Data Management Basics

The webinar will begin at 3pm

- You now have a menu in the top right corner of your screen.
- The red button with a white arrow allows you to expand and contract the webinar menu, in which you can write questions/comments.
- We will answer your questions at the end.
- If we do not get to a question, we will reply later by e-mail.
- You will be on mute throughout the webinar – we need to do this in order to ensure a high quality recording.
Data Management Basics

Anca Vlad
UK Data Service
Research Data Services Team

05 December, 2019
Overview

• UK Data Service

• Managing your data – background, why and how
  • GDPR
  • Consent
  • Anonymisation
  • Access controls
  • Documentation
  • Security
  • Encryption
  • Backups

• Your questions
Data Management at the UK Data Service

- Support and training for data creators with accessing, managing, and using data
- One-stop-shop for social science data

Website, data catalogue: https://www.ukdataservice.ac.uk/
More webinars available: https://www.ukdataservice.ac.uk/news-and-events/events.aspx
Background

• Data sharing is fast becoming a new paradigm in research across all disciplines, providing benefits to individual researchers, institutions, funders and more

• Good research data management habits are essential to creating data that are suitable for sharing and reuse

• Many funders and publishers now specify requirements for data handling, including the formulation of a data management plan
Why is data management important?

- Data creation in research is often expensive
- Data is the cornerstone of research
- Good quality data leads to good quality research
- Data underpins published findings
- Enables compliance with ethical codes, data protection laws, journal requirements and funder policies
- To protect data from loss, destruction and potential exposure
Practical steps researchers can take

• Write a data management plan

• Make sure data are shareable and can be understood:
  • Obtain consent to share
  • Do not disclose identities without consent
  • Use open source and standard formats
  • Provide context and documentation
  • Protect your data at all stages (secure storage, encryption)
Data management plan

- Assessment of existing data
- Information on new data
- Quality assurance of data
- Backup and security of data
- Difficulties in data sharing and measures to overcome these
- Consent, anonymization, re-use strategies
- Copyright/Intellectual Property Ownership
- Responsibilities
- Management and curation

Data management plan guidance
Multiple tools for protecting participants

1. Seek **informed consent**, also for data sharing and long-term preservation and curation
2. Protect identities e.g. **anonymization**, and (or) not collecting personal data (only collect data that is necessary)
3. Regulate **access** where needed (all or part of data) e.g. by group, use or time period
The GDPR - basics

- Applies to personal data, pseudonymised data and living persons only
- Personal data are ‘any information relating to an identified or identifiable natural person’
- Note that not all research data personal data
- Also note there may still be ethical reasons for wanting to protect this information though!
The GDPR – processing grounds

There are **six** grounds for the processing of personal data, and one of these must be present in order to process a data subject’s personal data:

1. Consent
2. Contract
3. Legal obligation
4. Vital interests
5. Public interest (public task)
6. Legitimate interest
Consent for sharing – one more small step

- Engagement in the research process
  - What activities are involved in participating in the project?
- Dissemination in presentations, publications, the web
  - Consent for use of quotes for articles and video publicity
- Data sharing and archiving
  - Consider future uses of data

Consent is always dependent on the research context – special cases of covert research and verbal consent.
In practice: Wording

Wording in consent forms and information sheets could be broken down into **three** key areas:

1. Taking part in the study
2. Use of the information in the study
3. Future use and reuse of the information by other

Model consent form (GDPR compliant):
https://www.ukdataservice.ac.uk/media/622375/ukdamodelconsent.doc
In practice: Wording

3. Future use and reuse of the information by others

I give permission for the [specify the data] that I provide to be deposited in [name of data repository] so it can be used for future research and learning.

Specify in which form the data will be deposited, e.g. anonymised transcripts, audio recording, survey database, etc.; and if needed repeat the statement for each form of data you plan to deposit.

Specify whether deposited data will be anonymised, and how. Make sure to describe this in detail in the information sheet.

