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 don’t get to a question, we will reply later by email.
- You will be on mute throughout – we need to do this in order to ensure a high quality recording.

Data Management Basics

Libby Bishop and Scott Summers
UK Data Service
Research Data Management Team

Webinar
11 February 2016

UK Data Service

Overview of this session

Presentation

- UK Data Service
- Managing your data – why & how
  - Consent, anonymisation, documentation, etc.
  - Security, backups, encryption, etc.
- More resources available (this webinar is highlights only)
- Your questions

Data Management at UK Data Service

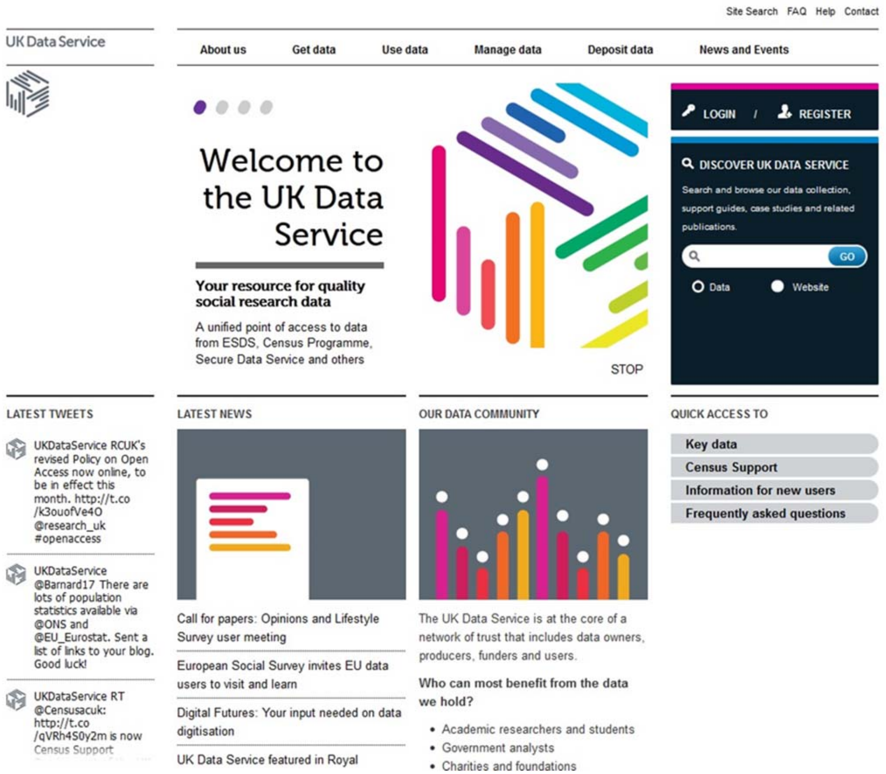
- support and training for data creators with accessing, managing, and using data

- one-stop-shop for social science data

https://discover.ukdataservice.ac.uk/

- more webinars available

https://www.ukdataservice.ac.uk/news-and-events/webinars



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

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

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

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

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

Why manage research data well ?

- Data creation in research is often expensive
- Data = cornerstone of research
- Data underpin published findings
- Good quality data = good quality research
- Protect data from loss, destruction,…
- Compliance with ethical codes, data protection laws, journal requirements, funder policies

Data sharing goes mainstream

David gave on overview of data sharing expectations from various angles. He started by referring to the Royal Society's report from 2012: *Science as an open enterprise*, which sets sharing as the standard for doing science. He then also mentioned other initiatives like the G8 Science Ministers' statement, the joint report from the Academy of Medical Sciences, BBSRC, MRC and Wellcome Trust on reproducibility and reliability of biomedical research and the UK Concordat on Open Research Data with a take-home message that sharing data and other research outputs is increasingly becoming a global expectation, and a core element of good research practice.

Wellcome Trust's policy for open data

https://unlockingresearch.blog.lib.cam.ac.uk/?p=525

Data management is sexy again #MITCDOIQ

by Elizabeth Kays | Jul 27, 2015 | 0 comments

Practical steps researchers can take

- Write a data management/sharing plan
- Make sure data are shareable and can be understood:
  - Obtain consent to share
  - Do not disclose identities without consent
  - Use open/standard formats
  - Provide context & documentation
  - Protect your data

ESRC 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, anonymisation, re-use strategies
- Copyright / Intellectual Property Ownership
- Responsibilities
- Management and curation

ESRC DMP guidance

Multiple tools for protecting identities

- Obtain informed consent, also for data sharing and long-term preservation / curation
- Protect identities e.g. anonymisation, not collecting personal data
- Regulate access where needed (all or part of data) e.g. by group, use, time period

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, video publicity
- Data sharing and archiving
  - Consider future uses of data

Always dependent on the research context – special cases of covert research, verbal consent, etc.

In practice: wording in consent form / information sheet

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 both be used. We will get your permission before using a quote from you or a photograph of you.

