity Network
- National Geoscience Data Centre
- NERC Earth Observation Data Centre
- NERC Environmental Bioinformatics Centre
- Publishing Network for Geoscientific and Environmental Data (PANGAEA)
- The Oxford Text Archive
- UK Data Archive
- UKDS ReShare (April 2014)
- UK Solar System Data Centre
- Visual Arts Data Service
INSTITUTIONAL REPOSITORIES

• Almost all universities now have an institutional repository where researchers can put their research outputs
• Generally designed for articles/publications, but increasingly providing for datasets too

• Project underway to develop a national registry for research data, which will bring together data collections from across national data centres and institutional data repositories (Jisc, DCC, UKDA)
DATA JOURNALS

• A fairly new phenomenon, but growing
• Publish a detailed journal style data article describing the data, how it was collected, reuse potential
• Recommends or provides a place of deposit (repository)
• Academic credit

Examples:
  • Journal of Open Public Health Data
  • Journal of Open Psychology Data
  • Journal of Open Archaeology Data
OPEN ACCESS

- Open access – freely available to anyone without restrictions on access
- Increasingly standard for academic publications
- Openness does not mean no credit – still citation
- Not all research data appropriate for sharing openly

- Open access pilot for data from EU Horizon 2020 projects
- Otherwise few mandates for open access data
MAXIMUM IMPACT FOR DATA SHARING

- Make your data citable through a persistent identifier, e.g. DataCite DOIs
- Identify yourself as unique, e.g. ORCID
Understanding Society: Waves 1-2, 2009-2011

SN: 6614
Title: Understanding Society: Waves 1-2, 2009-2011
Alternative title: United Kingdom Household Longitudinal Study; UKHLS
Persistent identifier: 10.5255/UKDA.SN.6614.4
Series: Understanding Society: Waves 1-, 2008-
Depositor: University of Essex. Institute for Social and Economic Research
Principal Investigator(s): University of Essex. Institute for Social and Economic Research
Investigator(s): NatCen Social Research
Data collector(s): NatCen Social Research
Sponsor(s): Economic and Social Research Council
Department for Work and Pensions
Department for Education
Department for Transport
Department for Culture, Media and Sport
Department for Communities and Local Government
Department of Health
Scotish Government
Welsh Assembly Government
Northern Ireland Executive
Department for Environment, Food and Rural Affairs
Food Standards Agency
Grant number: RES-586-47-0001-01
DATA CITATION at UK DATA ARCHIVE

Health Survey for England, 2009
UKDA study number:6732

Principal Investigators
National Centre for Social Research
University College London, Department of Epidemiology and Public Health

Sponsor
Information Centre for Health and Social Care

Distributed by
UK Data Archive, University of Essex, Colchester.
July 2011 (2nd Edition)

Bibliographic Citation
All works which use or refer to these materials should acknowledge those sources by means of bibliographic citation. To ensure that such source attributions are captured for bibliographic indexes, citations must appear in footnotes or in the reference section of publications. The bibliographic citation for this data collection is:

Acknowledgement
Any publication, whether printed, electronic or broadcast, based wholly or in part on these materials, should acknowledge the original data creators, depositors or copyright holders, the funders of the Data Collections (if different) and the UK Data Archive, and to acknowledge Crown Copyright where appropriate.

Any publication, whether printed, electronic or broadcast, based wholly or in part on these materials should carry a statement that the original data creators, depositors or copyright holders, the funders of the Data Collections (if different) and the UK Data Archive bear no responsibility for their further analysis or interpretation.

Copyright:
Crown copyright material is reproduced with the permission of the Controller of HMSO and the Queen’s Printer for Scotland

Disclaimer
Although all efforts are made to ensure the quality of the materials, neither the original data creators, depositors or copyright holders, the funders of the Data Collections, nor the UK Data Archive bear any responsibility for the accuracy or comprehensiveness of these materials.

All rights reserved. No part of these materials may be reproduced, stored in, or introduced into a retrieval system, or transmitted, in any form, or by any means (electronic, mechanical, photocopying, recording or otherwise) without the prior written permission of the UK Data Archive.

UK Data Archive
University of Essex
Wivenhoe Park
Colchester
Essex CO4 3SQ
United Kingdom
www.data-archive.ac.uk

6732, Health Survey for England, 2009
(HSE)

Depositor:
National Centre for Social Research

Principal Investigators:
National Centre for Social Research
DOI at UK DATA ARCHIVE

10.5255/UKDA-SN-6732-1

- Archive unique identifier
- Archive readable identifier
- Resource identifier prefix
- Resource identifier
- Resource version
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.