Specify whether use or access restrictions will apply to the data in future, e.g. exclude commercial use, apply safeguarded access, etc.; and discuss these restrictions with the repository in advance.

We expect to use your contributed information in various outputs, including a report and content for a website. Extracts of interviews and some photographs may be used as well. We will get your permission before using a quote from you or a photograph of you.

After the project has ended, we intend to archive the interviews at [name of repository]. Then the interview data can be disseminated for reuse by other researchers, for research and learning purposes.
## Anonymization

<table>
<thead>
<tr>
<th>Quantitative</th>
<th>Qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Remove direct identifiers [e.g. names, address, institution, photo]</td>
<td>• Remove direct identifiers (e.g names, address, institution, photos) or replace with pseudonyms</td>
</tr>
<tr>
<td>• Reduce the precision/detail of a variable through aggregation [e.g. birth year instead of date of birth, occupational categories rather than jobs; and, area rather than village]</td>
<td>• Avoid blanking out; use pseudonyms or replacements</td>
</tr>
<tr>
<td>• Generalise meaning of detailed text variable [e.g. occupational expertise]</td>
<td>• Identify replacements with [brackets]</td>
</tr>
<tr>
<td>• Restrict upper and lower ranges of a variable to hide outliers [e.g. income, age]</td>
<td>• Plan or apply editing at time of transcription</td>
</tr>
<tr>
<td>✅ Keep an anonymization log (and keep it separate from anonymised data files)</td>
<td>• Consistency throughout project</td>
</tr>
<tr>
<td></td>
<td>• Avoid over-anonymising – removing information in text can distort data, make them unusable, unreliable or misleading; so balance anonymization with the need to preserve context</td>
</tr>
<tr>
<td></td>
<td><a href="https://www.ukdataservice.ac.uk/deposit-data/stories/gush">https://www.ukdataservice.ac.uk/deposit-data/stories/gush</a></td>
</tr>
<tr>
<td></td>
<td>✅ Keep an anonymization log (and keep it separate from anonymised data files)</td>
</tr>
</tbody>
</table>
Anonymization - Audio-visual data

- Data manipulation of audio and image files can remove personal identifiers
  - e.g. voice alteration and image blurring (e.g. of faces)
- Labour intensive, expensive, may damage research potential of data
- Better alternatives:
  - Obtain consent to use and share data unaltered for research purposes (wish access restrictions in place)
  - Avoid mentioning disclosing information during audio recordings
In practice: example anonymisation

Health and Social Consequences of the Foot and Mouth Disease
Epidemic in North Cumbria, 2001-2003 (SN 5407)
M. Mort, Lancaster University. Institute for Health Research

NOTE: all identifying info contained in this transcript is fictional

Date of Interview: 21/02/02

Interview with Lucas Roberts, DEFRA field officer
Date of birth: 2 May, 1965

Gender: Male

Occupation: Frontline worker
Location: Plumpton, North Cumbria

Lucas was living at home with his parents, “but I'm hoping to move out soon” so we met at his parents’ small neat house. We sat in a very comfortable sitting room with an open fire and Lucas made me coffee and offered shortbread. Although at first Lucas seemed a little nervous, quick to speech and very watchful he seemed to relax as we spoke and to forget about the tape.

I will just start by asking you to tell me a little bit about yourself and your background.
### Managing access to data

<table>
<thead>
<tr>
<th>Category</th>
<th>Access Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open</strong></td>
<td>• available for download/online access under open licence without any registration</td>
</tr>
</tbody>
</table>
| **Safeguarded** | • available for download / online access to logged-in users who have registered and agreed to an End User Licence *(e.g. not identify any potentially identifiable individuals)*  
                        • special agreements (depositor permission; approved researcher)  
                        • embargo for fixed time period                                                                                             |
| **Controlled** | • available for remote or safe room access to authorised and authenticated users whose research proposal has been and who have received training |
In practice: data with access conditions