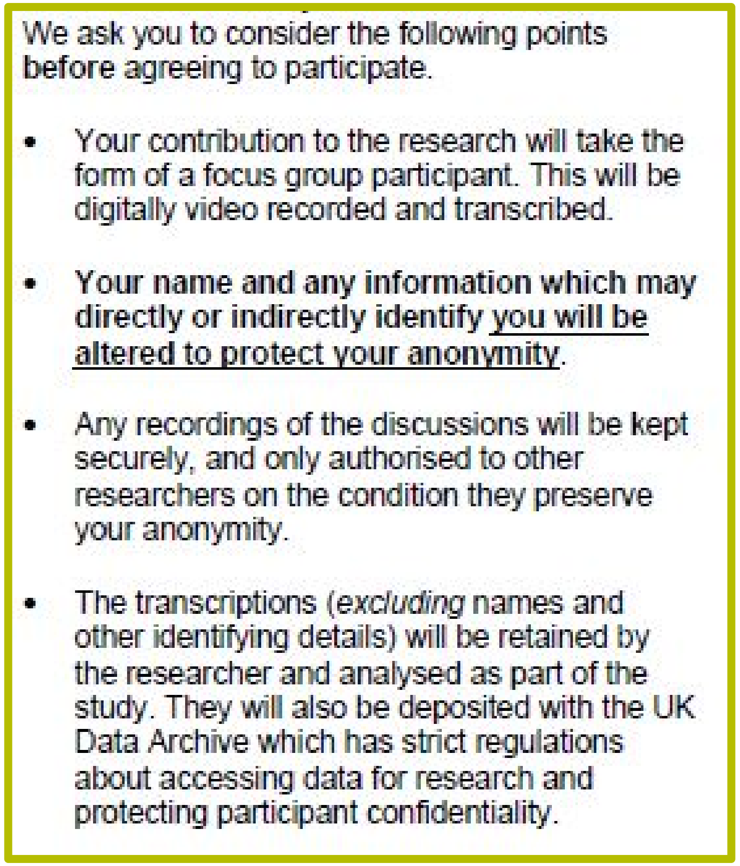
After the project has ended, we intend to archive the interviews at …. Then the interview data can be disseminated for reuse by other researchers, for research and learning purposes.

The interviews will be archived at ……. and disseminated so other researchers can reuse this information for research and learning purposes:

- ❑ I agree for the audio recording of my interview to be archived and disseminated for reuse
- ❑ I agree for the transcript of my interview to be archived and disseminated for reuse
- ❑ I agree for any photographs of me taken during interview to be archived and disseminated for reuse

In practice: wording in consent form / information sheet

Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public.



We ask you to consider the following points before agreeing to participate.

- Your contribution to the research will take the form of a focus group participant. This will be digitally video recorded and transcribed.
- **Your name and any information which may directly or indirectly identify you will be altered to protect your anonymity.**
- Any recordings of the discussions will be kept securely, and only authorised to other researchers on the condition they preserve your anonymity.
- The transcriptions (*excluding* names and other identifying details) will be retained by the researcher and analysed as part of the study. They will also be deposited with the UK Data Archive which has strict regulations about accessing data for research and protecting participant confidentiality.

ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/consent-forms.aspx

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

- Remove direct identifiers, or replace with pseudonyms – often not essential research info
- Avoid blanking out
- Identify replacements, e.g. with brackets e.g., [City A]
- Keep anonymisation log of all changes– store separately from data files
- Plan or apply editing at time of transcription
- Avoid over-anonymising –balance anonymisation with the need to preserve data integrity
- Consistency within research team and throughout project.

Audio-visual data

Digital manipulation of audio and image files can remove personal identifiers

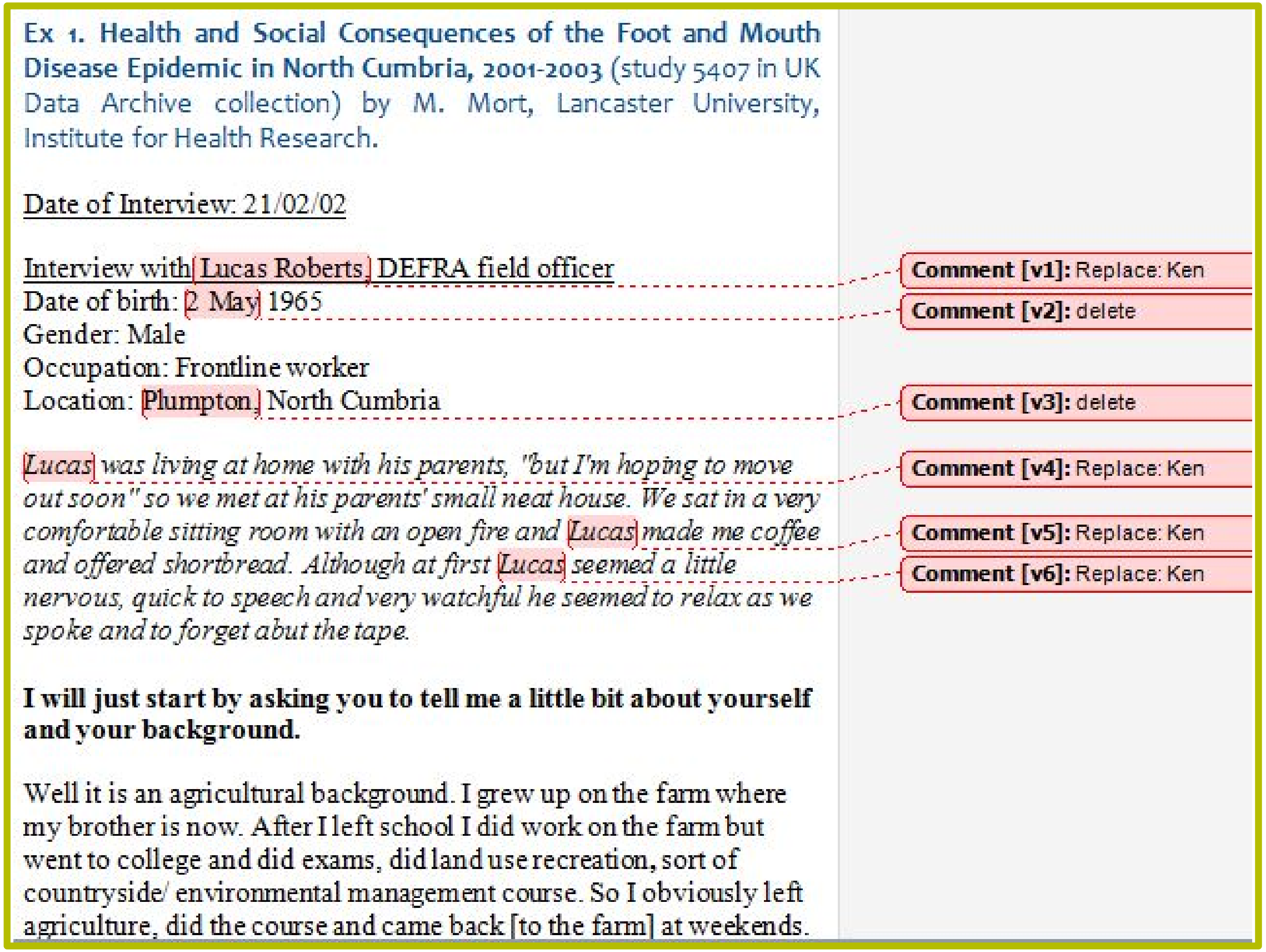
e.g. voice alteration, image blurring (e.g. of faces)

