Documenting and organising research data for archiving and reuse

Anca Vlad
Research Data Publishing Officer,
ReShare Administrator

Managing and sharing research data: best practice for data protection
London, 28-29 November 2018

UK Data Service

Overview

- Short overview of the UK Data Service
- Documenting and organising data
  - Project-level documentation
  - Data-level documentation
  - Metadata
  - File naming
  - File formats
- Bulk Rename – Exercise
- QAMyData – overview and demo

UK Data Service

UK Data Service - About

- Curate, preserve, provide access to social science data for reuse
- Funded by UKRI Economic and Social Research Council
- Data management advice for data creators
- Support for users of the service
- Information about the use to which data are put

Website, data catalogue: https://www.ukdataservice.ac.uk/

UK Data Service

Some statistics about our UK Service

- **7,300** datasets in the collection
- **1050** qualitative and mixed methods collections
- **400** new datasets added each year
- **219** case studies of data reuse
- **25,000** registered users
- **60,000** downloads worldwide per year
- **4000+** user support queries per year

UK Data Service

Documentation

Planning ahead will save you time and help keep things organised. A useful exercise is to think about the information that a stranger to the project would need in order to understand, replicate or reuse the data.

Planning to archive the data at the end of the project and where in particular can be useful to know in advance, as guidance/formats/metadata standards can vary across repositories.

- Project-level documentation includes information about the study, what were the main research questions, type of data was collected to answer these questions.

- Data-level documentation includes information at the level of individual data files, such as an interview transcript or a particular variable in a dataset.

Project-level documentation

- Purpose of collection (research questions)
- Data collection contents (summary of files: kind, size, formats, linked files) – ReadMe file: http://reshare.ukdataservice.ac.uk/help/index.html
- Data collection process (methodology, data sources, sample size, sampling design, population of interest)
- Data processing (cleaning, coding, anonymization, any project specific details)
- Quality assurance (transcription errors, data integrity checks, weighting, properly coding missing values)
- Accessibility (data repository, access conditions, confidentiality, copyright/ownership, citation, permanent identifiers)

Data-level documentation

Qualitative data

- information provided at the beginning of each data unit (each transcript/fieldnote)
- data list contains demographic information about the participants (also useful to map the data)

Quantitative data

- Variable-level annotation: variable type (numeric/string), description(label), values, description of derived variables, information about missing values for each variable;
- Variable labels should be succinct, and include the question number (if applicable from a questionnaire).

Transcript template

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?

UK Data Service

Data list example

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
21	Id2b	45-55	Idle	Dairy	150	1		Quentin	19	6377int021	
22	Id3	35-45	Idle	Arable	680	3	01.02.2008	Helena	27	6377int022	6377int022
23	Mo1	>65	Morda	Mixed	138	1	12.03.2007	Helena	31	6377int023	6377int023
24	Mo3	35-45	Morda	Arable	152	2	13.03.2007	Helena	16	6377int024	6377int024
25	Mo4	55-65	Morda	Mixed livestock	122	1	13.03.2007	Helena	19	6377int025	6377int025
26	Mo5	>65	Morda	Mixed	142	2	14.03.2007	Helena	14	6377int026	6377int026
27	Mo6	>65	Morda	Mixed livestock	19	1	14.03.2007	Helena	15	6377int027	6377int027
28	Mo7	<35	Morda	Dairy	74	1	30.05.2007	Helena	22	6377int028	6377int028
29	Mo8	55-65	Morda	Mixed livestock	48	4	30.05.2007	Helena	19	6377int029	6377int029
30	Mo9	>65	Morda	Mixed	278	1	31.05.2007	Helena	20	6377int030	6377int030
31	Mo10	35-45	Morda	Beef	81	1	01.06.2007	Helena	21	6377int031	6377int031
32	Mo11	35-45	Morda	Mixed	109	2	01.06.2007	Helena	22	6377int032	6377int032
33	Mo12	35-45	Morda	Mixed livestock	51	1	11.10.2007	Helena	34	6377int033	
34	Ro1	>65	Rother	Rent and let out	57	1	20.03.2007	Helena	19	6377int034	6377int034
35	Ro3	>65	Rother	Sheep	49	1	21.03.2007	Helena	33	6377int035	6377int035
36	Ro4	35-45	Rother	Mixed	182	1	21.03.2007	Helena	24	6377int036	6377int036