- 40 interview and diary transcripts are archived and available for reuse by registered users (**safeguarded access**)
- 3 interviews and 5 diaries were embargoed until 2015 (**Safeguarded – embargoed**)
- Audio files archived and only available with permission from depositor(s) (**Safeguarded – Special agreement**)

Documenting your data

• Enables you to understand the data if/when you return to it.
• Sufficient information for future researchers to understand and use the data

• 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 and research instruments
  • Ensure accurate processing and archiving
  • Create a catalogue record for a published data collection
Include as documentation

- Data collection methodology and processes: sampling, sample size, fieldwork protocol, experiment protocol, interviewer instructions
- Codebook, user guide (for quantitative data)
- Information sheet, consent form (blank versions)
- Questionnaires, show cards, topic guides
- Transcripts: header with context information: data and place of interview, interviewer, interviewee details (in line with consent form) etc.
- Data list: overview of key information about each interview, a map of the data collection (for qualitative data)
- Links to reports and publications (preferably DOIs where possible)
Data-level documentation

- All structured, tabular data should have adequate variable names, variable and value labels
- Variable names might include:
  - Question number system matching questions in the questionnaire used e.g. Q1a, Q1b, Q2, Q3b
  - 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’
  - For interoperability across platforms, variable names should not be longer than 8 characters and without spaces
Data-level documentation

Similar principles for variable labels:
• Be brief, maximum 80 characters
• Include unit of measurement where appropriate
• Reference the question number of a survey or questionnaire
e.g. variable ‘q11hexw’ with label ‘Q11b: hours spent taking physical exercise in a typical week’ – the label gives the unit of measurement and a reference to the questions number (Q11b)
• Coding or classification schemes used, with a bibliographic reference
e.g. Standard Occupational Classification 2000; ISO 3166 alpha-2 country codes

For value labels:
• 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(skipped)’, ’96= not known’, ’95=error’
In practice: user guide and documentation

A user guide should contain variety of documents that provide context: interview schedule, methodology, study findings, consent procedures, transcription notes, codebook etc.

User guide for
### In practice: data list

**Study Number 6377**

*Integrated Floodplain Management, 2006-2008*

Morris, J.