Labour intensive, expensive, may damage research potential of data

Better alternatives:

- obtain consent to use and share data unaltered for research purposes
- avoid mentioning disclosing information during audio recordings

In practice: example anonymisation



Ex 1. Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.

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 abut the tape.

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

Well it is an agricultural background. I grew up on the farm where my brother is now. After I left school I did work on the farm but went to college and did exams, did land use recreation, sort of countryside/ environmental management course. So I obviously left agriculture, did the course and came back [to the farm] at weekends.

Comment [v1]: Replace: Ken

Comment [v2]: delete

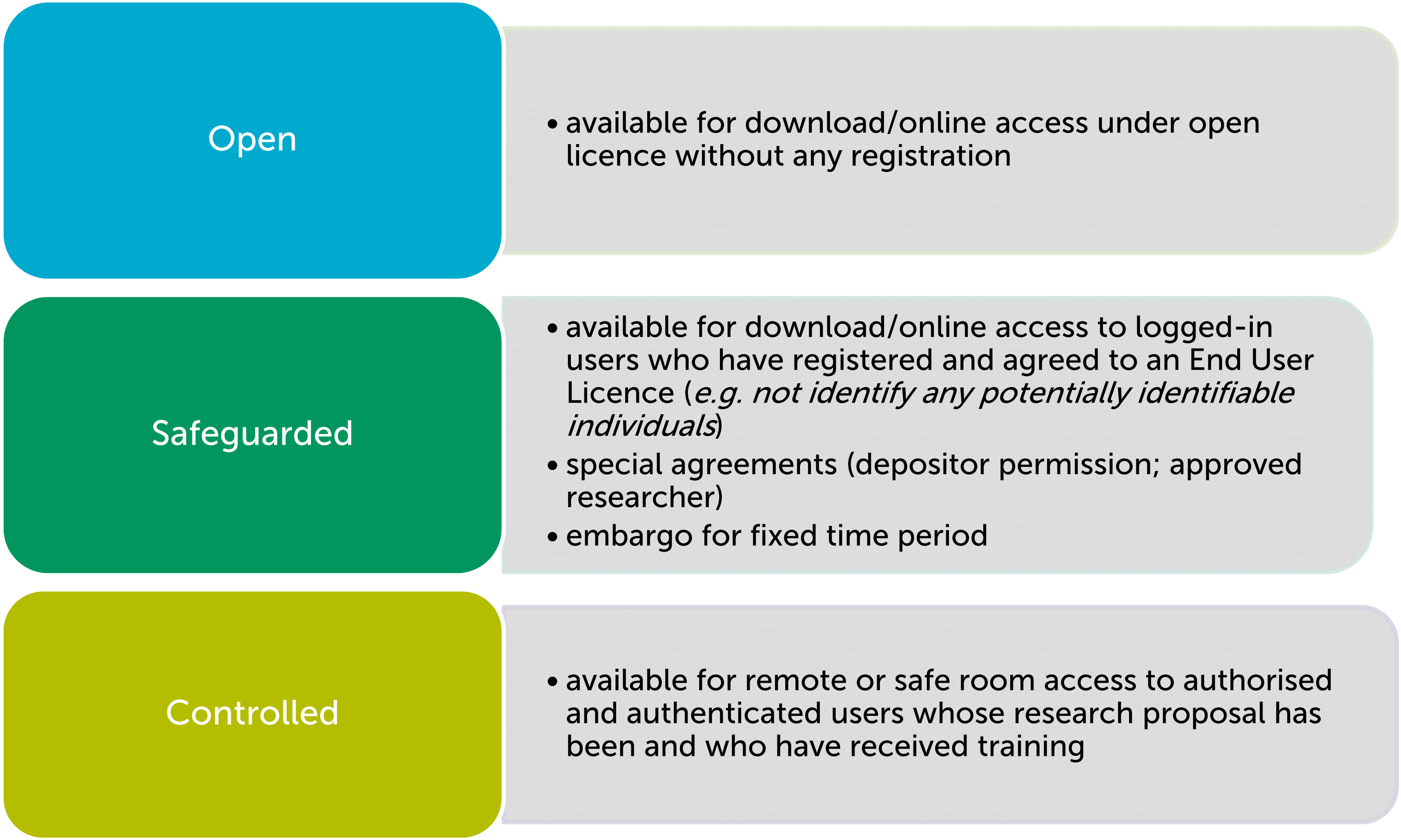
Comment [v3]: delete

Comment [v4]: Replace: Ken

Comment [v5]: Replace: Ken

Comment [v6]: Replace: Ken

Managing access to data



Open

- available for download/online access under open licence without any registration