Topics:
- Data storage
- Backing up
- Data security
- Encryption
- File sharing
- Version control
- Data disposal
LOCAL DATA STORAGE

- All digital media are fallible
- Optical (CD, DVD) & magnetic media (hard drives, tape) degrade – lifespans even lower if kept in poor conditions
- Physical storage media become obsolete e.g. floppy disks

- Copy data files to new media two to five years after first created
Example: robocopy/rsync synchronising

- Synchronise files between two computers using ‘robocopy’
- Alternative for Mac OSX called ‘rsync’

```
call robocopy <source> <target> /MIR /Z /W:5
```

- Options (or ‘switches’)
  - `/MIR = mirror`
  - `/Z = resume large files if broken`
  - `/W:5 = wait time between retries`
Example: SyncToy synchronising

- Similar features but more user friendly
- Free download from Microsoft
STORAGE AND FILE SHARING SERVICES

- Online or ‘cloud’ services increasingly popular
  - Google Drive, DropBox, Microsoft SkyDrive etc.
  - Accessible anywhere
  - Background syncing
  - Mirror files
  - Mobile apps available
  - Very convenient

- Everyone uses them, and that’s ok BUT precautions must be taken..
- Consider if appropriate, as services can be hosted outside the EU (DPA for personal data)
- Encrypt anything sensitive or avoid services altogether
OTHER STORAGE AND FILE SHARING OPTIONS

• Institutional / departmental drive with access for external researchers e.g. via virtual private network (VPN)

• Shared secure file transfer protocol (FTP) server

• Internally managed ‘cloud’ e.g. ownCloud, ZendTo

• Data repository or archive
  • A repository acts as more of a ‘final destination’ for data
  • Institutional repository e.g. SOAS Research Online
  • UK Data Service has it’s own launching shortly, called ‘ReShare’ for social science data of any kind
BACKING-UP DATA

• Protect against: software failure, hardware failure, malicious attack, natural disasters

• Back-ups are additional copies that can be used to restore originals

• What is your back-up strategy: institutional or personal
  • where? - original copy, external local and remote copies
  • what media? - CD, DVD, external hard drive, tape, etc.
  • how often?
  • for how long is it kept?
VERIFICATION AND INTEGRITY CHECKS

- Ensure that your backup method is working as intended
- Be wary when using sync tools in particular
  - Mirror in the wrong direction and you could lose new files completely
- Applies to online DropBox-like syncing services too

- You can use checksums to verify the integrity of a backup
- Also useful when transferring files
- Checksum somewhat like a files ‘fingerprint’
- … that changes when the file changes
VERSION CONTROL

- Keep track of different copies or versions of data files
  - Useful for files kept in multiple locations
  - Or which have multiple users
  - A way to safeguard against accidental changes

- File names are a good way to do this
  - Unique descriptive names for files
  - Include date and/or version number in name
  - Indicate relationships between files

  e.g. *FoodInterview_1_draft; FoodInterview_1_final; HealthTest_06-04-2008; BGHSurveyProcedures_00_04*
Example: version control table

<table>
<thead>
<tr>
<th>Version</th>
<th>Responsible</th>
<th>Notes</th>
<th>Last amended</th>
</tr>
</thead>
<tbody>
<tr>
<td>00_05</td>
<td>Sally Watsley</td>
<td>Version 00_03 and 00_04 compared by SW</td>
<td>25/11/2007</td>
</tr>
<tr>
<td>00_04</td>
<td>Vani Yussu</td>
<td>Entries checked by VY, independent from previous</td>
<td>17/10/2007</td>
</tr>
<tr>
<td>00_03</td>
<td>Steve Knight</td>
<td>Entries checked by SK</td>
<td>29/07/2007</td>
</tr>
<tr>
<td>00_02</td>
<td>Karin Mills</td>
<td>Test results 81-120 entered</td>
<td>05/07/2007</td>
</tr>
<tr>
<td>00_01</td>
<td>Karin Mills</td>
<td>Test results 1-80 entered</td>
<td>04/07/2007</td>
</tr>
</tbody>
</table>
Example: Google Drive version control

- Collaboratively edit documents in ‘the cloud’ while tracking version history
ENCRIPTION

Always encrypt personal or sensitive data
Encrypt anything you would not send on a postcard
  • for moving files e.g. transcripts
  • for storing files e.g. shared areas, mobile devices
Basic principles
  • use an algorithm to transform information (A=1)
  • need a ‘key’ to decrypt
Free softwares that are easy to use
  • Safehouse
  • Truecrypt
  • Axcrypt
These softwares
  • encrypt hard drives, partitions, files and folders
  • encrypt portable storage devices such as USB flash drives
DATA DESTRUCTION

When you delete data and documentation from a hard drive: it is probably not gone

- files need to be overwritten to ensure they are irretrievably deleted:
  - BCWipe - uses ‘military-grade procedures to surgically remove all traces of any file’
  - Axcrypt
- if in doubt, physically destroy the drive using an approved secure destruction facility
- physically destroy portable media, as you would shred paper
DATA SECURITY

Protect data from unauthorised access, use, change, disclosure and destruction

Personal data need more protection – always keep separate

Control access to computers

- passwords
- anti-virus and firewall protection, power surge protection
- networked vs non-networked PCs
- all devices: desktops, laptops, memory sticks, mobile devices
- all locations: work, home, travel
- restrict access to sensitive materials e.g. consent forms, patient records

Proper disposal of equipment (and data)

- even reformatting the hard drive is not sufficient

Control physical access to buildings, rooms, cabinets
DEMO: DATA ENCRYPTION

Create an encrypted storage space using free software SafeHouse

www.data-archive.ac.uk/media/312652/storingyourdata_encryptionexercise.pdf
Calculate the MD5 checksum value of a file to check its integrity, e.g. after back-up

www.data-archive.ac.uk/media/361550/storingyourdata_checksumexercise.pdf
DEMO: VERSIONING

Keep track of different versions of documents in MS Word

www.data-archive.ac.uk/media/375814/formattingyourdata_versioningexercise.pdf
Synchronise files between two folders using SyncToy software

www.data-archive.ac.uk/media/375817/formatingyourdata_synchronisingexercise.pdf
DISCUSSION & QUIZ
FORMATTING & ORGANISING 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.

