UK Data Service

Data Curation for Quantitative and Qualitative data

Anca Vlad

UKRI Economic and Social Research Council

Overview

- Practical steps
- Prepare for archiving
- Documentation
- Metadata
- Anonymisation
- Transcription
- File formats
- Further resources
- Menti exercise

Practical steps for depositors, prepare to archive

- Write a data management plan

- Make sure data are shareable and can be understood:
  - What legal gateway will you use if you are collecting/using personal data
  - If you use consent, ensure it covers data sharing
  - 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)

Prepare for archiving

- Organise files to deposit by:
  - File type (documentation, data, readme)
  - Data type
- File formats – ensure all files you are planning to archive follow UKDS recommended and acceptable formats *
- Data quality and integrity checks – check data to ensure it meets certain quality checks such as: missing or odd characters in variable or value labels, duplicate ids, define missing values, odd characters or personal/sensitive information in string values, out of range values for categorical variables etc.
- Documentation – ensure you prepare and include rich, sufficient documentation to accompany your data.
- Start deposit process (in ReShare) as soon as possible.

*Please note, this may vary for other archives/repositories

3 Step Approach 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

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

What to 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)

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 *Mort, M. (2006). Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003. [data collection]. UK Data Service. SN: 5407* http://doi.org/10.5255/UKDA-SN-5407-1

In practice: data list

Study Number 6377

Integrated Floodplain Management, 2006-2008

Morris, J.

Floodplain farm survey

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

Metadata

- Collating and recording metadata is important for the purposes of cataloguing, citing, discovering and retrieving data collections.
- Metadata are a subset of core data documentation providing standardised, structured information.
- Metadata are intended for reading by machines, and help to explain the purpose, origin, time references, geographic location, creator, access conditions and terms of use of a data collection.

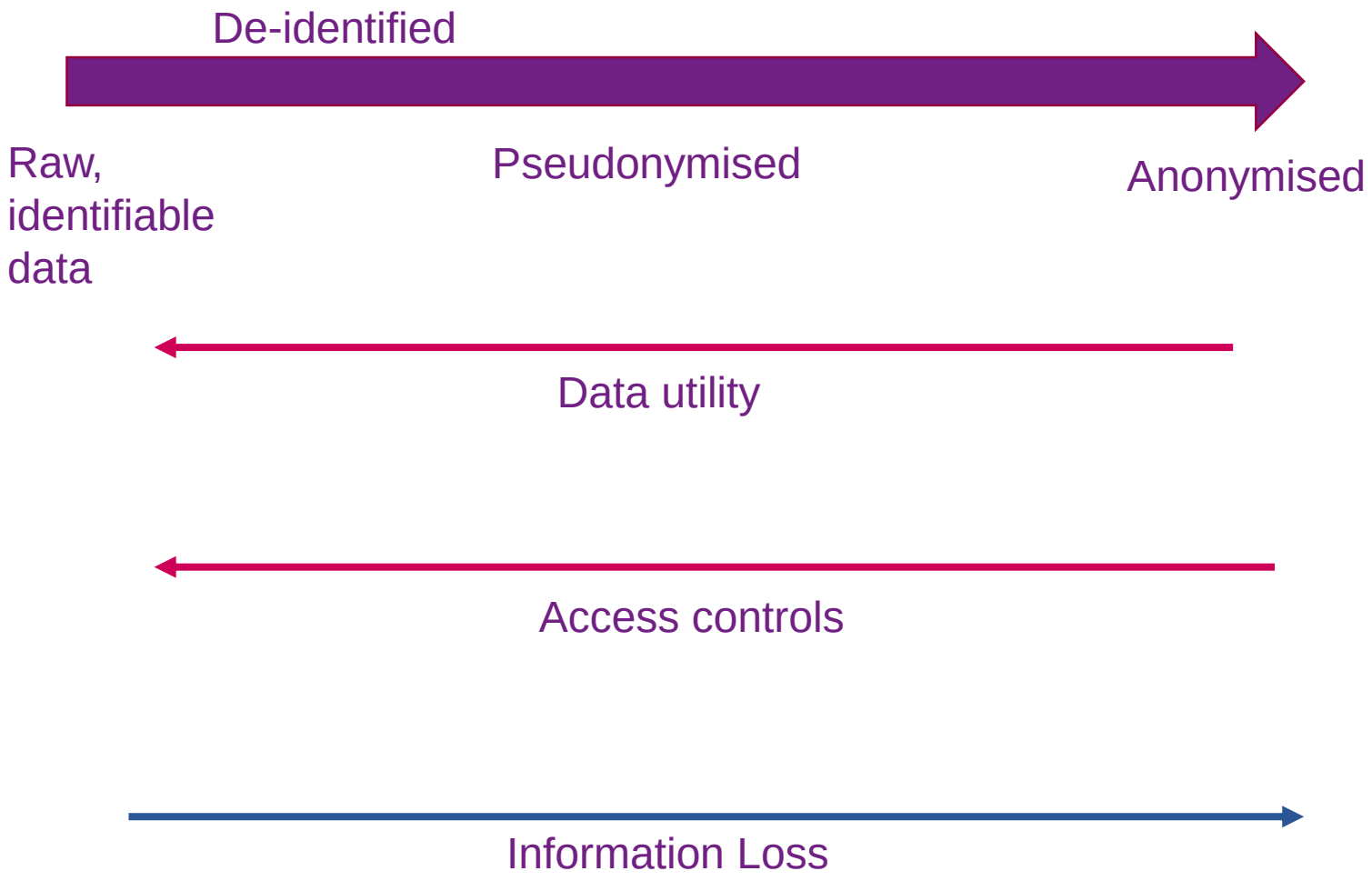
- The Data Documentation Initiative (DDI) is a rich and detailed metadata standard originally designed for describing social, behavioural and economic sciences data. It is used by most social science data archives in the world.