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

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

- Interviews (audio + transcript) and written diaries with 54 people
- 40 interview and diary transcripts are archived and available for re-use by registered users
- 3 interviews and 5 diaries are embargoed until 2015
- audio files archived and only available by permission from researchers

discover.ukdataservice.ac.uk/catalogue/?sn=5407

doc.ukdataservice.ac.uk/doc/5407/mrdoc/pdf/q5407userguide.pdf

Documenting your data

- Enables you to understand data 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 such as a user guide(s) and data listing
  - ensure accurate processing and archiving
  - create a catalogue record for a published data collection

Include as documentation

- Data collection methodology and processes: sampling, sampling size, fieldwork protocol, interviewer instructions
- Information sheet / consent form
- Questionnaire, showcards, question lists
- Transcripts: header with context information: date & place interview, interviewee name, etc.
- Data list: overview of key information about each interview, as ‘at-a-glance’ summary of the data collection
- Links to reports, publications

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

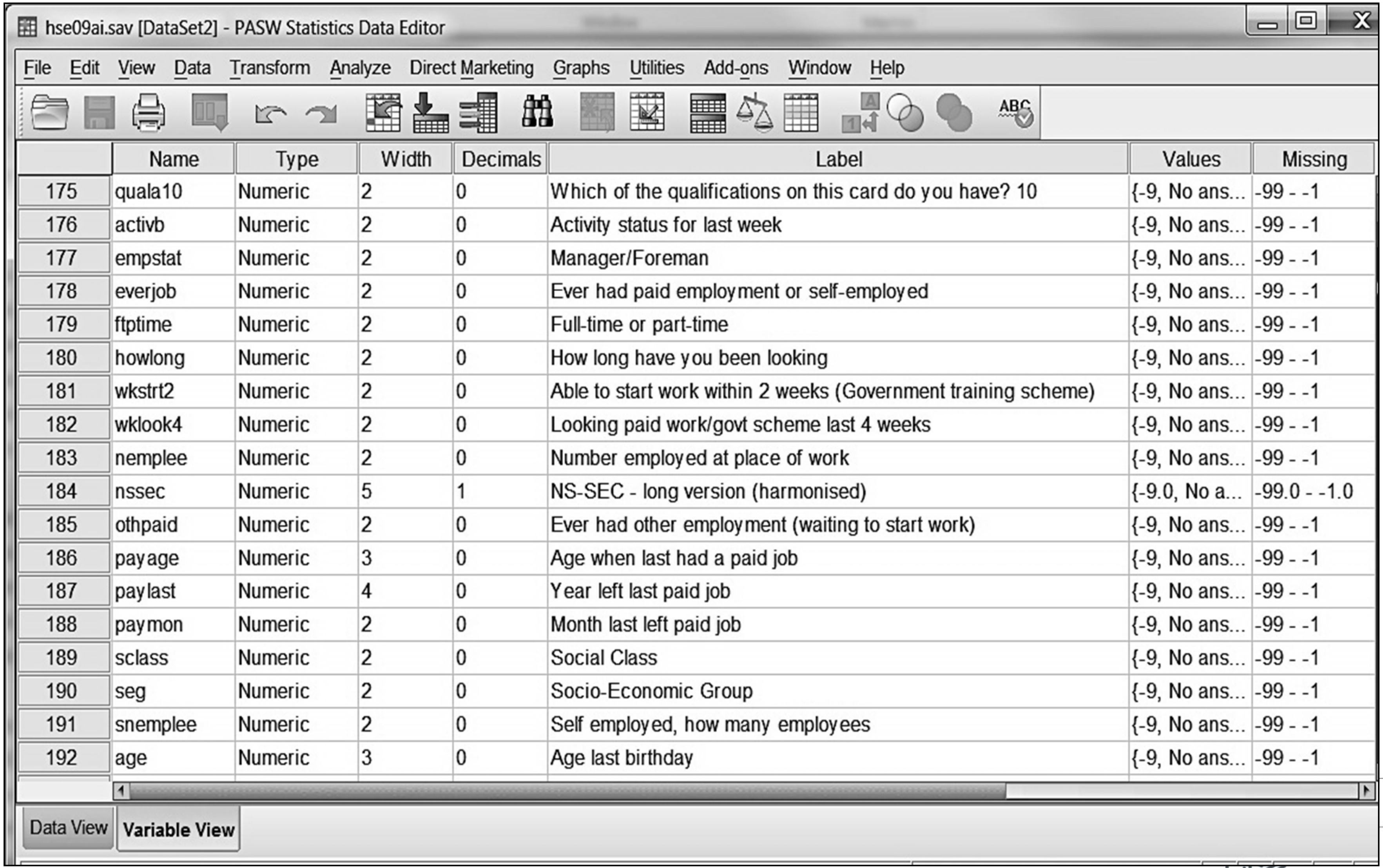
- Similar principles for variable labels:
  - be brief, max. 80 characters
  - include unit of measurement where applicable
  - reference the question number of a survey or questionnaire

