
Managing and Sharing Research Data

Louise Corti
Research Data Management Team

Wise Up Wednesdays
26 November 2014

UK Data Service



Overview of this session

Presentation

- About the UK Data Service
- Managing your data – some key points
- Issues in sharing your data
- Training materials available

Exercises

- ‘Fun’ quiz on documenting, formatting, organising, and storing your data
- Anonymising qualitative text



What is the UK Data Service?

- a comprehensive resource funded by the Economic and Social Research Council (ESRC)
- a single point of access to a wide range of secondary social science data
- support, training and guidance throughout the data life cycle
- listen to our recorded webinars at

<http://ukdataservice.ac.uk/news-and-events/videos.aspx>



UK Data Service

Site Search FAQ Help Contact

UK Data Service

About us Get data Use data Manage data Deposit data News and Events

Welcome to the UK Data Service

Your resource for quality social research data

A unified point of access to data from ESDS, Census Programme, Secure Data Service and others

STOP

LOGIN / REGISTER

DISCOVER UK DATA SERVICE

Search and browse our data collection, support guides, case studies and related publications.

GO

Data Website

LATEST TWEETS

UKDataService RCUK's revised Policy on Open Access now online, to be in effect this month. <http://t.co/k3ouofVe4O> @research_uk #openaccess

UKDataService @Barnard17 There are lots of population statistics available via @ONS and @EU_Eurostat. Sent a list of links to your blog. Good luck!

UKDataService RT @Censusacuk: <http://t.co/qVRh4S0y2m> is now Census Support

LATEST NEWS

Call for papers: Opinions and Lifestyle Survey user meeting

European Social Survey invites EU data users to visit and learn

Digital Futures: Your input needed on data digitisation

UK Data Service featured in Royal

OUR DATA COMMUNITY

The UK Data Service is at the core of a network of trust that includes data owners, producers, funders and users.

Who can most benefit from the data we hold?

- Academic researchers and students
- Government analysts
- Charities and foundations

QUICK ACCESS TO

Key data

Census Support

Information for new users

Frequently asked questions

Integrates ESDS, Survey Question Bank and Census.ac.uk

ukdataservice.ac.uk

UK Data Service



What does the UK Data Service do?

- put together a **collection of the most valuable data** and enhance these over time
- **preserve data in the long term** for future research purposes
- make the **data and documentation** available for reuse
- **provide data management advice** for data creators
- **provide support** for users of the service
- information about **how data are used**
- easy access through **website**



Who is it for?

- academic researchers and students
- government analysts
- charities and foundations
- business consultants
- independent research centres
- think tanks
- citizen scientists, where skills enable analysis



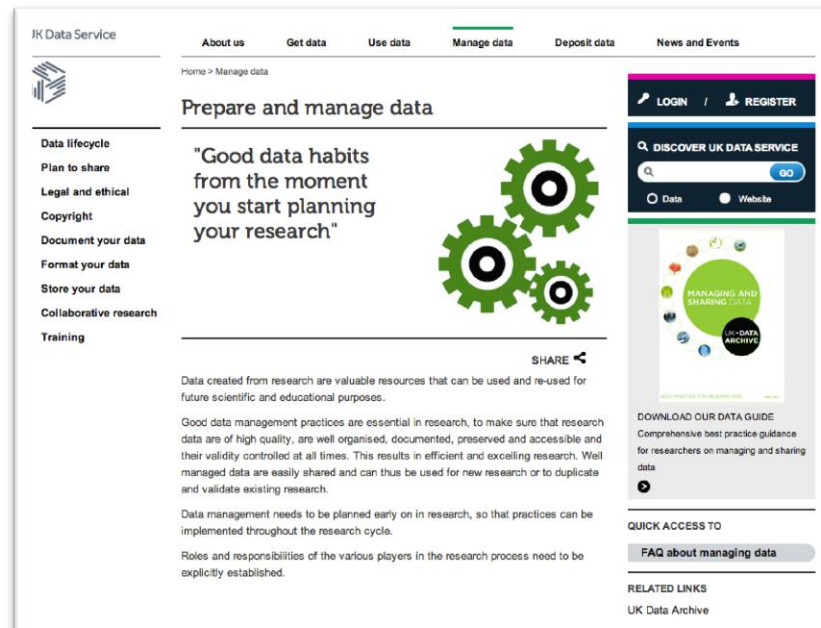
Our approach to data sharing

- Our organisation has **45 years** experience in selecting, ingesting, curating and providing access to social science data
- Extensive experience of **supporting researchers and data creators** of social science data and related disciplines
- Been managing data sharing for the **ESRC Data Policy** since 1995, including cross council programmes
- Our best practice approaches to making data shareable based on:
 - challenges faced by researchers to share data
 - handling research data – quantitative and qualitative