#### Floodplain farm survey

<table>
<thead>
<tr>
<th>Interview ID</th>
<th>Farmer code</th>
<th>Age</th>
<th>Farm scheme</th>
<th>Farm type</th>
<th>Size of farm (hectare)</th>
<th>Number of holdings</th>
<th>Date of interview</th>
<th>Interviewer name</th>
<th>No of pages</th>
<th>Text file name</th>
<th>Audio file name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Be1</td>
<td>35-45</td>
<td>Beckingham</td>
<td>Beef</td>
<td>360</td>
<td>1</td>
<td>04.12.2006</td>
<td>Helena</td>
<td>28</td>
<td>6377int001</td>
<td>6377int001</td>
</tr>
<tr>
<td>2</td>
<td>Be2</td>
<td>45-55</td>
<td>Beckingham</td>
<td>Arable</td>
<td>364</td>
<td>1</td>
<td>05.12.2006</td>
<td>Helena</td>
<td>21</td>
<td>6377int002</td>
<td>6377int002</td>
</tr>
<tr>
<td>3</td>
<td>Be3</td>
<td>45-55</td>
<td>Beckingham</td>
<td>Arable</td>
<td>372</td>
<td>2</td>
<td>06.12.2006</td>
<td>Helena</td>
<td>22</td>
<td>6377int003</td>
<td>6377int003</td>
</tr>
<tr>
<td>4</td>
<td>Be4</td>
<td>45-55</td>
<td>Beckingham</td>
<td>Arable</td>
<td>194</td>
<td>3</td>
<td>06.12.2006</td>
<td>Helena</td>
<td>18</td>
<td>6377int004</td>
<td>6377int004</td>
</tr>
<tr>
<td>5</td>
<td>Be5</td>
<td>55-65</td>
<td>Beckingham</td>
<td>Arable</td>
<td>108</td>
<td>1</td>
<td>07.12.2007</td>
<td>Helena</td>
<td>21</td>
<td>6377int005</td>
<td>6377int005</td>
</tr>
<tr>
<td>6</td>
<td>Be6</td>
<td>45-55</td>
<td>Beckingham</td>
<td>Arable</td>
<td>1254</td>
<td>2</td>
<td>01.02.2008</td>
<td>Helena</td>
<td>19</td>
<td>6377int006</td>
<td>6377int006</td>
</tr>
<tr>
<td>7</td>
<td>Bu1</td>
<td>55-65</td>
<td>Bushley</td>
<td>Mixed</td>
<td>101</td>
<td>2</td>
<td>13.02.2007</td>
<td>Quentin</td>
<td>29</td>
<td>6377int007</td>
<td>6377int007</td>
</tr>
<tr>
<td>8</td>
<td>Bu2</td>
<td>&gt;65</td>
<td>Bushley</td>
<td>Mixed</td>
<td>97</td>
<td>1</td>
<td>15.02.2007</td>
<td>Quentin</td>
<td>15</td>
<td>6377int008</td>
<td>6377int008</td>
</tr>
<tr>
<td>9</td>
<td>Bu3</td>
<td>&gt;65</td>
<td>Bushley</td>
<td>Arable</td>
<td>194</td>
<td>4</td>
<td>13.02.2007</td>
<td>Quentin</td>
<td>21</td>
<td>6377int009</td>
<td>6377int009</td>
</tr>
<tr>
<td>10</td>
<td>Bu4</td>
<td>55-65</td>
<td>Bushley</td>
<td>Mixed</td>
<td>202</td>
<td>1</td>
<td>15.03.2007</td>
<td>Helena</td>
<td>19</td>
<td>6377int010</td>
<td>6377int010</td>
</tr>
<tr>
<td>11</td>
<td>Cu1</td>
<td>35-45</td>
<td>Cuddychurch</td>
<td>Dairy</td>
<td>64</td>
<td>1</td>
<td>08.05.2007</td>
<td>Helena</td>
<td>19</td>
<td>6377int011</td>
<td>6377int011</td>
</tr>
<tr>
<td>12</td>
<td>Cu2</td>
<td>55-65</td>
<td>Cuddychurch</td>
<td>Dairy</td>
<td>189</td>
<td>2</td>
<td>08.05.2007</td>
<td>Helena</td>
<td>18</td>
<td>6377int012</td>
<td>6377int012</td>
</tr>
<tr>
<td>13</td>
<td>Cu3</td>
<td>55-65</td>
<td>Cuddychurch</td>
<td>Mixed livestock</td>
<td>76</td>
<td>1</td>
<td>08.05.2007</td>
<td>Helena</td>
<td>13</td>
<td>6377int013</td>
<td>6377int013</td>
</tr>
<tr>
<td>14</td>
<td>Cu5</td>
<td>45-55</td>
<td>Cuddychurch</td>
<td>Mixed livestock</td>
<td>198</td>
<td>1</td>
<td>09.05.2007</td>
<td>Helena</td>
<td>24</td>
<td>6377int014</td>
<td>6377int014</td>
</tr>
<tr>
<td>15</td>
<td>Cu6</td>
<td>55-65</td>
<td>Cuddychurch</td>
<td>Dairy</td>
<td>89</td>
<td>1</td>
<td>09.05.2007</td>
<td>Helena</td>
<td>14</td>
<td>6377int015</td>
<td>6377int015</td>
</tr>
<tr>
<td>16</td>
<td>Cu7</td>
<td>&gt;65</td>
<td>Cuddychurch</td>
<td>Mixed livestock</td>
<td>190</td>
<td>4</td>
<td>11.05.2007</td>
<td>Helena</td>
<td>20</td>
<td>6377int016</td>
<td>6377int016</td>
</tr>
<tr>
<td>17</td>
<td>Cu8</td>
<td>55-65</td>
<td>Cuddychurch</td>
<td>Mixed livestock</td>
<td>109</td>
<td>2</td>
<td>11.05.2007</td>
<td>Helena</td>
<td>22</td>
<td>6377int017</td>
<td>6377int017</td>
</tr>
<tr>
<td>18</td>
<td>Id1</td>
<td>55-65</td>
<td>Idle</td>
<td>Arable</td>
<td>158</td>
<td>3</td>
<td>07.02.2007</td>
<td>Quentin</td>
<td>17</td>
<td>6377int018</td>
<td>6377int018a</td>
</tr>
<tr>
<td>18</td>
<td>Id1</td>
<td>55-65</td>
<td>Idle</td>
<td>Arable</td>
<td>158</td>
<td>3</td>
<td>07.02.2007</td>
<td>Quentin</td>
<td>17</td>
<td>6377int018</td>
<td>6377int018b</td>
</tr>
<tr>
<td>19</td>
<td>Id1b</td>
<td>55-65</td>
<td>Idle</td>
<td>Arable</td>
<td>158</td>
<td>3</td>
<td>Quentin</td>
<td>22</td>
<td>6377int019</td>
<td>6377int019</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Id2</td>
<td>45-55</td>
<td>Idle</td>
<td>Dairy</td>
<td>150</td>
<td>1</td>
<td>08.02.2007</td>
<td>Quentin</td>
<td>17</td>
<td>6377int020</td>
<td>6377int020</td>
</tr>
</tbody>
</table>
Transcription template