    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

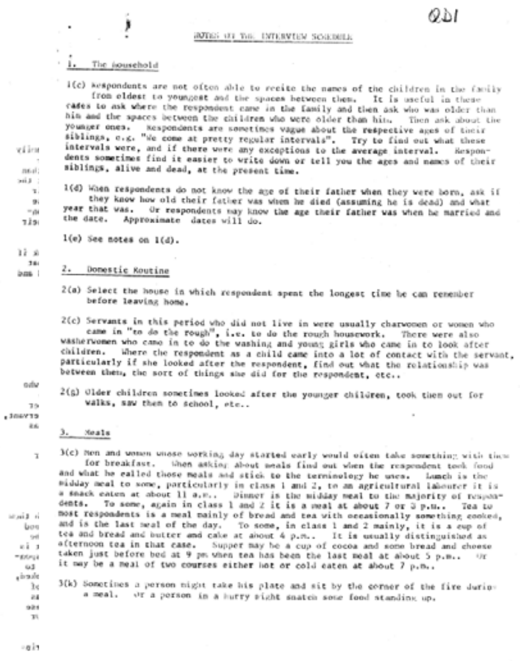
Embedded data-level metadata in SPSS file


hse09ai.sav [DataSet2] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View
Variable View

	Name	Type	Width	Decimals	Label	Values	Missing
175	quala10	Numeric	2	0	Which of the qualifications on this card do you have? 10	{-9, No ans...	-99 - -1
176	activb	Numeric	2	0	Activity status for last week	{-9, No ans...	-99 - -1
177	empstat	Numeric	2	0	Manager/Foreman	{-9, No ans...	-99 - -1
178	everjob	Numeric	2	0	Ever had paid employment or self-employed	{-9, No ans...	-99 - -1
179	ftptime	Numeric	2	0	Full-time or part-time	{-9, No ans...	-99 - -1
180	howlong	Numeric	2	0	How long have you been looking	{-9, No ans...	-99 - -1
181	wkstrt2	Numeric	2	0	Able to start work within 2 weeks (Government training scheme)	{-9, No ans...	-99 - -1
182	wklook4	Numeric	2	0	Looking paid work/govt scheme last 4 weeks	{-9, No ans...	-99 - -1
183	nemplee	Numeric	2	0	Number employed at place of work	{-9, No ans...	-99 - -1
184	nssec	Numeric	5	1	NS-SEC - long version (harmonised)	{-9.0, No a...	-99.0 - -1.0
185	othpaid	Numeric	2	0	Ever had other employment (waiting to start work)	{-9, No ans...	-99 - -1
186	payage	Numeric	3	0	Age when last had a paid job	{-9, No ans...	-99 - -1
187	paylast	Numeric	4	0	Year left last paid job	{-9, No ans...	-99 - -1
188	paymon	Numeric	2	0	Month last left paid job	{-9, No ans...	-99 - -1
189	sclass	Numeric	2	0	Social Class	{-9, No ans...	-99 - -1
190	seg	Numeric	2	0	Socio-Economic Group	{-9, No ans...	-99 - -1
191	snemplee	Numeric	2	0	Self employed, how many employees	{-9, No ans...	-99 - -1
192	age	Numeric	3	0	Age last birthday	{-9, No ans...	-99 - -1

In practice: user guide and documentation

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



QD1

NOTES ON THE INTERVIEW SCHEDULE

1. The household

1(c) Respondents are not often able to recite the names of the children in the family from oldest to youngest and the spaces between them. It is useful in these cases to ask where the respondent came in the family and then ask who was older than him and the spaces between the children who were older than him. Then ask about the younger ones. Respondents are sometimes vague about the respective ages of their siblings, e.g. "We come at pretty regular intervals". Try to find out what these intervals were, and if there were any exceptions to the average interval. Respondents sometimes find it easier to write down or tell you the ages and names of their siblings, alive and dead, at the present time.

1(d) When respondents do not know the age of their father when they were born, ask if they know how old their father was when he died (assuming he is dead) and what year that was. Or respondents may know the age their father was when he married and the date. Approximate dates will do.

1(e) See notes on 1(d).

2. Domestic Routine

2(a) Select the house in which respondent spent the longest time he can remember before leaving home.

2(c) Servants in this period who did not live in were usually charwomen or women who came in "to do the rough", i.e. to do the rough housework. There were also washerwomen who came in to do the washing and young girls who came in to look after children. Where the respondent as a child came into a lot of contact with the servant, particularly if she looked after the respondent, find out what the relationship was between them, the sort of things she did for the respondent, etc..

2(g) Older children sometimes looked after the younger children, took them out for walks, saw them to school, etc..