Topics:
- File formats
- File conversions
- Organising files and folders
- File naming
- Quality control
CAN YOU UNDERSTAND/USE THESE DATA?

SrvMthdDraft.doc
SrvMthdFinal.doc
SrvMthdLastOne.doc
SrvMthdRealVersion.doc
Choice of software format for digital data:
• planned data analyses
• software availability
• hardware used – e.g. audio
• discipline-specific standards and customs

Digital data = software dependent

Digital data endangered by obsolescence of software/hardware

Best formats for long-term preservation - standard formats, interchangeable formats, open formats
  e.g. tab-delimited, comma-delimited (CSV), ASCII, RTF, PDF/A, OpenDocument format, SPSS portable, XML

UK Data Archive optimal file formats for various data types
FILE FORMAT CONVERSIONS

Convert data for preservation or back-up: export, save as

Beware of conversion errors losses:

• loss of internal metadata
  
  e.g. convert MS Access to tab-delimited tables

• loss of editing, formatting, formulae
  
  e.g. convert MS Word to RTF

• truncation or loss of data
  
  e.g. string variables lost in SPSS – Stata conversion; MS Access memo fields truncated in conversion to CSV

Check for errors and changes after conversion
EXAMPLE: FORMAT CONVERSION

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Timber volumes in m³</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Logging private land</td>
<td>20346.345</td>
<td>47005.223</td>
<td>26001.754</td>
<td>11468897</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>Logging forest reserves</td>
<td>4060.567</td>
<td>1777.783</td>
<td>804.997</td>
<td>0.000</td>
<td>3329.653</td>
</tr>
<tr>
<td>6</td>
<td>Logging state land</td>
<td>0.000</td>
<td>1200.000</td>
<td>569.162</td>
<td>2077.567</td>
<td>356.935</td>
</tr>
<tr>
<td>7</td>
<td>Total</td>
<td>61119.912</td>
<td>87065.006</td>
<td>64802.913</td>
<td>51354.464</td>
<td>5686.588</td>
</tr>
</tbody>
</table>

MS Excel format

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Timber volumes in m³</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Logging private land</td>
<td>20346.345</td>
<td>47005.223</td>
<td>26001.754</td>
<td>11468897</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>Logging forest reserves</td>
<td>4060.567</td>
<td>1777.783</td>
<td>804.997</td>
<td>0.000</td>
<td>3329.653</td>
</tr>
<tr>
<td>6</td>
<td>Logging state land</td>
<td>0.000</td>
<td>1200.000</td>
<td>569.162</td>
<td>2077.567</td>
<td>356.935</td>
</tr>
<tr>
<td>7</td>
<td>Total</td>
<td>61119.912</td>
<td>87065.006</td>
<td>64802.913</td>
<td>51354.464</td>
<td>5686.588</td>
</tr>
</tbody>
</table>

Tab–delimited text format

Loss of annotation
ORGANISING DATA

Plan in advance how best to organise data

Examples

• hierarchical structure of files, grouped in folders, e.g. images
• survey data – spreadsheet, SPSS, relational database
• interview transcripts - individual well-named files
FILE NAMING

- file name = principal identifier of file
- logical naming - easy to identify, locate, retrieve, access
- naming provides organisation, context & consistency
- name elements: version nr, date, content description, creator name

Best practice
- name independent of location
- brief & relevant
- no special characters, dots or spaces
- for separation use underscores _
- versioning via filename: ordinal and decimal version numbers
- use names to classify broad types of files
- avoid very long file names
EXAMPLE DIRECTORY STRUCTURE

- ENBIOproject
  - Data
    - Databases
      - ConsumerSurvey
      - StakeholderNetworkAnalysis
      - StakeholderSurvey
    - Images
      - FocusGroupImages
      - LandscapeImages
    - Models
    - Sound
      - FocusGroupRecordings
      - InterviewRecordings
    - Text
      - FocusGroupTranscripts
      - InterviewTranscripts
  - Documentation
    - ConsentForms
      - CF_FocusGroups
      - CF_Interviews
    - InformationSheets
      - IS_ConsumerSurvey
      - IS_FocusGroups
      - IS_Interviews
      - IS_StakeholderSurvey
  - Methodology
    - Method_ConsumerSurvey
    - Method_FocusGroups
    - Method_Interviews
    - Method_Modelling
    - Method_SNA
    - Method_StakeholderSurvey
QUALITY CONTROL

Data quality control at various stages during research process:

1. data collection – data must reflect facts, responses, observations, events
   - instrument calibration
   - multiple observations/measurements
   - expert opinion
   - standardised protocols
   - computer assisted interviews
   - interview recording and transcription
QUALITY CONTROL

2. data entry, digitisation, transcription and coding – avoid errors - use standardised and consistent procedures

- validation rules for data entry
- data entry screens
- double entry
- variable and record labelling to avoid confusions
- missing value coding
- controlled vocabularies or choice lists to minimise manual data entry
- purpose-built database structure to organise data and data files
3. data checking and verifying - automated and/or manual

- check for out-of-range values
- check data completeness
- statistical analyses to detect errors or anomalous values (descriptives, frequencies, means, range, clustering)
- peer review of data
DOCUMENTING & CONTEXTUALISING DATA

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

Topics:
• Documenting data
• Study-level documentation and context
• Data-level documentation
WHY DOCUMENT DATA?