File naming conventions and best practice

- how to best organize files depends on the plan and organization of the study
- file name/base name(not including file format extension) = principal identifier of file
- use logical naming i.e. easy to identify and retrieve the file
- naming provides organisation, context & consistency
- name elements: version number, date, content description, creator name

Best practice:

- meaningful & brief
- name independent of location
- relevant to content
- no special characters, dots or spaces
- avoid using space, for separation use underscores
- dates used should be in format YYYY-MM-DD;
- include versioning (when appropriate) via filename: ascending, decimal version numbers
- avoid very long file names

Exercise – Bulk Rename

- file renaming software
- for Windows computers
- makes it possible to easily rename multiple files at once using your own criteria

- Features:
  - Add date/time stamps
  - Add location
  - Replace unwanted characters with acceptable ones
  - Delete unwated characters
  - Add preffixes and/or suffixes
  - Convert Cases
  - Remove or change file extensions

- Exercise: Use the Bulk Rename Utility to edit the names of all files in the folder Day2\Exercises\BulkRename\Transcripts to **InterviewTranscript_London_Nov2018.rtf**

Metadata

‘Data about data’ – machine readable information needed to catalogue and discover the data. This information can contain (DDI compliant):

- Title (and alternate title)
- Funding
- Principal investigator
- Kind of data (numeric/text/audio/ video etc.)
- Keywords
- Topic
- Unit of observation
- Data type (historical/census/business microdata/survey/ longitudinal etc.)
- Temporal coverage (time period, date of collection)
- Geographic coverage
- Access

FAIR principles for repositories

- Findable
- Accessible
- Interoperable
- Re-usable

https://www.force11.org/group/fairgroup/fairprinciples

UK Data Service

FAIR principles for repositories

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

QAMyData

- Software designed to assess data quality and review disclosure risk in quantitative data
- Useful to prepare high quality data for publishing (to allow replication & reuse) and check unknown data prior to reuse
- Ensures consistency and improves quality, it enables correct and meaningful analysis and it can reduce disclosure risk
- Thresholds can be customised, to indicate what version is prepared to accept
- Issues are identified in a summary/detailed report useful to improve the data file.

QAMyData

- File checks: virus check, file open, bad file name check, document embedded in file
- Metadata checks: missing variable labels, invalid variable names, missing value labels.
- Data integrity checks:
  - Number of cases and variables, number of numeric and string variables
  - Incorrect format of numeric variables, odd characters
  - Spelling mistakes and truncation, format same for all cases
  - Values outlying the listed code values, empty variables, undefined missing values
- Disclosure checks:
  - Direct identifiers
  - Disclosive outliers
  - Frequencies less than agreed threshold
  - Unique values in continuos variables
  - Key-identifiers and non-identifying variables (combinations of key variables)
  - String variables
  - Open-ended String variables

UK Data Service

QAMyData – Example checks

- Number of variables (columns)
- Number of cases (rows)
- Variable has all rows missing
- Length of variable label e.g. > 79
- Length of category label e.g. > 39
- If a date type variable and includes day of month
- If a string variable and each row is unique
- If likely to be open text
- String or open text contains a postcode
- String (any label, name, category, open text) contains invalid characters
  - Variable name includes trailing spaces, #, ! etc
  - Non ASCII characters

- Demo

UK Data Service

Questions?

https://pbs.twimg.com/media/B7ZUntnCUAEQAgR.jpg

UK Data Service

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