3. Meals

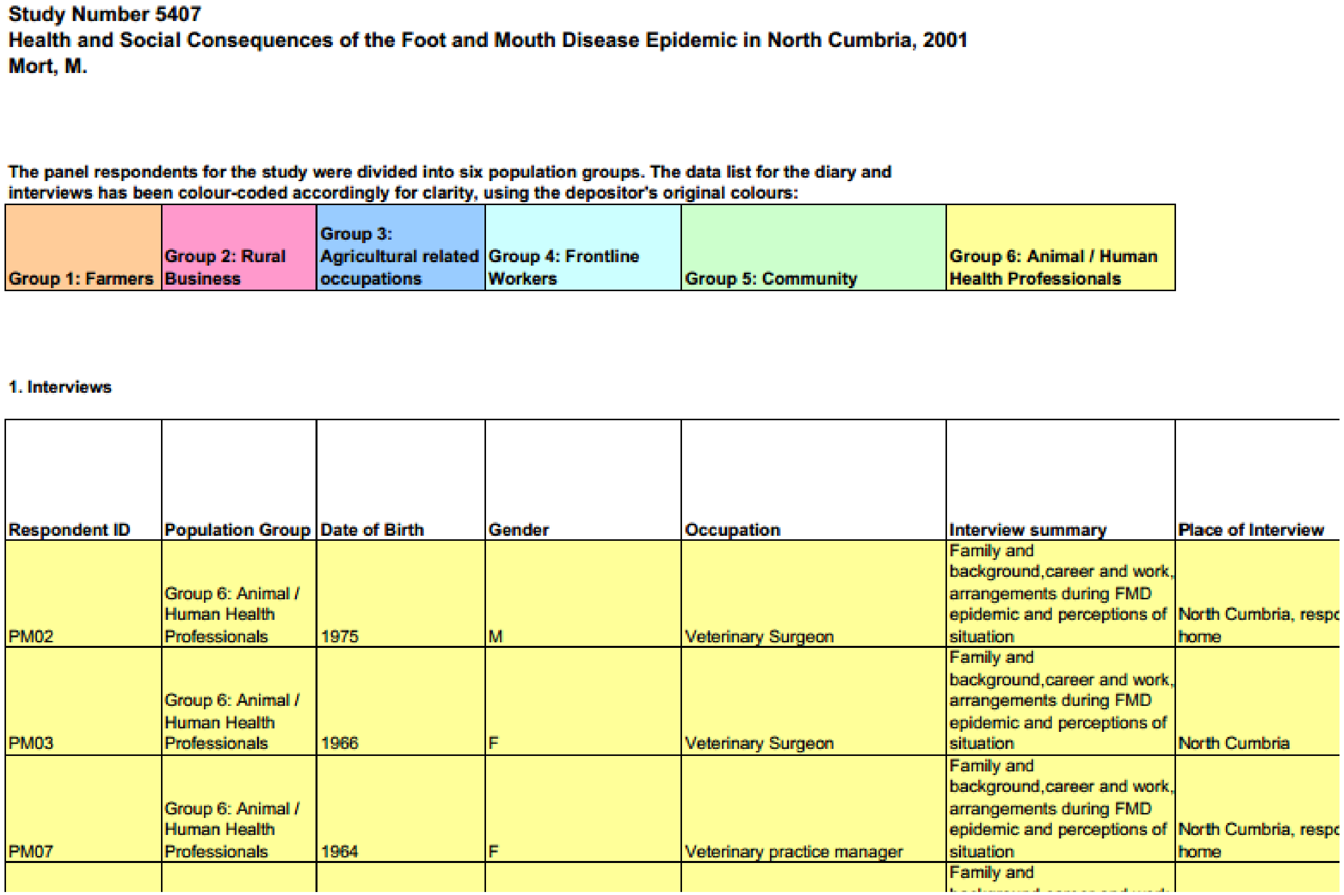
3(c) Men and women whose working day started early would often take something with them for breakfast. When asking about meals find out when the respondent took food and what he called those meals and stick to the terminology he uses. Lunch is the midday meal to some, particularly in class 1 and 2, to an agricultural labourer it is a snack eaten at about 11 a.m.. Dinner is the midday meal to the majority of respondents. To some, again in class 1 and 2 it is a meal at about 7 or 8 p.m.. Tea to most respondents is a meal mainly of bread and tea with occasionally something cooked, and is the last meal of the day. To some, in class 1 and 2 mainly, it is a cup of tea and bread and butter and cake at about 4 p.m.. It is usually distinguished as afternoon tea in that case. Supper may be a cup of cocoa and some bread and cheese taken just before bed at 9 pm when tea has been the last meal at about 5 p.m.. Or it may be a meal of two courses either hot or cold eaten at about 7 p.m..

3(k) Sometimes a person might take his plate and sit by the corner of the fire during a meal. Or a person in a hurry might snatch some food standing up.

UK Data Service

In practice: data list

- Data listing provides an at-a-glance summary of interview sets



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, respo 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, respo home
					Family and background,career and work,	

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:

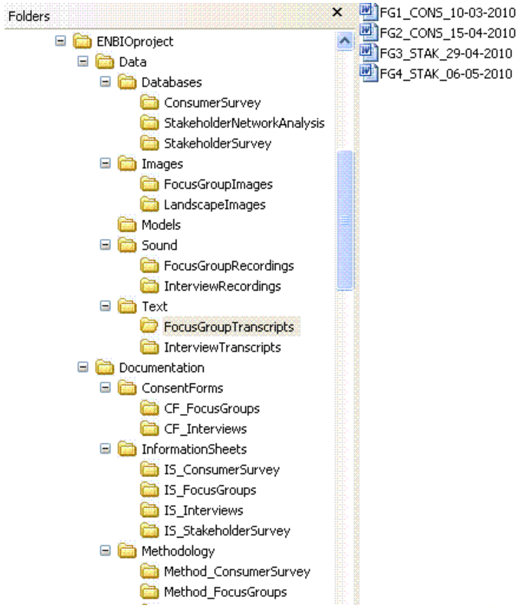
- standard, interchangeable, open
- *e.g. tab-delimited, comma-delimited (CSV), ASCII, RTF, PDF/A, OpenDocument format, XML*
- UK Data Service optimal file formats for various data types
- Digital Preservation Coalition guidance on preservation formats