- Enables you to understand/interpret data
- Needed to make data independently understandable
- Ensures informed and correct use, reduces chance of incorrect use/misinterpretation
- If using your data for the first time, what would you need to know?

- The UK Data Archive and other repositories use data documentation to:
  - create/supplement catalogue record for dataset
  - create user guide(s) and data listing for dataset
  - ensure accurate processing and archiving
WHAT SHOULD BE CAPTURED?

Wider contextual information about project and data
- background, project history, aims, objectives, hypotheses
- publications based on dataset

Data collection methodology and processes
- sampling
- data collection process - fieldwork, interviewer instructions
- instruments used - questionnaires, showcards, interview schedules
- temporal/geographic coverage
- data validation - cleaning, error-checking
- derived variables – compilation
- weighting: factors and variables, weighting process
- secondary data sources used

Useful documents are:
- final report, published reports, user guide, working paper, publications, lab books
WHAT SHOULD BE CAPTURED?

Information on dataset structure
• data files
• relationships between files
• records, cases…
• data listing of attributes for interviewees (quali data)

Variable-level documentation
• labels, codes, classifications
• missing values
• derivations and aggregations
WHAT SHOULD BE CAPTURED?

Data confidentiality, access and use conditions
- anonymisation carried out
  - aggregation, banding, coding and top-coding, disclosure control
  - editing of sensitive material in interview transcripts
- consent conditions/procedures
- access or use conditions of data
LABELLING

- All structured, tabular data should have cases or records and variables adequately documented with
- Names, labels and descriptions for all variables, fields, records and their values
- Variable names
  - question number system related to questions in a survey/questionnaire
    e.g. Q1a, Q1b, Q2, Q3a
  - numerical order system
    e.g. V1, V2, V3
  - meaningful abbreviations or combinations of abbreviations referring to meaning of the variable
    e.g. oz%=percentage ozone, GOR=Government Office Region, moocc=mother occupation, faocc=father occupation
  - for interoperability across platforms - variable names max 8 characters without spaces (absolute maximum is 32 characters)
CODE LABELLING

• Code labels
  • brief, max. 80 characters
  • unit of measurement
  • reference the question number of a survey or questionnaire
    e.g. variable 'p1sex' = 'sex of respondent' with codes '1=female', '2=male', '-8=don't know', '-9=not answered'
    e.g. variable 'q11hexw' with label 'Q11: hours spent taking physical exercise in a typical week' - the label gives the unit of measurement and a reference to the question number (Q11b)

• Codes of, and reasons for, missing data
  • avoid blanks, system-missing or '0' values
    e.g. '99=not recorded', '98=not provided (no answer)', '97=not applicable', '96=not known', '95=error'

• Coding or classification schemes used, with a bibliographic ref
  e.g. Standard Occupational Classification 2000 - a list of codes to classify respondents' jobs; ISO 3166 alpha-2 country codes - an international standard of 2-letter country codes
DATA-LEVEL DOCUMENTATION

• Embed annotations in data files:
  • Quantitative data: variable/value labels; worksheet information; table relationships and queries
  • Qualitative data/text documents:
    • Interview transcript speech demarcation (speaker tags)
    • Document header with brief details of interview date, place, interviewer name, interviewee details, context

• Examples (see visual screenshots):
  • SPSS: variable attributes documented in Variable View (label, code, data type, missing values)
  • MS Access: variable descriptions and attributes documented in Design View; relationships
  • ArcGIS: shapefiles (layers) and tables in geodatabase; metadata created in ArcCatalog
  • MS Excel: base worksheet data-related documentation
METADATA

- core metadata for archived datasets, e.g. UK Data Archive:
  - dataset title
  - name, institution and contact details of data owner/researcher
  - abstract
  - keywords/subject categories, indexed using in-house thesaurus
  - funding source and award number
  - temporal coverage (data collection start and end dates)
  - geographic coverage (country, region, longitude/latitude)
  - data availability/access conditions
  - copyright holder
  - data parameters: type, format, sample size/units
  - publications
  - access conditions
  - DOI

- at UK Data Archive: DDI metadata, rich detailed content, enables sharing (OAI)
HOW TO CREATE METADATA FOR DATA

- metadata: through data deposit forms/tools

- Data Documentation Initiative (DDI) documentation derived from software packages using DDI tools
  tools.ddialliance.org/

  - www.ddialliance.org/codebook/index.html
  - metadata tools: tools.ddialliance.org
  - e.g. SPSS to DDI, use Nesstar Publisher
    www.nesstar.com/software/publisher.html
  - German Institute for Educational Progress (IQB) – educational data codebooks
    www.iza.org
TIPS FOR DESCRIBING DATA WELL

- good data documentation and metadata enables easy discovery of data – quality depends on what the depositor can provide
- provide meaningful information (titles, descriptions, abstract, keywords)
- submitting good documentation and metadata enables data archives/data centres to create rich and detailed metadata
- good data documentation leads to good data re-use
- detailed yet comprehensive documentation
- where possible integrate within dataset
EXAMPLES OF DOCUMENTATION FOR RE-USE

- Quantitative dataset
  - documentation - questionnaire, variable list, codebook etc.