  UK Data Service uses DDI to structure catalogue records.

Metadata - UKDS catalogue records

- UKDS DDI records contain mandatory and optional metadata elements on the following:
  - Study description – information about the context of the data collection, such as bibliographic citation of the study and data, the scope of the study (topics, geography, time), methodology of data collection, sampling and processing, data access information, and information on accompanying materials
  - Data file description – information on data format, file type, file structure, missing data, weighting variables and software
  - Variable descriptions
- Where researchers can provide detailed and meaningful data collection titles, descriptions, keywords, contextual and methodological information in the deposit form, it helps create rich resource-discovery metadata for their deposited collections.

Anonymisation / Pseudonymisation



QAMyData

- Developed by UKDS, it is a free and easy-to-use open source tool known as QAMyData that provides a ‘health check’ for numeric data
- Uses automated methods to detect and report on some of the most common problems in survey or numeric data, such as:
  - ✓ missingness
  - ✓ duplication
  - ✓ outliers
  - ✓ direct identifiers.

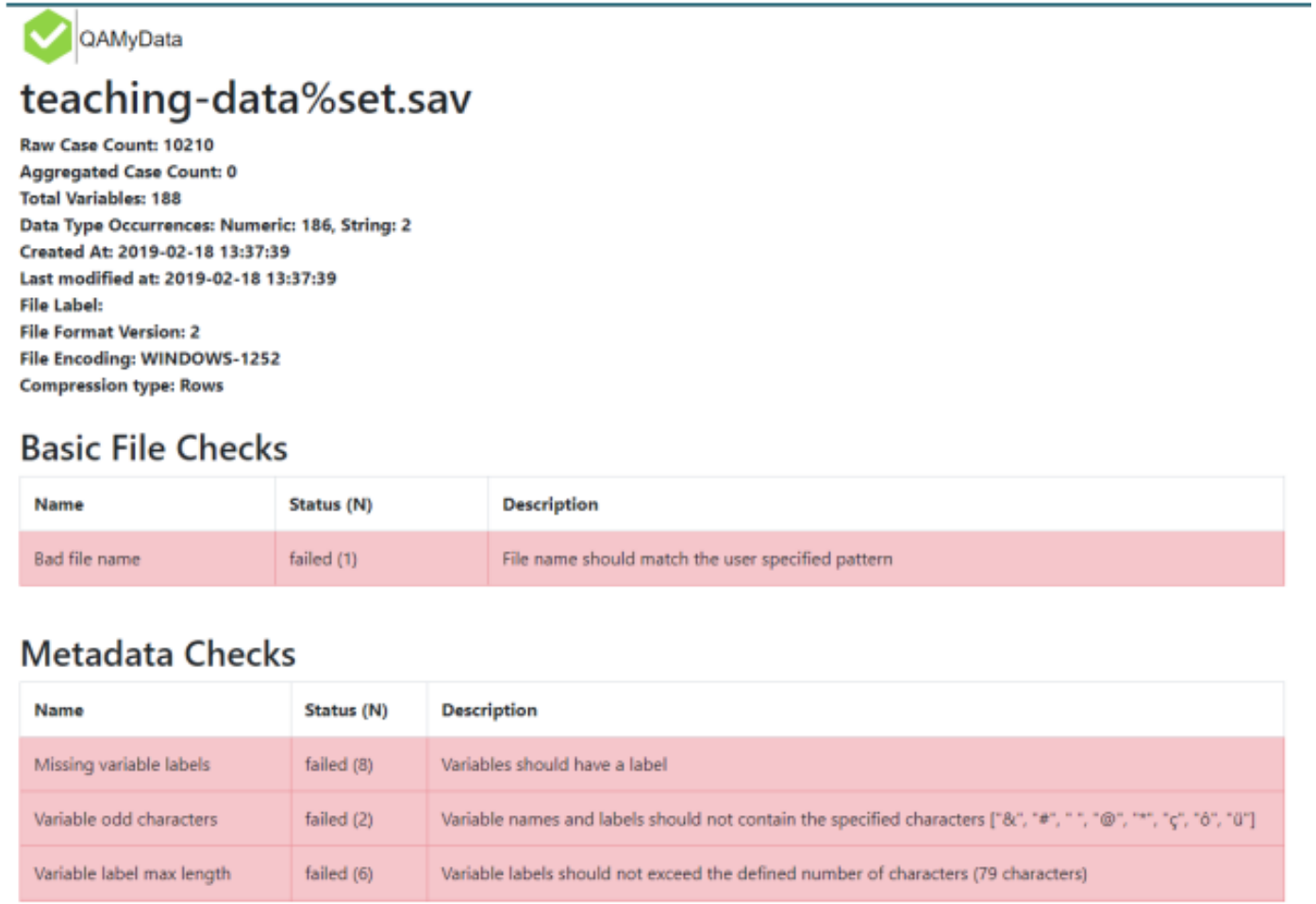
- Requirements were scoped through a series of engagement exercises with the Service’s own data curation team, other data publishers, managers and quantitative researchers to create a comprehensive list of ‘tests’ that are typically used when quality assessing numeric data files.

- For more information, see UKDS QAMyData webpage

QAMyData

- a number of configurable tests that have been categorised into four types: file, metadata, data integrity, and identifiers, which can be run on popular file formats, including SPSS, Stata, SAS and CSV
- standard config file has default settings for each test, such as a threshold for pass or fail on various tests (e.g. detect value label that are truncated, email addresses identified as a string, or undefined missing values) which can be easily adapted to meet the user's own desired thresholds.
- The software creates a 'data health check' that details errors and issues as both a summary and detailed report, providing a location of the failed test. New tests can easily be added. Data depositors and publishers can act on the results and resubmit the file until a clean bill of health is produced.