Should:
• Possess a unique identifier
• Adopt a uniform layout throughout the research project
• Make use of speaker tags – turn-taking
• Carry line breaks
• Be page numbered
• Carry a document header giving brief details of the interview: data, place, interviewer name, interviewee details, etc.

Other considerations:
• Cover page
• Compatibility with import featured of Computer Assisted Qualitative Data Analysis Software (CAQDAS)
Model Interview Transcript:
https://www.ukdataservice.ac.uk/media/622380/ukdamodeltranscript.pdf
File formats

Choice of software format for digital data:
• Planned data analyses
• Software availability / cost
• Hardware used – e.g. audio capture
• Discipline – specific standards and customs

*Digital data is software dependent, so endangered by obsolescence of software/hardware.*

Best formats for long-term preservation:

standard, interchangeable and open

[UK Data Service optimal file formats](#) for various data types

[Digital Preservation Coalition](#) guidance on preservation formats
Organising data

- Plan in advance how to best organise data (project specific)
- 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
Data security and storage

Protect data from unauthorised:
• Access
• Use
• Change
• Disclosure
• Destruction

Who knows who is watching, listening or attempting to access your data…
Data security strategy:

• Control access to computers:
  • Use passwords and lock your machine when away from it
  • Run up-to-date anti-virus and firewall protection
  • Power surge protection
  • Restrict access to sensitive materials e.g. consent forms and patient records
  • Personal data need more protection – always keep them separate and secure
  • Utilise encryption
    • on all devices: desktops, laptops, memory sticks and mobile devices
    • at all locations: work, home and travel

• Control physical access to buildings, rooms and filing cabinets

• Properly dispose of data and equipment once the project is finished
Encryption software can be easy to use and enables users to:

- Encrypt hard drives, partitions, files and folders
- Encrypt portable storage devices such as USB flash drives

Encryption software examples:

- VeraCrypt
- BitLocker
- Axcrypt
- FileVault2

Data encryption tutorials:

https://www.youtube.com/playlist?list=PLG87Imnep1SmnFGrhAijFVHonQSvMlplpHkV
Video tutorials