- Qualitative dataset – depends on size and scale
  - user guide, data listing
QUANTITATIVE STUDY

Smaller-scale study - user guide may just contain survey questionnaire, methodology information
Example from Health Survey for England 2007 – documents separated, bigger study

**Documentation:**

<table>
<thead>
<tr>
<th>Format</th>
<th>Name</th>
<th>Size in Kilo Bytes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF</td>
<td>6112datadocs.pdf</td>
<td>808</td>
<td>Lists of Variables and Derived Variables</td>
</tr>
<tr>
<td>PDF</td>
<td>6112interviewingdocs.pdf</td>
<td>1655</td>
<td>Questionnaires, Showcards, Coding Frames and Consent Booklet</td>
</tr>
<tr>
<td>PDF</td>
<td>6112supportingdocs.pdf</td>
<td>1573</td>
<td>Interviewer, Nurse, Coding and Editing Instructions</td>
</tr>
<tr>
<td>PDF</td>
<td>6112userguide.pdf</td>
<td>288</td>
<td>User Guide</td>
</tr>
<tr>
<td>HTML</td>
<td>UKDA_Study_6112_Information.htm</td>
<td>25</td>
<td>Study information and citation</td>
</tr>
</tbody>
</table>
QUALITATIVE STUDY

User guide contains variety of documents that provide context

R000239794 - The Market for Migrant Domestic and Sex Workers

This project, led by the University of Nottingham and COMPAS, University of Oxford, gathers data on the demand for migrant domestic and sex workers in Spain and the UK.

Key Findings

Demand for and attitudes toward migrant labour

- Demand for sex-workers and domestic workers is very closely related to supply. Where sex and domestic workers are easily and cheaply available
Data listing provides an at-a-glance summary of data collection

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Gender</th>
<th>Nationality</th>
<th>Income</th>
<th>Marital Status</th>
<th>Ethnicity</th>
<th>Interview</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pep</td>
<td>36</td>
<td>Male</td>
<td>Spanish/Catalan</td>
<td>Medium</td>
<td>Single</td>
<td>White</td>
<td>Face-to-face transcript</td>
<td>English translation</td>
</tr>
<tr>
<td>Jordi</td>
<td>34</td>
<td>Male</td>
<td>Spanish/Catalan</td>
<td>Medium</td>
<td>Single</td>
<td>White</td>
<td>Face-to-face transcript</td>
<td>English translation</td>
</tr>
<tr>
<td>Erik</td>
<td>38</td>
<td>Male</td>
<td>French</td>
<td>Medium</td>
<td>Divorced (co-habit)</td>
<td>White</td>
<td>Face-to-face transcript</td>
<td>English translation</td>
</tr>
<tr>
<td>Enrique</td>
<td>35</td>
<td>Male</td>
<td>Spanish/Catalan</td>
<td>Medium/high</td>
<td>Married</td>
<td>White</td>
<td>Face-to-face transcript</td>
<td>English translation</td>
</tr>
<tr>
<td>Sergio</td>
<td>31</td>
<td>Male</td>
<td>Spanish</td>
<td>Medium</td>
<td>Single</td>
<td>White</td>
<td>Face-to-face transcript</td>
<td>English translation</td>
</tr>
<tr>
<td>~</td>
<td>20</td>
<td>Male</td>
<td>British (Guernsey)</td>
<td>Medium</td>
<td>Single</td>
<td>White</td>
<td>Notes</td>
<td>English</td>
</tr>
<tr>
<td>Mr and Mrs</td>
<td>~</td>
<td>Male and Female</td>
<td>~</td>
<td>~</td>
<td>Married</td>
<td>~</td>
<td>Face-to-face transcript/notes</td>
<td>English</td>
</tr>
<tr>
<td>K and D</td>
<td>~</td>
<td>Male and Female</td>
<td>~</td>
<td>~</td>
<td>Married</td>
<td>~</td>
<td>Notes</td>
<td>English</td>
</tr>
<tr>
<td>~</td>
<td>37</td>
<td>Female</td>
<td>~</td>
<td>High</td>
<td>Married</td>
<td>White</td>
<td>Face-to-face transcript</td>
<td>English (some Spanish)</td>
</tr>
</tbody>
</table>
EXERCISES & QUIZ
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

Topics:

- Legal and ethical aspects
- Informed consent for data sharing
- Anonymising data
- Controlling access to data
ETHICAL ARGUMENTS FOR ARCHIVING DATA

• store and protect data securely
• not burden over-researched, vulnerable groups
• make best use of hard-to-obtain data (e.g., elites, socially excluded, over-researched)
• extend voices of participants
• provide greater research transparency
• enable fullest ethical use of rich data

In each, ethical duties to participants, peers and public may be present
DUTY OF CONFIDENTIALITY AND DATA SHARING

• Duty of confidentiality exists in common law and may apply to research data

• If participant consents to share data, then sharing does not breach confidentiality

• Public interest can override duty of confidentiality; best practice is to avoid vague or general promises in consent forms
DATA PROTECTION ACT, 1998

- Personal data:
  - relate to living individual
  - individual can be identified from those data or from those data and other information
  - include any expression of opinion about the individual