QAMyData result



QAMyData

teaching-data%set.sav

Raw Case Count: 10210
Aggregated Case Count: 0
Total Variables: 188
Data Type Occurrences: Numeric: 186, String: 2
Created At: 2019-02-18 13:37:39
Last modified at: 2019-02-18 13:37:39
File Label:
File Format Version: 2
File Encoding: WINDOWS-1252
Compression type: Rows

Basic File Checks

Name	Status (N)	Description
Bad file name	failed (1)	File name should match the user specified pattern

Metadata Checks

Name	Status (N)	Description
Missing variable labels	failed (8)	Variables should have a label
Variable odd characters	failed (2)	Variable names and labels should not contain the specified characters ["&", "#", " ", "@", "*", "ç", "ö", "ü"]
Variable label max length	failed (6)	Variable labels should not exceed the defined number of characters (79 characters)

For more information see QAMyData User Guide

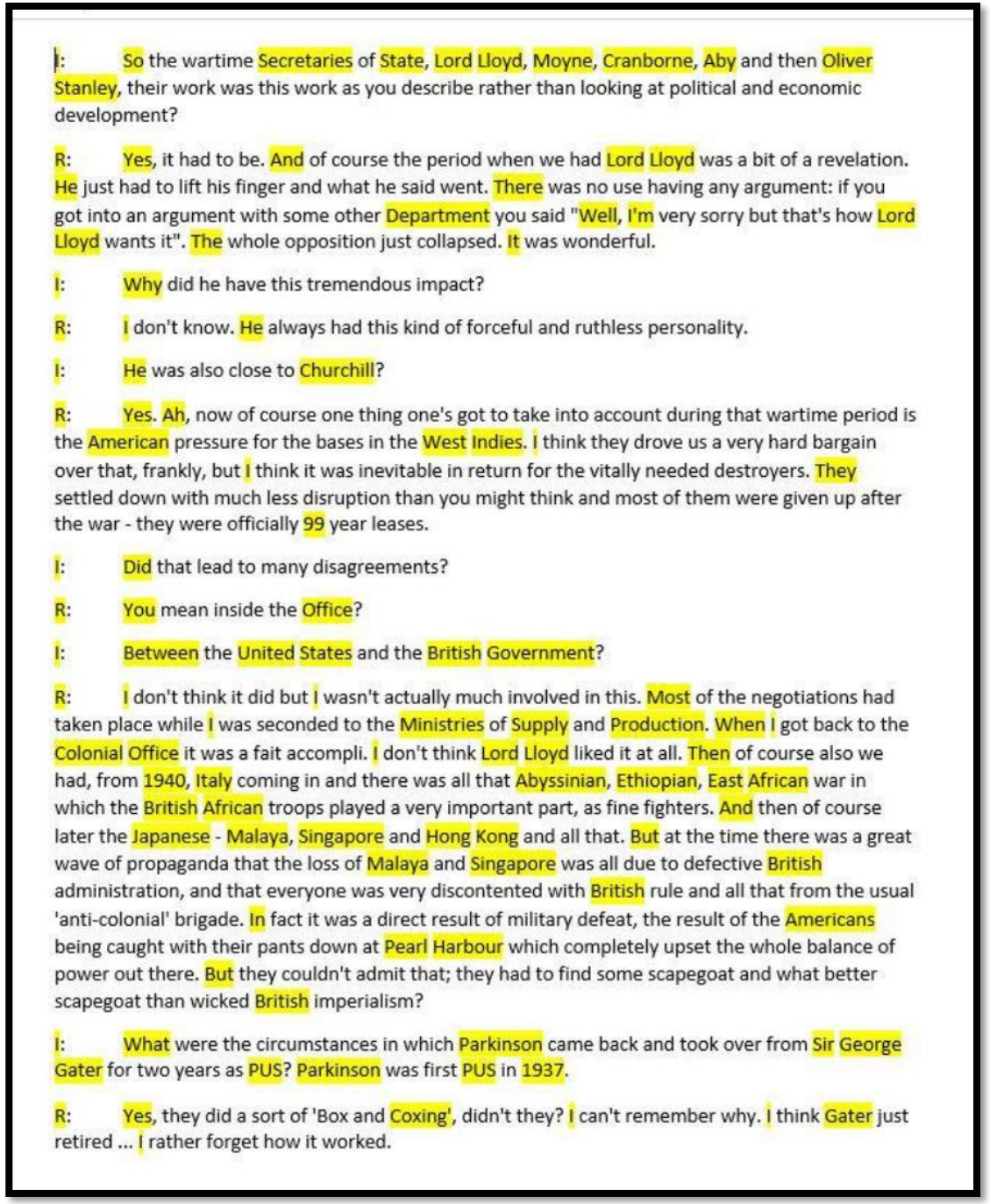
Anonymising qualitative data: some tips

- Plan or apply editing at time of transcription (*except: longitudinal studies to ensure linkages)*
- 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
- Avoid blanking out; use pseudonyms or replacements
- Avoid over-anonymising - removing/aggregating information in text can distort data, make them unusable, unreliable or misleading
- See our Workshop on Data Anonymisation our join our next online training event.

Controlling access is a better option than over-anonymizing!

Text Anonymisation Tool for qualitative data

- Aid to find and remove disclosive information from the dissemination copy of data files
- Finds all numbers and words starting with capital letters
- Does not identify names, organisations, dates or place names
- Does not make changes to data files
- For more info and access see here and here.



I: So the wartime Secretaries of State, Lord Lloyd, Moyne, Cranborne, Aby and then Oliver Stanley, their work was this work as you describe rather than looking at political and economic development?

R: Yes, it had to be. And of course the period when we had Lord Lloyd was a bit of a revelation. He just had to lift his finger and what he said went. There was no use having any argument: if you got into an argument with some other Department you said "Well, I'm very sorry but that's how Lord Lloyd wants it". The whole opposition just collapsed. It was wonderful.

I: Why did he have this tremendous impact?

R: I don't know. He always had this kind of forceful and ruthless personality.

