Love Research - Love Data: getting the most out of sharing and using data

Louise Corti
UK Data Service

Exeter Qestival
13 September 2019

UK Data Service

Research scenario: a controversial area

You have undertaken a research and clinical trials programme that investigates the epidemiology and treatment of Chronic Fatigue Syndrome CFS/ME in young adults.

The work is funded by the Medical Research Council (MRC) and you are expected to share the data

Can you share the data and what are the key issues?

Divide into 2 teams:

Opportunities

Problems

10 minutes

UK Data Service

Opportunities

- Improved health outcomes through validated research
- Help progress science, building on a foundation of trusted evidence, which can be connected
- Avoid expensive data collection and duplication
- Help reduce the burden for already over-researched patient groups
- Opportunities for novel research through new unanticipated analyses
- Provide greater research transparency/reproducibility/accountability
- Enhance academic impact and credibility, with opportunities for further funding (and risk of not getting research papers published or losing funds if data not shared
- Enable data harmonisation
- Do due diligence for patients who have voluntarily contributed data
- Real life secondary data can bring teaching to life

Problems

- Risk of disclosing personal, confidential/ sensitive information, especially when linked to public sources
- Securing data sharing contracts with local legal entities
- Being ‘forced’ to share data can lead to risk of misuse of data, with the possibility of damaging outcomes for patients and for data owners
- Lack of detailed knowledge of how to prepare a clean dataset for sharing
- Lack of funds and resource to prepare older datasets
- Limiting consent statements in earlier studies

Motivating factors to share data in a repository

MORE IMPORTANT

LESS IMPORTANT

Extra funding to cover costs

established researchers ~ cellular, developmental and physiological sciences, genetic and molecular science, neuroscience and mental health, population health

infection and immunobiology

Enhanced academic reputation

early career researchers ~ researchers not sharing data now

Knowing how other people use data

early career researchers ~ LMIC researchers ~ cellular, developmental and physiological sciences, humanities, infection and immuno-biology, population health

genetic and molecular science

Co-authorship on reuse papers

early career researchers clinical, population health, social science researchers cellular, developmental & physiological sciences, neuroscience and mental health

biomedical and humanities researchers, genetic and molecular science, infection and immunobiology

Case study that showcase data

LMIC researchers ~ humanities, Infection and immuno-biology, population health

cellular, developmental and physiological sciences, genetic and molecular science, neuroscience and mental health

Data deposit leads to data paper publication

early career researchers; LMIC researchers ~ cellular, developmental and physiological sciences, infection and immuno-biology, neuroscience and mental health

genetic and molecular science, humanities and social sciences

✗ Too difficult to share data widely?

Ethical, legal and research integrity challenges

- Personal, confidential or sensitive information
- Linkage risk of data with other available sources
- Lack of trust in others using your data
- Efforts to prepare compete with the science
- Lack of practical experience/knowledge in preparing and publishing a dataset

✓ Ethical arguments *for* sharing data

- Provide greater research transparency/reproducibility/accountability to funders
- Not burden already over-researched, vulnerable groups
- Make best use of hard-to-obtain data
- Extend voices of participants

In each, ethical duties to participants, peers and public may be present

Strategies for enabling safe access to data

- ✓ Obtain informed consent for participation or sharing personal data
- ✓ Protection of identities when promised
- ✓ Processing ground for personal data
- ✓ Regulated access where needed (all or part of data) by group, use, time
- ✓ Safeguards and security

Open where possible, closed when necessary

Managing research data well?

- Good quality data leads to good quality research
- Data underpins published findings
  - Documentation can be used in dissertation write-up
  - Documentation can be used in a viva
- Helps promote discussion in dissertation supervision about how to collect and analyse data
- Protect data from loss, destruction and potential exposure
- Enables compliance with ethical codes/data protection law
- Enhances transparency of research and can authenticate your dissertation progress

Practical steps you can take

- Consider how to manage your data early

- Make sure you can understand your data and it is protected:
  - obtain consent to share data with your supervisor/ project colleagues/
  - do not disclose identities without consent
  - provide clear documentation
  - create a datalist
  - store your data safely at all stages

Promising ‘anonymity’

- Once ‘anonymised’, data falls out of data protection legislation
- Not all research data can be fully or easily anonymised/de-identified
  - Combinations of unique key attributes
  - Rich textual data
  - Combining data from different sources