- Requirements for handling personal data
  - processed fairly and lawfully
  - obtained and processed for a specified purpose
  - adequate, relevant and not excessive for the purpose
  - accurate
  - not kept longer than necessary
  - processed in accordance with the rights of data subjects, e.g. right to be informed about how data will be used, stored, processed, transferred, destroyed, …; right to access info and data held
  - kept secure
  - not transferred abroad without adequate protection

- Only disclosed if consent has been given to do so (except legal duty)
Exceptions for personal data collected as part of research:

- can be retained indefinitely (if needed)
- can be used for other purposes in some circumstances
- people should still be informed

If data are anonymised (personal identifiers removed) then DP laws will not apply as these no longer constitute ‘personal data’

DPA is not intended to, and does not, inhibit ethical research
SENSITIVE DATA

• Data regarding an individual's race or ethnic origin, political opinion, religious beliefs, trade union membership, physical or mental health, sex life, criminal proceedings or convictions (DPA 1998)

• Can only be processed for research purposes if:
  • explicit consent (ideally in writing) has been obtained; or
  • medical research by a health professional or equivalent with duty of confidentiality; or
  • analysis of racial/ethnic origins for purpose of equal opportunities monitoring; or
  • in substantial public interest and not causing substantial damage and distress
OPTIONS FOR SHARING CONFIDENTIAL DATA

Researchers to consider

• obtaining informed consent, also for data sharing and preservation / curation
• protecting identities e.g. anonymisation, not collecting personal data
• restricting / regulating access where needed (all or part of data) e.g. by group, use, time period
• securely storing personal or sensitive data

Consider jointly and in dialogue with participants

Plan early in research
INFORMED CONSENT FOR ETHICAL PURPOSES

• What does it mean for consent to be “informed”?  
  • purpose of the research  
  • what is involved in participation  
  • benefits and risks  
  • mechanism of withdrawal  
  • data uses – primary research, storing, processing, re-use, sharing, archiving,…  
  • strategies to ensure confidentiality of data where this is relevant – anonymisation, access restrictions…

• RCUK also expects data to be accessible for other uses  
  RCUK Policy on Access to Research Outputs

• Now a requirement for ESRC awards:  
  “Where research data are considered confidential or contain sensitive personal data, award holders must seek to secure consent for data sharing or alternatively anonymise the data in order to make sharing possible.”  
  ESRC Research Data Policy 2010 2.4(32)
DO PARTICIPANTS CONSENT TO SHARE DATA?

- Timescapes
  - data on personal relationships
  - 95%+ consent rate
- Foot and mouth disease in N. Cumbria
  - sensitive community information
  - UK Data Archive consultation; pilot with 4 participants
  - 40/54 interviews; 42/54 diaries; audio restricted
- Finnish research on consent
  - Re-contact project: life stores, gender, etc.
  - 165/169 (98%) agreed
- Bereaved relatives want others to benefit from their data
INFORMED CONSENT FOR UNKNOWN FUTURE USES

• In fact, a great deal of information can be provided
  • who can access the data – only *bona fide* researchers
  • purposes – research or teaching or both
  • confidentiality protections, undertakings of future users
  • general consent (similar to consent with emergent research topics)

• Medical research and biobank models – enduring, broad, open consent
  • no time limits; no recontact required
  • unspecified hypotheses and procedures
  • 99% consent rate (2500+ patients) – Wales Cancer Bank

*ESRC expects that others will also use it [data], so consent should be obtained on this basis and the original researcher must take into account the long-term use and preservation of data. (ESRC Framework for Research Ethics, 1.17.5.1)*
CONSENT NEEDED ACROSS THE DATA LIFE CYCLE

- Engagement in the research **process**
  - decide who approves final versions of transcripts

- **Dissemination** in presentations, publications, the web
  - decide who approves research outputs

- Data **sharing** and archiving
  - consider future uses of data

Always dependent on the research context

[UK Data Archive model consent form](#)
A GOOD INFORMATION SHEET & CONSENT FORM

- Meets requirements of Data Protection laws
  - purpose of the research
  - what is involved in participation
  - benefits and risks
  - mechanism of withdrawal
  - usage of data – for primary research and sharing
  - strategies to ensure confidentiality of data (anonymisation, access,....) where this is relevant

- Simple

- Avoids excessive warnings

- Complete for all purposes: use, publishing, sharing
# WHEN TO ASK FOR CONSENT

<table>
<thead>
<tr>
<th></th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-off</td>
<td>Simple</td>
<td>Research outputs (even questions, not known in advance)</td>
</tr>
<tr>
<td></td>
<td>Least hassle of participant</td>
<td>Participants will not know all content they will contribute</td>
</tr>
<tr>
<td>Process</td>
<td>Most complete for assuring active consent</td>
<td>Might not get consent needed before losing contact</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Repetitive, can annoy participant</td>
</tr>
</tbody>
</table>
ANONYMISATION PREVENTS IDENTITY DISCLOSURE

A person’s identity can be disclosed through:

- **direct identifiers**
  - *e.g. name, address, postcode, telephone number, voice, picture*

Often NOT essential research information (administrative)

- **indirect identifiers** – possible disclosure in combination with other information
  - *e.g. occupation, geography, unique or exceptional values (outliers) or characteristics*
KEY POINTS FOR ANONYMISING