Organising data

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

Examples

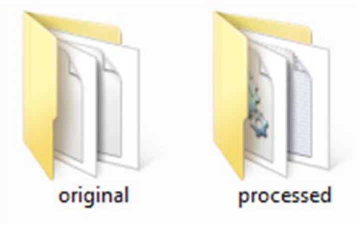
- hierarchical structure of files, grouped in folders, e.g. audio, transcripts and annotated transcripts
- survey data: spreadsheet, SPSS,
  relational database
- interview transcripts: individual
  well-named files

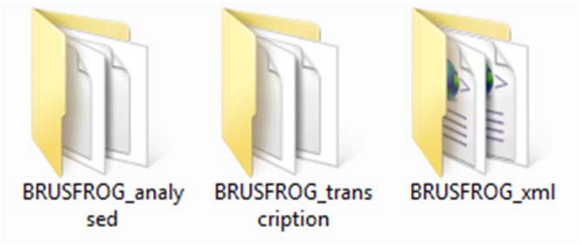


Folders

- 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

FG1_CONS_10-03-2010
FG2_CONS_15-04-2010
FG3_STAK_29-04-2010
FG4_STAK_06-05-2010





Transcription template

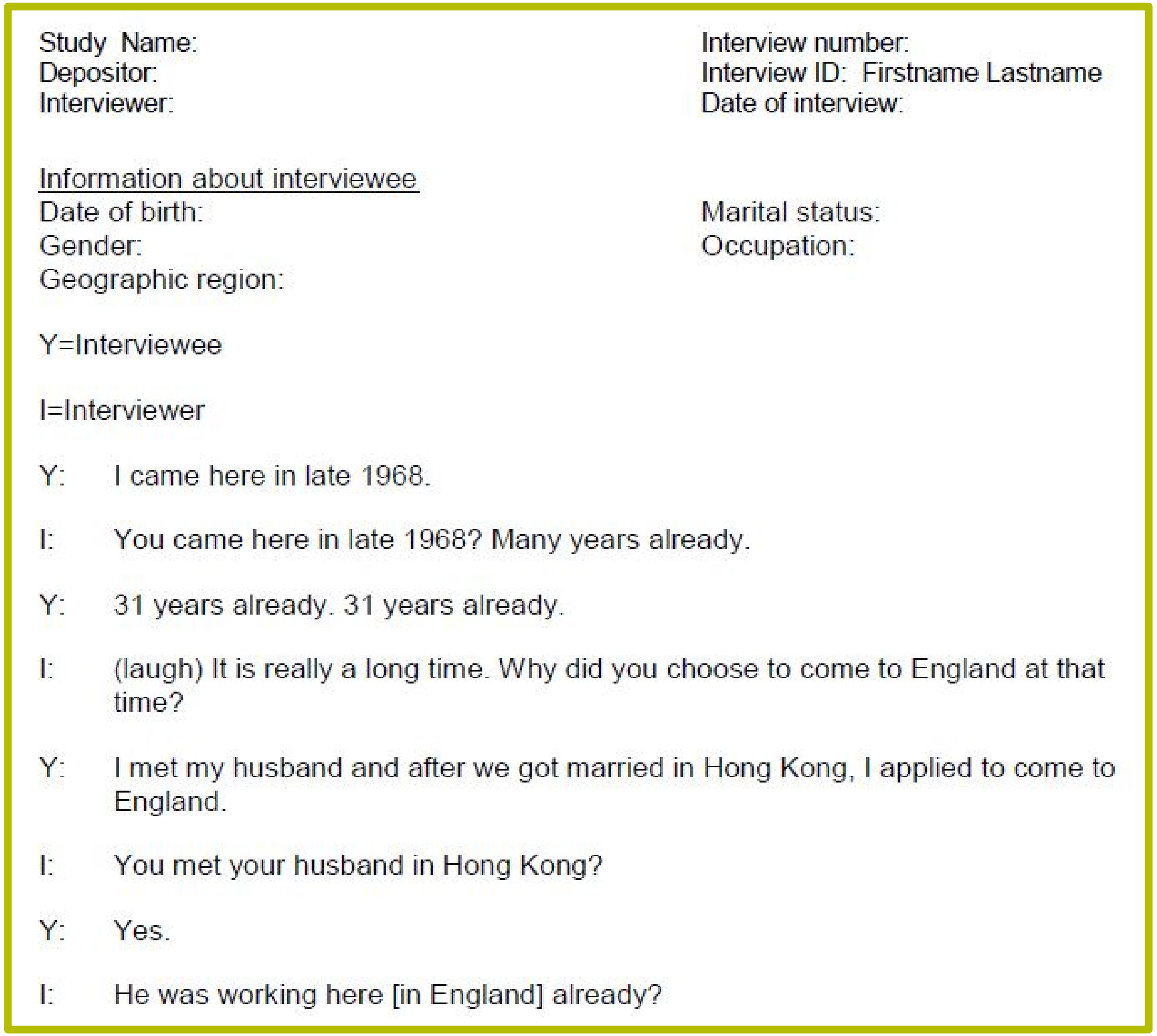
Should:

- 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: date, place, interviewer name, interviewee details, etc.

Other considerations:

- cover page
- compatibility with import features of Computer Assisted Qualitative Data Analysis Software (CAQDAS)

In practice: transcript format



Study Name:
Depositor:
Interviewer:

Interview number:
Interview ID: Firstname Lastname
Date of interview:

Information about interviewee

Date of birth:
Gender:
Geographic region:

Marital status:
Occupation:

Y=Interviewee

I=Interviewer

Y: I came here in late 1968.

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

Y: 31 years already. 31 years already.

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

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

I: You met your husband in Hong Kong?

Y: Yes.