Our managing and sharing data resources

- Online best practice guidance: ukdataservice.ac.uk/manage-data.aspx
- Sage Managing and Sharing Research Data – a Guide to Good Practice: www.uk.sagepub.com/books/9781446267264
- Busy training programme



UK Data Service



Key skills that enhance methods training

- Policy landscape and data sharing
- Writing and implementing a data management plan
- Documenting and contextualising data
- Formatting and organising data
- Storing and transferring data, incl. encryption and security
- Legal, ethical issues in handling and sharing data – consent, anonymisation and access control
- Rights relating to research data
- Publishing and citing research data



Why data management planning

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

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 (Fol, funder)



ESRC research data policy

Research data should be openly available to the maximum extent possible through long-term preservation and high quality data management.
(ESRC Research Data Policy, 2010)

- ESRC grant applicants planning to create data during their research include a [data management plan](#) with their application, as an attachment to the Je-S form
- ESRC award holders offer their research data to the [ESRC Data Store](#) (managed by UK Data Service) within three months of the end of their grant, to preserve them and to make them available for new research.

Researchers who collect the data initially should be aware that ESRC expects that others will also use it, 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, 2012)

ESRC data management plan

Assessment of existing data

Information on new data

Quality assurance of data

Backup and security of data

Expected difficulties in data sharing

Copyright / Intellectual Property Right

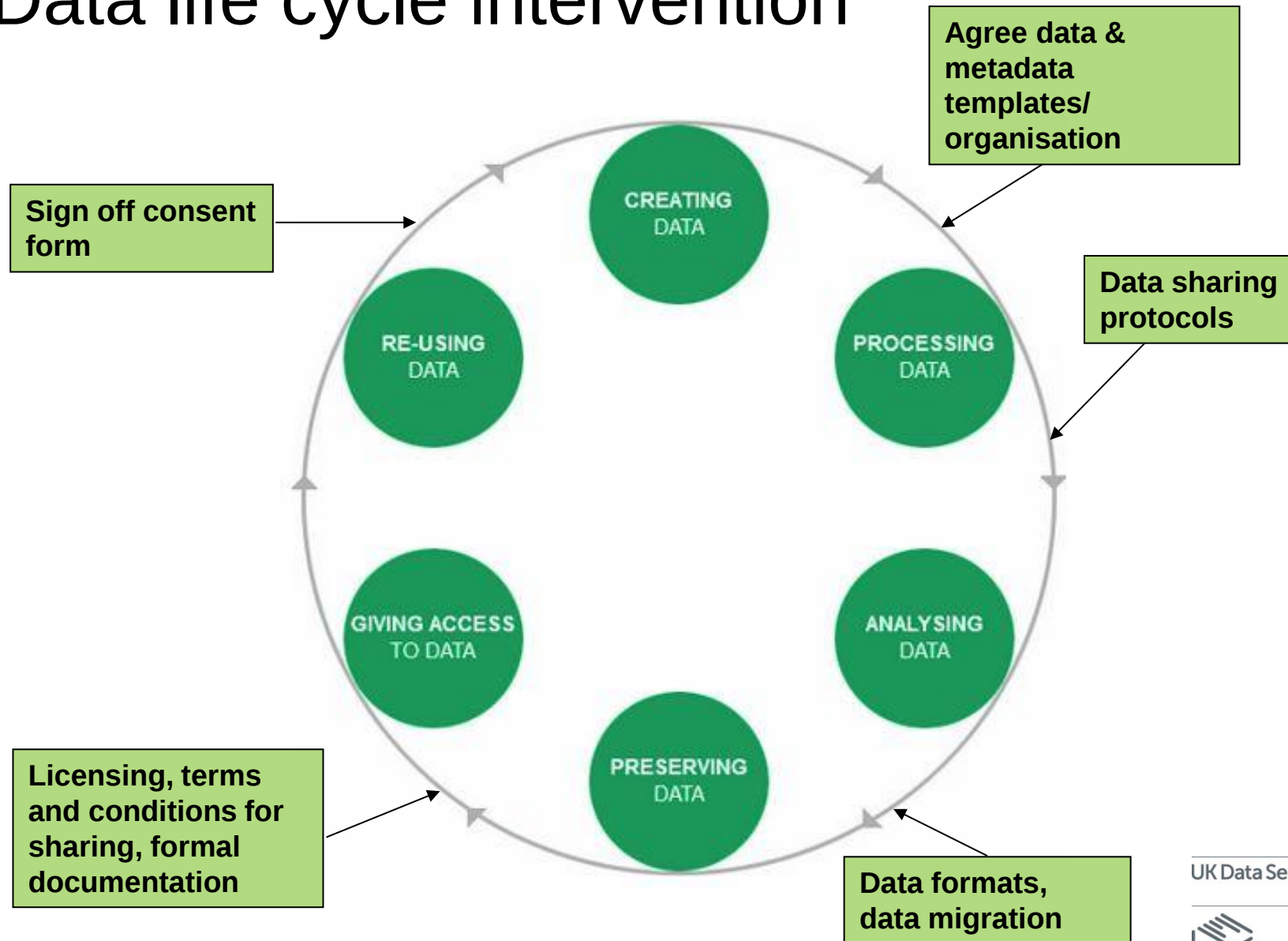
Responsibilities

Preparation of data for sharing and archiving

[ESRC DMP guidance](#)



Data life cycle intervention



Key planning issues

- Know your legal, ethical and other obligations towards research participants, colleagues, research funders and institutions
- Know your institution's policies and services: storage and backup strategy, research integrity framework, IPR policy, institutional data repository
- Assign roles and responsibilities to relevant parties
- Incorporate data management into research cycle
- Implement and review management of data during project meetings and review



How to cost data management

STEP 1

- check data management activities in table and *tick* what applies to your proposed research; we propose 18 essential RDM activities

STEP 2

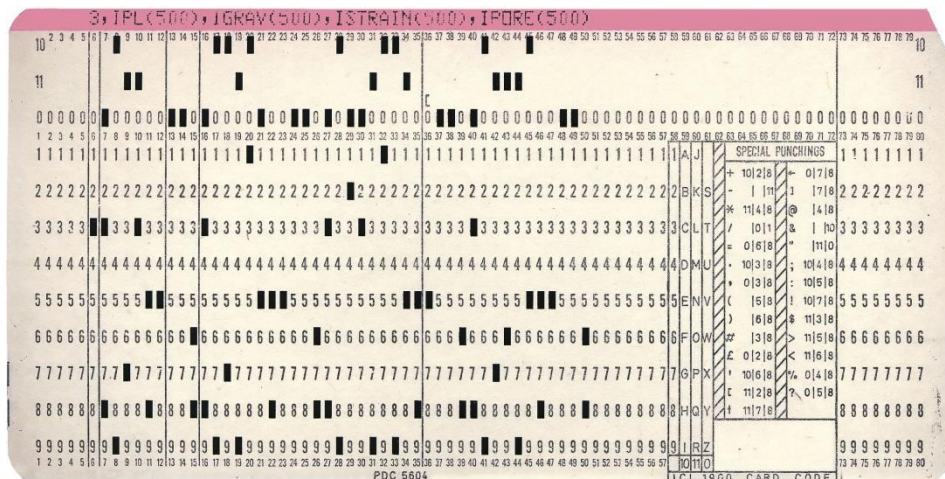
- for each selected activity, estimate / calculate additional time and/or resources needed and cost this

STEP 3

- add data management costs to your research application; coordinate resourcing and costing with your institution, research office and institutional IT services



Can you understand/use these data?