- VeraCrypt: https://www.youtube.com/watch?v=Ogm9QHQpFqU
- AxCrypt: https://www.youtube.com/watch?v=ACcRlnsoYZg
- FileVault 2: https://www.youtube.com/watch?v=JIZ9EFMS0ic
- BitLocker: https://www.youtube.com/watch?v=y4losu-Yfsw
- Time Machine: https://www.youtube.com/watch?v=hlsQaVj7WtA
- MD5summer: https://www.youtube.com/watch?v=VcBfkB6N7-k
Digital back-up strategy

Consider

• **What’s backed-up?** - all, some or just the bits you change?
• **Where?** - original copy, external local and remote copies
• **What media?** - DVD, external hard drive, USB, Cloud?
• **How often?** - hourly, daily, weekly? Automate the process?
• **What method / software?** - duplicating, syncing or mirroring?
• **For how long is it kept?** - data retention policies that might apply?
• **Verify and recover** - never assume, regularly test and restore

Backing-up need not be expensive

• 2Tb external drives are around £50, with back-up software

Also consider non-digital storage too!
File sharing and collaborative environments

Sharing data between researchers
• Too often sent as insecure email attachments

Other options:
• Virtual Research Environments
  • MS SharePoint
• Locally managed; ownCloud and ZendTo
• File transfer protocol (FTP)
• Physical media
• Cloud solutions
  • Google Drive, DropBox, Microsoft OneDrive and iCloud (insecure?)
  • More secure options? Mega.nz, SpiderOak, Tresorit

• Assess risks of using cloud storage
Data disposal

Proper disposal of equipment and media
- Even reformatting a hard drive is *not* sufficient
- If in doubt, physically destroy the drive

**BCWipe** - uses ‘military-grade procedures to surgically remove all traces of any file’
  - can be applied to entire disk drives

**AxCrypt** - free open source file and folder shredding
  - Integrates into Windows well, useful for single files
UKDS data management guidance

- Best practice guidance: [www.ukdataservice.ac.uk/manage-data.aspx](http://www.ukdataservice.ac.uk/manage-data.aspx)
- Managing and Sharing Research Data – a Guide to Good Practice (Sage Publications Ltd)
- Training: [www.ukdataservice.ac.uk/news-and-events/events](http://www.ukdataservice.ac.uk/news-and-events/events)
- Twitter: @UKDSRDM
Tools and templates

Model consent form:
http://www.dataarchive.ac.uk/media/112638/ukdamodelconsent.pdf

• Survey consent statement:
http://dataarchive.ac.uk/media/147338/ukdasurveyconsent.doc

• Transcription template:
http://dataarchive.ac.uk/media/136055/ukdamodeltranscript.pdf

• Transcription instructions: http://dataarchive.ac.uk/media/285633/ukda-example-transcriptioninstructions.pdf

• Transcription confidentiality agreement:
http://dataarchive.ac.uk/media/285636/ukda-transcriber-confidenceagreement.pdf

• Data list template:
http://dataarchive.ac.uk/media/2989/UK%20Data%20Archive%20Example%20Data%20List.pdf
Training

Recurring workshops and webinars

Webinar: Data management basics
Webinar: Key issues in reusing data
Webinar: Finding and accessing data in the UK Data Service
Webinar: Guided walk through ReShare
Webinar: Key data: UK and cross-national surveys
Keep connected

• Subscribe to UK Data Service list: www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE

• Follow UK Data Service on Twitter: @UKDataService

• Follow our RDM account on Twitter: @UKDSRDM

• Youtube: www.youtube.com/user/UKDATASERVICE
Contact

Enquiries/ Help Desk:

http://ukdataservice.ac.uk/help/get-in-touch.aspx
help@ukdataservice.ac.uk

Follow us on:
https://twitter.com/UKDataService
https://www.facebook.com/UKDataService
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE
Questions?

https://pbs.twimg.com/media/B7ZUntrCUAEQAgR.jpg

https://pbs.twimg.com/media/B7ZUntrCUAEQAgR.jpg