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

Data security

Protect data from unauthorised access, change and disclosure

- control physical access to buildings, rooms, cabinets
- control access to all computers devices
  - Use passwords and lock your machine
  - Up-to-date anti-virus and firewall protection
- always encrypt personal or sensitive data
  - when moving data files
  - when or storing files

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

VeraCrypt

VeraCrypt

BitLocker

Axcrypt

FileVault2

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?
- 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

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

Also consider non-digital storage too!

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

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)
  - Securer options? - Mega.nz, SpiderOak and Tresorit
  - Assess risks of using cloud storage

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

Data Disposal

Proper disposal of equipment and media

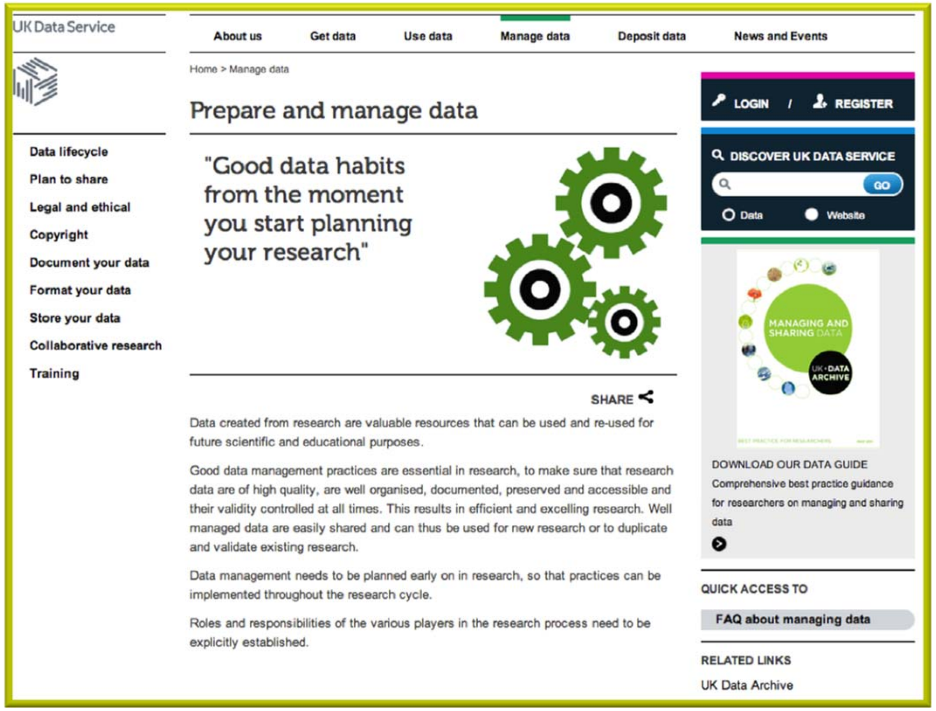
- even reformatting a hard drive is ***not*** sufficient

  - **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

- If in doubt, physically destroy the drive

Our data management guidance

- online best practice guidance: ukdataservice.ac.uk/manage-data.aspx
- Managing and Sharing Research Data – a Guide to Good Practice: (Sage Publications Ltd)
- helpdesk for queries: ukdataservice.ac.uk/help/get-in-touch.aspx
- training: www.data-archive.ac.uk/create-manage/advice-training/events



Prepare and manage data

"Good data habits from the moment you start planning your research"

Data created from research are valuable resources that can be used and re-used for future scientific and educational purposes.

Good data management practices are essential in research, to make sure that research data are of high quality, are well organised, documented, preserved and accessible and their validity controlled at all times. This results in efficient and excelling research. Well managed data are easily shared and can thus be used for new research or to duplicate and validate existing research.

Data management needs to be planned early on in research, so that practices can be implemented throughout the research cycle.

Roles and responsibilities of the various players in the research process need to be explicitly established.

DOWNLOAD OUR DATA GUIDE

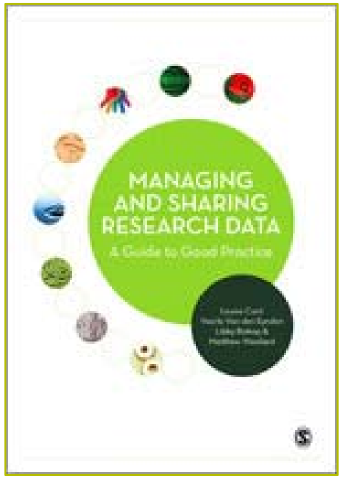
Comprehensive best practice guidance for researchers on managing and sharing data

QUICK ACCESS TO

FAQ about managing data

RELATED LINKS

UK Data Archive



Tools & templates

- Model consent form: http://www.data-archive.ac.uk/media/112638/ukdamodelconsent.pdf
- Survey consent statement: http://data-archive.ac.uk/media/147338/ukdasurveyconsent.doc
- Transcription template: http://data-archive.ac.uk/media/136055/ukdamodeltranscript.pdf
- Transcription instructions: http://data-archive.ac.uk/media/285633/ukda-example-transcription-instructions.pdf
- Transcription confidentiality agreement: http://data-archive.ac.uk/media/285636/ukda-transcriber-confidentiality-agreement.pdf
- Data list template: http://data-archive.ac.uk/media/2989/UK%20Data%20Archive%20Example%20Data%20List.pdf
- RDM costing tool: www.data-archive.ac.uk/media/247429/costingtool.pdf

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

Questions?

UK Data Service

University of Essex

ukdataservice.ac.uk/help/get-in-touch.aspx