I: He was also close to Churchill?

R: Yes. Ah, now of course one thing one's got to take into account during that wartime period is the American pressure for the bases in the West Indies. I think they drove us a very hard bargain over that, frankly, but I think it was inevitable in return for the vitally needed destroyers. They settled down with much less disruption than you might think and most of them were given up after the war - they were officially 99 year leases.

I: Did that lead to many disagreements?

R: You mean inside the Office?

I: Between the United States and the British Government?

R: I don't think it did but I wasn't actually much involved in this. Most of the negotiations had taken place while I was seconded to the Ministries of Supply and Production. When I got back to the Colonial Office it was a fait accompli. I don't think Lord Lloyd liked it at all. Then of course also we had, from 1940, Italy coming in and there was all that Abyssinian, Ethiopian, East African war in which the British African troops played a very important part, as fine fighters. And then of course later the Japanese - Malaya, Singapore and Hong Kong and all that. But at the time there was a great wave of propaganda that the loss of Malaya and Singapore was all due to defective British administration, and that everyone was very discontented with British rule and all that from the usual 'anti-colonial' brigade. In fact it was a direct result of military defeat, the result of the Americans being caught with their pants down at Pearl Harbour which completely upset the whole balance of power out there. But they couldn't admit that; they had to find some scapegoat and what better scapegoat than wicked British imperialism?

I: What were the circumstances in which Parkinson came back and took over from Sir George Gater for two years as PUS? Parkinson was first PUS in 1937.

R: Yes, they did a sort of 'Box and Coxing', didn't they? I can't remember why. I think Gater just retired ... I rather forget how it worked.

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)

In practice: transcript format

Study Name: Interview ID:
Depositor: Date of Interview:

Information about interviewee:
(e.g. Age, Gender, Occupation, Marital Status, Geographic region, etc. as relevant /appropriate)

R= Respondent/Interviewee *(if more than one respondent, use R1, R2, etc.)*
I=Interviewer

R: I came here in late 1968.