SrvMthdDraft.doc

SrvMthdFinal.doc

SrvMthdLastOne.doc

SrvMthdRealVersion.doc



UK Data Service



File formats

- Choice of software format for digital data:
 - hardware used e.g. audio capture
 - discipline-specific customs and planned data analyses
 - software availability/cost
- Digital data endangered software/ hardware obsolesce
- Best formats for long-term preservation are standard, interchangeable and open formats
 - tab-delimited, comma-delimited (CSV), ASCII
 - SPSS portable, XML
 - RTF, OpenDocument format, PDF/A,
 - See [Recommended formats](#)
- Beware of errors/losses of data when converting!



Format conversion

	A	B	C	D	E	F
1		Timber volumes in m3				
2	Year	1994	1995	1996	1997	1998
3	Date recorded	20/01/1995	23/01/1996	11/01/1997	16/01/1998	14/12/1998 ¹
4	Logging private land	20346.345	47005.223	26001.754	11468.897	0.000
5	Logging forest reserves	4060.567	1777.783	804.997	0.000	3329.653
6	Logging state land	0.000	1200.000	559.162	2077.567	358.935
7	Total	61119.912	87065.006	64802.913	51354.464	5686.588
8						
9		Data missing				
10		Estimate				
11						
12	¹ temporary volumes					
13						

MS Excel (.XLSX) format using colour highlighting for annotation

Loss of
annotation

	A	B	C	D	E	F
1		Timber volumes in m3				
2	Year	1994	1995	1996	1997	1998
3	Date recorded	20/01/1995	23/01/1996	11/01/1997	16/01/1998	14/12/1998 ¹
4	Logging private land	20346.345	47005.223	26001.754	11468.897	0
5	Logging forest reserves	4060.567	1777.783	804.997	0	3329.653
6	Logging state land	0	1200	559.162	2077.567	358.935
7	Total	61119.912	87065.006	64802.913	51354.464	5686.588
8						
9		Data missing				
10		Estimate				
11						
12	¹ temporary volumes					
13						

Tab-delimited text format, and loss of colour annotation



Formatting for high quality data:

- use consistent templates for the same kind of data
- are well organised – folders
- are suitably named
- are properly versioned
- have the authenticity of master files identified
 - tools available for versioning and synchronising files



How do you document your data?

- Data doesn't mean anything without documentation
 - A table becomes just a block of meaningless numbers
 - An interview becomes a block of meaningless text
- Data documentation might include:
 - A survey questionnaire
 - An interview schedule
 - Records of interviewees and their demographic characteristics in a qualitative study
 - Variable labels in a table
 - Published articles that provides background information
 - Description of the methodology used to collect the data



Why document your data?

- Enables you to understand data when you return to it!
- Makes data independently understandable i.e. reusable
- Helps avoid incorrect use/misinterpretation
- If using your data for the first time, what would a new user need to know to make sense of it?
- The UK Data Archive uses data documentation to:
 - supplement a data collection with documents such as a user guide(s) and data listing
 - ensure accurate processing and archiving
 - create a catalogue record for a published data collection



Consider documentation early on

- Entirely dependent on the creator (you) can provide!
- Start gathering meaningful information from as early on in the research process as possible
- Consider during data management planning
- Identify likely documentation to generate and retain, such as:
 - **Useful documents** such as methodological descriptions and end of award reports
 - Information on **dataset structure** e.g. relationships between files
 - Metadata inside files i.e. **variable-level documentation**
 - **Contextual** information about project and data
 - Data collection **methodology and processes**
 - Data confidentiality, access and use conditions



Quantitative study

- Smaller-scale study – single user guide may contain compiled survey questionnaire, methodology information
- Example from Understanding Society, a bigger study - many documents presented separately:

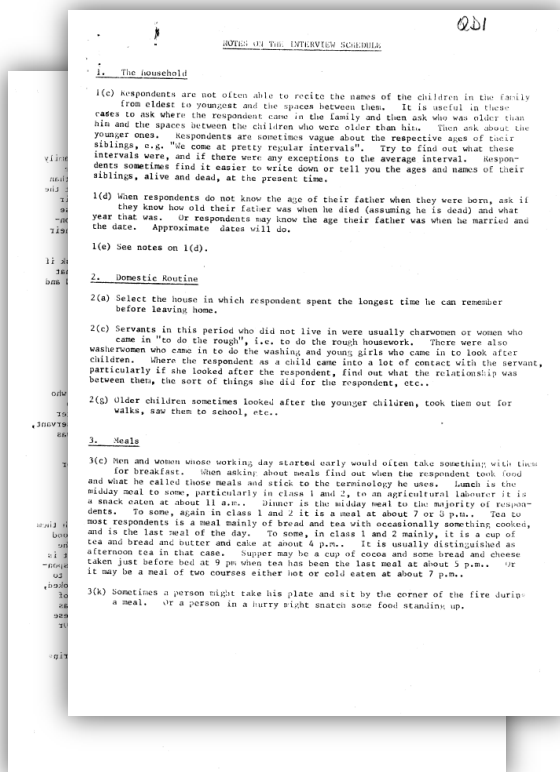
DOCUMENTATION

Title	File Name	Size (KB)
Cognitive Ability Measures	6614_cognitive_ability_measures_v1-1.pdf	348
Revisions November 2013	6614_ukhls_2013_revisions.pdf	375
Wave 1 Adult Main Questionnaire	6614_understanding_society_wave1_questionnaire.v04.pdf	2802
Wave 2 Adult Main Questionnaire	6614_understanding_society_wave2_questionnaire_v04.pdf	3726
Waves 1-3 User Manual	6614_usermanual_wave1to3_v1-1.pdf	883
Wave 3 Youth Self-Completion Questionnaire (GB)	6614_w3_youthquestionnaire_gbritain_annotated.pdf	1469
Wave 1 Consent Package	6614_wave1_consent_package.pdf	645
Wave 1 Adult Self-Completion Questionnaire	6614_wave1_main_adult_sc_questionnaire.pdf	429
Wave 1 Youth Self-Completion Questionnaire	6614_wave1_main_youth_sc_questionnaire.pdf	750
Wave 1 Project Instructions for Interviewers	6614_wave1_project_instructions_interviewers.pdf	2426
Wave 1 Showcards	6614_wave1_showcards.pdf	199



Qualitative study – user guide

- A user guide could contain a variety of documents that provide context: interview schedule, transcription notes, even photos



Qualitative study – data list

- Data listing provides an at-a-glance summary of interviews

Study Number 5407

Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001

Mort, M.

The panel respondents for the study were divided into six population groups. The data list for the diary and interviews has been colour-coded accordingly for clarity, using the depositor's original colours:

Group 1: Farmers	Group 2: Rural Business	Group 3: Agricultural related occupations	Group 4: Frontline Workers	Group 5: Community	Group 6: Animal / Human Health Professionals
------------------	-------------------------	---	----------------------------	--------------------	--

1. Interviews

Respondent ID	Population Group	Date of Birth	Gender	Occupation	Interview summary	Place of Interview
PM02	Group 6: Animal / Human Health Professionals	1975	M	Veterinary Surgeon	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria, respondent's home
PM03	Group 6: Animal / Human Health Professionals	1966	F	Veterinary Surgeon	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria
PM07	Group 6: Animal / Human Health Professionals	1964	F	Veterinary practice manager	Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	North Cumbria, respondent's home
					Family and background, career and work, arrangements during FMD epidemic and perceptions of situation	

UK Data Service



Data-level documentation

- Certain types of data file may contain important information:
 - variable/value labels; document metadata; table relationships and queries in relational databases; GIS data layers/tables
- Some examples:
 - SPSS: variable attributes documented in Variable View (label, code, data type, missing values)
 - MS Access: relationships between tables
 - ArcGIS: shapefiles (layers) and tables in geodatabase; metadata created in ArcCatalog
 - MS Excel: document properties, worksheet labels (where multiple)

Data-level documentation: variable names

- All structured, tabular data should have cases/records and variables adequately documented with names, labels and descriptions
- Variable names might include:
 - 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 should be max 8 characters and without spaces



Data-level documentation: variable labels

- Similar principles for variable labels:
 - be brief, max. 80 characters
 - include unit of measurement where applicable
 - reference the question number of a survey or questionnaire
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 ; ISO 3166 alpha-2 country codes



Data-level documentation: transcripts

- Qualitative data/text documents:
 - interview transcript speech demarcation (speaker tags)
 - document header with brief details of interview date, place, interviewer name, interviewee details, context