• never disclose personal data - unless consent for disclosure
• reasonable/appropriate level of anonymity
• maintain maximum meaningful information
• where possible replace rather than remove
• identifying information may provide context, do not over-anonymise
• re-users of data have the same legal and ethical obligation to NOT disclose confidential information as primary users
ANONYMISING QUANTITATIVE DATA

• remove direct identifiers  
  *e.g. names, address, institution, photo*

• reduce the precision/detail of a variable through aggregation  
  *e.g. birth year vs. date of birth, occupational categories, area rather than village*

• generalise meaning of detailed text variable  
  *e.g. occupational expertise*

• restrict upper lower ranges of a variable to hide outliers  
  *e.g. income, age*

• combining variables  
  *e.g. creating non-disclosive rural/urban variable from place variables*
GEO-REFERENCED DATA

Spatial references (point coordinates, small areas) may disclose position of individuals, organisations, businesses

Remove spatial references - prevents disclosure; also all geographical and related information lost

Better

• reduce precision - replace point co-ordinates with larger, non-disclosing geographical areas
  e.g. km² area, postcode district, ward, road

• reduce precision - replace point coordinate with meaningful variable typifying the geographical position; or summary statistics of location
  e.g. catchment area, poverty index, population density

• keep spatial references and impose access restrictions on data
ANONYMISING QUALITATIVE DATA

• not collect disclosive data unless necessary
• plan or apply editing at time of transcription
  
  except: longitudinal studies - anonymise when data collection complete (linkages)

• avoid blanking out; use pseudonyms or replacements
• avoid over-anonymising - removing/aggregating information in text can distort data, make them unusable, unreliable or misleading
• consistency within research team and throughout project
• identify replacements, e.g. with [brackets]
• keep anonymisation log of all replacements, aggregations or removals made – keep separate from anonymised data files
• xml mark-up can be used for anonymisation
  
  <seg type="anonymised">word to be anonymised</seg>
### Example: Anonymisation log interview transcripts

<table>
<thead>
<tr>
<th>Interview / Page</th>
<th>Original</th>
<th>Changed to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Int1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p1</td>
<td>Spain</td>
<td>European</td>
</tr>
<tr>
<td>p1</td>
<td>E-print Ltd</td>
<td>Printing</td>
</tr>
<tr>
<td>p2</td>
<td>20(^{th}) June</td>
<td>June</td>
</tr>
<tr>
<td>p2</td>
<td>Amy</td>
<td>Moira</td>
</tr>
<tr>
<td>Int2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p1</td>
<td>Francis</td>
<td>my friend</td>
</tr>
</tbody>
</table>
Digital manipulation of audio and image files can remove personal identifiers

*e.g.* *voice alteration, image blurring (e.g. of faces)*

Labour intensive, expensive, may damage research potential of data

Better:

- obtain consent to use and share data unaltered for research purposes
- avoid mentioning disclosing information during audio recordings
WHAT IF ANONYMISING IS IMPOSSIBLE?

- obtain consent for sharing non-anonymised data
- regulating/restricting user access, e.g. at UK Data Archive:
  - archived data NOT in public domain
  - use of data for specific purposes only after user registration
  - data users sign legally binding End User Licence – e.g. not identify any potentially identifiable individuals
  - stricter access regulations for confidential data (case by case basis):
    - access to approved researchers only
    - requiring data access authorisation from data owner prior to data release
    - confidential data under embargo for given time period
    - secure access to data

- researchers - consider access to data and safe storage
ACCESS CONTROLS ON DATA

• Essential when anonymisation ineffective or damaging to quality
  • visual or audio data
  • disclosive microdata

• Multiple access controls can apply to different data types within one study
EXERCISE: CONSENT FOR DATA SHARING

How effective do you think the consent forms are? Is anything missing? What would you change?

p. 75-81
EXERCISE: ANONYMISING DATA

A. Can this interview transcript be shared with other researchers? What anonymisation would be needed before sharing it?

B. Compare the (partial) data table with the survey questionnaire. What needs anonymising before sharing the data?

p. 91-101
RIGHTS FOR RESEARCH DATA
RIGHTS & RESEARCH DATA

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
COPYRIGHT

• Research data have IPR:
  • literary or artistic work, e.g. texts, maps, recordings
  • database structure

• IPR when created, gives control over exploitation
COPYRIGHT

- Data owner (researcher) has copyright of research data
  - joint
  - employer – check!
- Copyright permissions sought and granted prior to data sharing / archiving
- Clearing copyright = reach agreement with copyright holder
- Data archives publish data – hold no copyright
  - copyright holders give permission to preserve and disseminate data
- Copyright clearance before data can be reproduced
- Except - fair dealing for data use for non-commercial research, private study, teaching, quotation, criticism, review
  - author and source must be cited
  - copyright clearance not needed
Consider the data copyright scenarios

p. 105-108
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

TOPICS:
• Why DMP
• How DMP
• Costing data management
WHY DATA MANAGEMENT PLANNING

- Research funders require planning for data management and data sharing, e.g. UK Research Councils
  - which data created during research
  - which policies apply (legal, institutional,....)
  - which data standards
  - how document data
  - ownership, copyright, IPR
  - data storage, backup, security
  - how share, preserve, incl. access
  - roles & responsibilities
  - costing resources needed