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

R: 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?

R: 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?

R: Yes.

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

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

I: You were very brave to think about going abroad as you were so young at that time.

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

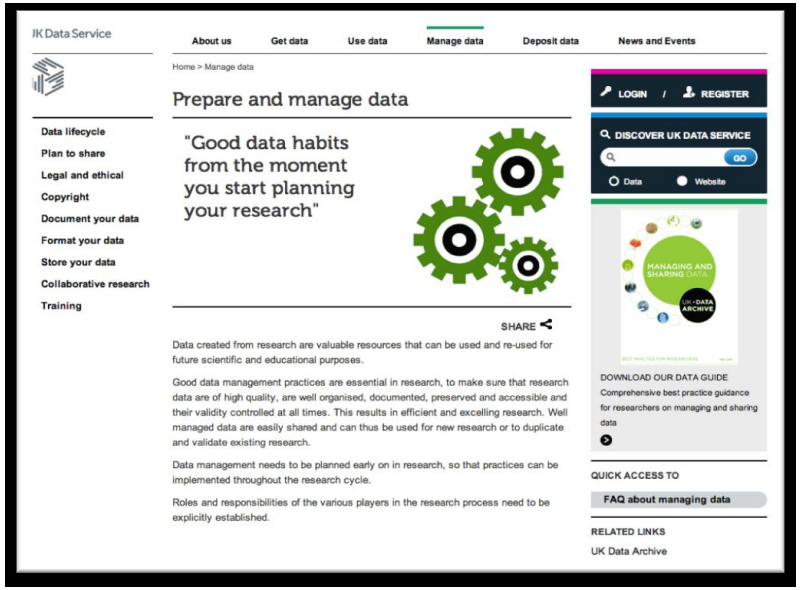
Digital data is software dependent, so endangered by obsolescence of software/hardware.

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

UKDS data management guidance

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



Tools and templates

- Model consent form and survey consent statement: https://ukdataservice.ac.uk/learning-hub/research-data-management/ethical-issues/consent-for-data-sharing/
- Transcription template: https://dam.ukdataservice.ac.uk/media/622380/ukdamodeltranscript.pdf
- Transcription instructions: https://ukdataservice.ac.uk//app/uploads/ukda-example-transcription-instructions.pdf
- Transcription confidentiality agreement: https://dam.ukdataservice.ac.uk/media/622354/ukda-transcriber-confidentiality-agreement.pdf
- Data list template: https://ukdataservice.ac.uk/uk_data_archive_data_listing_template

Further resources

- Research data management - UKDS Guidance
- Data Management Expert Guide – CESSDA
- DMPOnline - Digital Curation Centre
- Anonymising Research Data - ESRC National Centre for Research Methods, Working Paper 7/06
- Guide to Social Science Preparation and Archiving from the Inter-University Consortium for Political and Social Research
- UKDS Data management costing tool

Past and upcoming events

Workshops and Conferences

Past events: (slides and recording available)

How to anonymise qualitative and quantitative data

Data management basics: Introduction to data management and sharing

Data management basics: Ethical and legal issues in data sharing

Depositing your data with ReShare

Introduction to copyright: Copyright issues in secondary data use

Mapping crime data in R: An Introduction to GIS and spatial data

Safe Researcher Training

Census 2021: What to expect and when

Family Finance Surveys User Conference 2022

12 July 2022, 9:30 AM - Health Studies User Conference 2022

Get connected

http://ukdataservice.ac.uk/about-us/contact.aspx

https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKdataservice

https://twitter.com/UKDataService
@UKDataService

https://www.facebook.com/UKDataService

https://www.youtube.com/user/UKDATASERVICE

UK Data Service

Thank you.

Anca Vlad
advlad@essex.ac.uk

Mentimeter exercise

- You can use any device (phone, iPad etc.)
- Go to www.menti.com
- Use code: 9178 8667