Audio transcription

- adopt a uniform layout throughout the research project
- compatibility with import features of Computer Assisted Qualitative Data Analysis Software (CAQDAS)
- role of transcription varies by discipline. What to transcribe?
 - verbal and non-verbal?
 - turn-taking?
 - 'interruptions'
- who does it – researcher, service? Need rules
- implications of technologies – video, multiple camera, screen capture, webcams



Metadata – data about data

- Similar to documentation in that it provides context and description, but is much more **structured**
- Standard data collection metadata includes:
 - Components of a bibliographic reference
 - Core information that a search engine indexes to make the data findable
- International standards/schemes
 - Data Documentation Initiative (DDI)
 - ISO19115 (geographic)
 - Dublin Core
 - Metadata Encoding and Transmission Standard (METS)
 - Preservation Metadata Maintenance Activity (PREMIS)

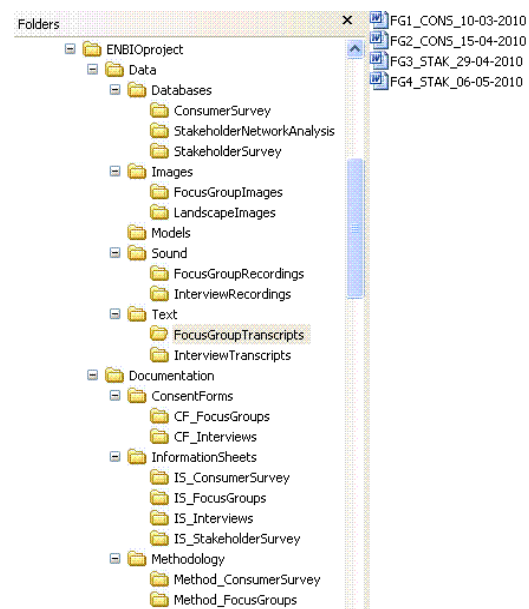
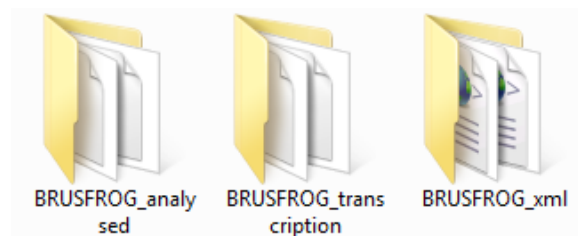
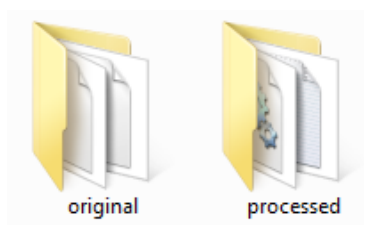


Organising data

- Plan in advance how best to organise data
- Use a logical structure and ensure collaborators understand

Examples

- hierarchical structure of files, grouped in folders, e.g. audio, transcripts and annotated transcripts
- survey data: spreadsheet, SPSS, relational database
- interview transcripts: individual well-named files



Storing data safely

- Looking after research data - protecting them from damage and loss
- Strategies in place for:
 - backing-up
 - transmission
 - secure storage
 - disposal



8 December 2010 Last updated at 11:43



Nasa sells shuttle PCs without wiping secret data

US space agency Nasa has been left red-faced after selling off computers without ensuring that highly sensitive data had been removed.

An internal investigation found 10 cases where PCs were sold despite failing data removal procedures.

Another four PCs - which were about to be sold - were found to contain data restricted under arms control rules.



Nasa is selling of hundreds of PCs used Space Shuttle programme



Digital back-up strategy

Consider

- what's backed-up? - all, some, just the bits you change?
- where? - original copy, external local and remote copies
- what media? - CD, DVD, external hard drive, tape, etc.
- how often? – assess frequency and automate the process
- for how long is it kept? Data retention policies that might apply?
- verify and recover - never assume, regularly test a restore

Backing-up need not be expensive

- 1Tb external drives are around £50, with back-up software

Consider non-digital storage too!



"We back up our data on sticky notes because sticky notes never crash."