DCC: [UK research funders' DMSP expectations](#)
WHY DATA MANAGEMENT PLANNING

• Research benefits
  • think what to do with research data, how collect, how look after
  • keep track of research data (e.g. staff leaving)
  • identify support, resources, services needed
  • plan storage, short & long-term
  • plan security, ethical aspects
  • be prepared for data requests (FoI, funder)
HOW

- Funder template for DMP
  - [ESRC DMP requirements in data policy](#) and [DMP guidance](#)
  - [MRC DMP guidance and template](#)
  - [AHRC technical appendix requirements](#)

- DCC’s [DMPonline](#) tool
DATA LIFE CYCLE

Sign off consent form

Creating data
- design research
- plan data management (formats, storage etc)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

Processing data

Analysing data

Preserving data

Giving access to data

Re-using data

Licensing, terms and conditions for sharing, formal documentation

Shared data sharing protocols

Data formats, data migration

Agree data & metadata templates/organisation

UK DATA ARCHIVE
DATA MANAGEMENT CHECKLIST

- points to relevant to consider when planning appropriate data management for research
- select what is relevant for your research

- 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?

www.data-archive.ac.uk/create-manage/planning-for-sharing/data-management-checklist
COSTING

- Identify resources needed to make research data shareable beyond primary research team - above planned standard research procedures and practices
- Resources = people, equipment, infrastructure, tools to manage, document, organise, store and provide access to data
- Early planning can reduce costs

See our data management costing tool

<table>
<thead>
<tr>
<th>ADDITIONAL ACTIVITY</th>
<th>COMMENTS AND SUGGESTIONS</th>
</tr>
</thead>
</table>
| CONSENT FOR DATA SHARING             | • when consent for data sharing is considered as part of standard consent procedures - very low or no additional cost  
  Do you need to ask participants for their consent for data to be shared?  
  Essential for qualitative interviews; possibly less so in quantitative surveys where data can be more easily anonymised.  
  • when participants need to be re-contacted or re-visited after research to obtain retro-active consent for data sharing - could be high cost  
  • does this require preparation of information sheets, consent forms or extra time for consent discussions/staff training? |
| DIGITISATION                         | • is additional equipment or software needed for conversion?  
  Do analogue or paper-based research data need to be digitised to increase their potential for sharing?  
  • if simply image scanning text - may be low cost  
  • if Optical Character Recognition required then checking for accuracy needed - may be high cost |
ROLES & RESPONSIBILITIES

Assign, not presume roles or responsibilities for data management

Who?

- project director: design and oversee research
- research staff: design research; collect, process, analyse data; decide where data will be held, who will have access
- laboratory or technical staff: generate metadata / documentation
- database designer
- external contractors: data collection, data entry, transcribing, processing or analysis
- support staff: manage / administer research and research funding; ethical review; assess IPR
- institutional IT services: data storage, security, back-up services
- external data centres or archives: facilitate data sharing
<table>
<thead>
<tr>
<th>ACTIVITY</th>
<th>COMMENTS AND SUGGESTIONS</th>
<th>COST</th>
</tr>
</thead>
</table>
| Data description                                                        | - Are data in a spreadsheet or database clearly marked with variable and value labels, code descriptions, missing value descriptions, etc.?  
- Are labels consistent?  
- Do textual data like interview transcripts need description of context, e.g. included as a heading page? | if data description is carried out as part of data creation, data input or data transcription – low or no additional cost  
if needed to be added afterwards – higher cost  
codebooks for datasets can often be easily exported from software packages |               |
| Data cleaning                                                            | - Do quantitative data need to be cleaned, checked or verified before sharing, e.g. check validity of codes used, check for anomalous values?  
- Will data match documentation, e.g. same number of variables, cases, records, files?  
- Does textual information in data need to be spell-checked? | if carried out as part of data entry and preparation before data analysis – low or no additional cost  
if needed afterwards – higher cost |               |
| Documentation                                                            | - Do you have documentation for the data that describes the context and methodology of how data were gathered, created, processed and quality controlled? | often essential contextual and methods documentation will be written up in publications and reports  
if all data creation steps are well documented and documentation is kept well organised during research – low or no additional cost  
if documentation to be written or compiled specifically afterwards – higher cost |               |
| Metadata                                                                 | - Do structured metadata need to be created when data are shared via a data centre or archive, e.g. completing a deposit form for the UK Data Archive? | completing a UK Data Archive deposit form may take one to two hours  
other data centres will have their own metadata forms |               |
| Formatting and organising                                               | - Are your data files, spreadsheets, interview transcripts, records etc. all in a uniform format or style?  
- Are files, records and items in the collection clearly named with unique file names and well organised? | if planned beforehand by developing templates and data entry forms for individual data files (transcripts, spreadsheets, databases) and by constructing clear file structures – low or no additional cost  
if needed afterwards – higher cost  
free software exists for batch file renaming to harmonise file names |               |
| Transcription                                                            | - Will you transcribe qualitative data (e.g. | if part of research practice – very low or no additional cost |               |
EXERCISE: DATA MANAGEMENT PLANNING

Use the data management checklist to help point to relevant data management topics you need to consider when planning your research project.
• Data management for qualitative research data using Nvivo9 software: [http://www.data-archive.ac.uk/media/262066/ukda-datamanagement-nvivo.pdf](http://www.data-archive.ac.uk/media/262066/ukda-datamanagement-nvivo.pdf)