Encryption and security

Always encrypt personal or sensitive data:

- when moving data files
- when or storing files

Free softwares that are easy to use:

- encrypt hard drives, partitions, files and folders
- encrypt portable storage devices such as USB flash drives
- Safehouse, Truecrypt, Axcrypt

Protect data from unauthorised access, change, disclosure, destruction

- control access to all computers devices
- control physical access to buildings, rooms, cabinets
- restrict access to sensitive materials e.g. consent forms

Proper disposal of equipment and media

- even reformatting the hard drive is **not** sufficient



File sharing and collaborative environments

Sharing data between researchers and teams

- too often email attachments
- Virtual Research Environments
 - MS SharePoint
- Cloud solutions
 - Google Drive, DropBox, Microsoft SkyDrive etc.
 - Base camp
 - Locally managed; ownCloud and ZendTo
- file transfer protocol (ftp)
- physical media



By David Fletcher
<http://www.cloudtweaks.com/2011/05/the-lighter-side-of-the-cloud-data-transfer/>

Quiz

Some light-hearted questions covering:

- Documenting p.51
- Formatting p. 74
- Back up and storage p. 100

Corti et al. (2014)

Ethical arguments **for** sharing and 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)
- 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
- An exception to the duty of confidentiality occurs when an informant has consented to information being used in specified ways
- Public interest can override duty of confidentiality, so best to avoid vague and general promises in consent forms
 - police investigations
 - child welfare



Data Protection Act, 1998

- Personal data:
 - relate to a living individual
 - individual can be identified from those data or from those data and other information
 - include any expression of opinion about the individual
- Only disclose personal data if consent given to do so (and if legally required to do so)
- DPA does not apply to anonymised data

- processed fairly and lawfully
- obtained and processed for specified purpose
- adequate, relevant and not excessive for 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



Data Protection Act and research

- 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

The Data Protection Act 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



Sharing confidential data

Researchers to consider:

- obtaining **informed consent** for data sharing and preservation /curation
- **protecting identities** not collecting personal data or anonymisation
- **restricting / regulating access** where needed (all or part of data). UK Data Service uses a spectrum of access

Consider jointly and in dialogue with participants

Plan early in research

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

Anonymising quantitative data - tips

- 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



Anonymising qualitative data

- 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 or make it 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

Anonymising qualitative data

Example: Anonymisation log interview transcripts		
Interview / Page	Original	Changed to
Int1		
p1	Spain	European country
p1	E-print Ltd	Printing company
p2	20 th June	June
p2	Amy	Moira
Int2		
p1	Francis	my friend

Case study of anonymisation

SN 5407 Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (SN5407 at UK Data Archive)

Maggie Mort, Lancaster University

[Discover.ukdataservice.ac.uk/catalogue](https://discover.ukdataservice.ac.uk/catalogue)

- funded by Department of Health
- recruit panel of 54 local people in affected area at time of FM crisis: farmers, agricultural professionals, small businesses, health professionals, vets, residents
- weekly diaries for 18 months describing how their life was affected by the crisis and process of recovery observed around them (handwritten)
- in-depth interviews and group discussions (audio recordings, transcripts)
- at end of research – feeling by researchers that data should be archived



Data sharing outcome

- sought advice from copyright specialist re. terms of agreement for archiving
- met with UK Data Service for advice data archiving
- developed separate consent forms for written and audio material, with opt in / opt out and an embargo option
- piloted discussion on data archiving with 4 panel members to explore:
 - feelings re. data anonymisation, confidentiality, copyright, ownership
 - understanding of archiving by participants and information required
 - user options of archived data - scholarly / educational purposes
- discussed archiving individually with each panel member
 - 7 panel members declined archiving their data
 - 40 interview and diary transcripts were made available for re-use by registered users
 - 3 interviews and 5 diaries were embargoed until 2015
 - audio files archived and only available by permission from researchers

Detailed information:

UK Data Service



Exercise

- Anonymising qualitative data
p. 133/4
- Consent forms
P. 128
- Group discussion on challenges of sharing your own data



Keep connected

- Subscribe to UK Data Service list:
www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE
- Follow UK Data Service on Twitter: @UKDataService
- Facebook
- Youtube: www.youtube.com/user/UKDATASERVICE



Contact

Collections Development and Producer Relations team
UK Data Service
University of Essex
ukdataservice.ac.uk/help/get-in-touch.aspx

UK Data Service