Anonymising quantitative data

- remove direct identifiers

  e.g. names, address, institution and photos
- reduce the precision/detail through aggregation

  e.g. birth year instead of date of birth; occupational categories rather than job; and, area rather than village
- generalise meaning of detailed text variable

  e.g. occupational expertise
- restrict upper lower ranges of a variable - hide outliers

  e.g. income and age
- combining variables

  e.g. creating non-disclosive rural / urban variable from place variables

Anonymising qualitative data

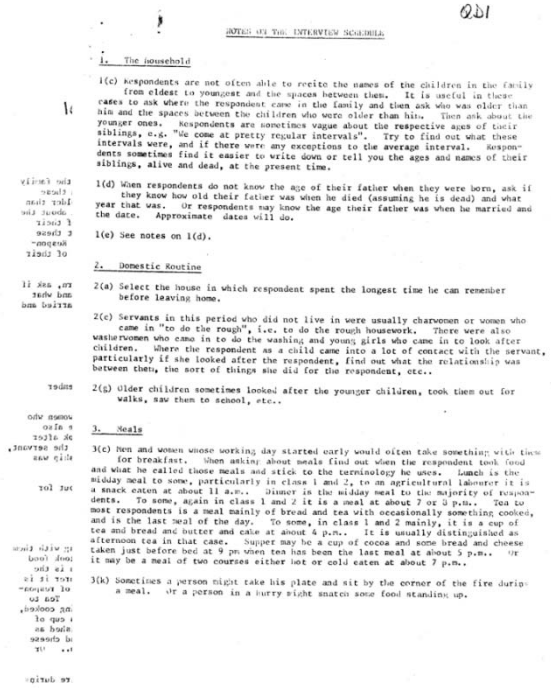
- plan or apply editing at time of transcription
- avoid blanking out; use pseudonyms or replacements
- identify replacements, e.g. with [brackets]
- avoid over-anonymising – removing / aggregating information in text can distort data or make it misleading
- consider keeping an anonymisation log of all replacements, aggregations or removals made and keep it *separate* from anonymised data files

What's useful data documentation?

- Data collection methodology and processes: sampling methods, sampling size, fieldwork protocol and interviewer instructions
- Information sheet / consent form
- Fieldwork tools: questionnaire, showcards and interview schedule
- Data list: overview of key information about each interview, as 'at-a-glance' summary of the data collection
- Analysis tools: codebook, memos, variable listing
- Annotated code for syntax for derived variables

In practice: user guide and documentation

- A user guide could include all kinds of context: transcription notes, photos



Q.D1

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 eldest 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

Embedded metadata in an SPSS file

hse09ai.sav [DataSet2] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	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

Data View | Variable View

PASW Statistics Processor is ready

UK Data Service

Documenting metadata on interviews

Information about interviewee

Date of birth : 1902
Gender : M
Marital status : Married
Occupation : Postman
Geographic region : Colchester, Essex

I : I'd like to start, if I may, by asking you your birth date.

K : November 9th, 1902.

I : Could you tell me how many children there were in your family?

K : There were 11 of us. I was the eldest.

I : Could you tell me, if you remember, how they went after that and roughly the space between them and whether they were boys or girls.

K : Well, the first 3 of us were boys, then I had a sister, another brother, three more sisters and twin brothers at the end.

I : So you were approximately 7 boys, is that right, and 4 girls?

K : That's right, yes.

I : And do you know approximately how old your parents were when you were born?

K : Oh, maybe 21, 22.

I : And when the last child was born?

K : Oh, I suppose they were 45.

I : Did they lose any?

UK Data Service

Documenting data: the data list

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	

Transcription template

- 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.

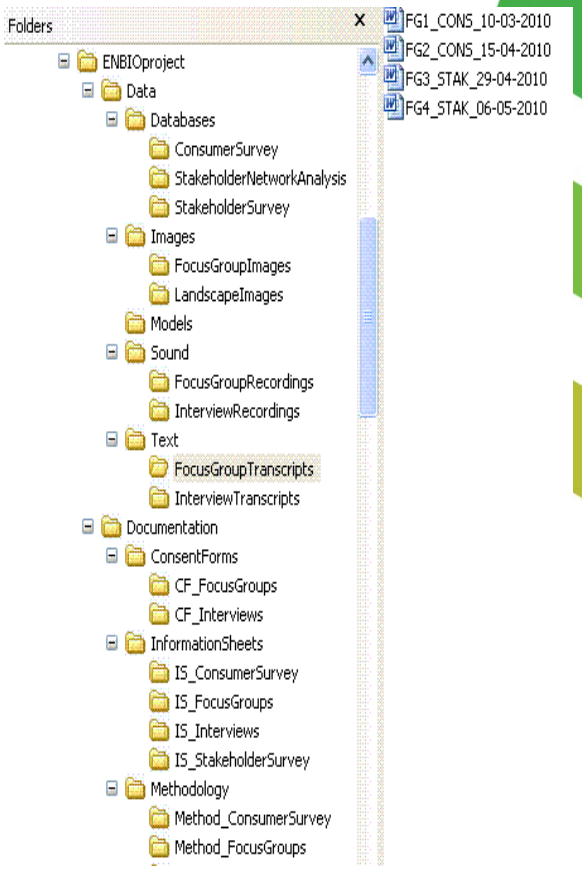
- Cover page or header
- Compatibility with import features of Computer Assisted Qualitative Data Analysis Software (CAQDAS)

Organising data

- Plan in advance how best to organise data
- Use a logical structure
- Use logical names and version control e.g. V1.0, V2.1, ‘FINAL’
- 2018-01-30_Interview_01

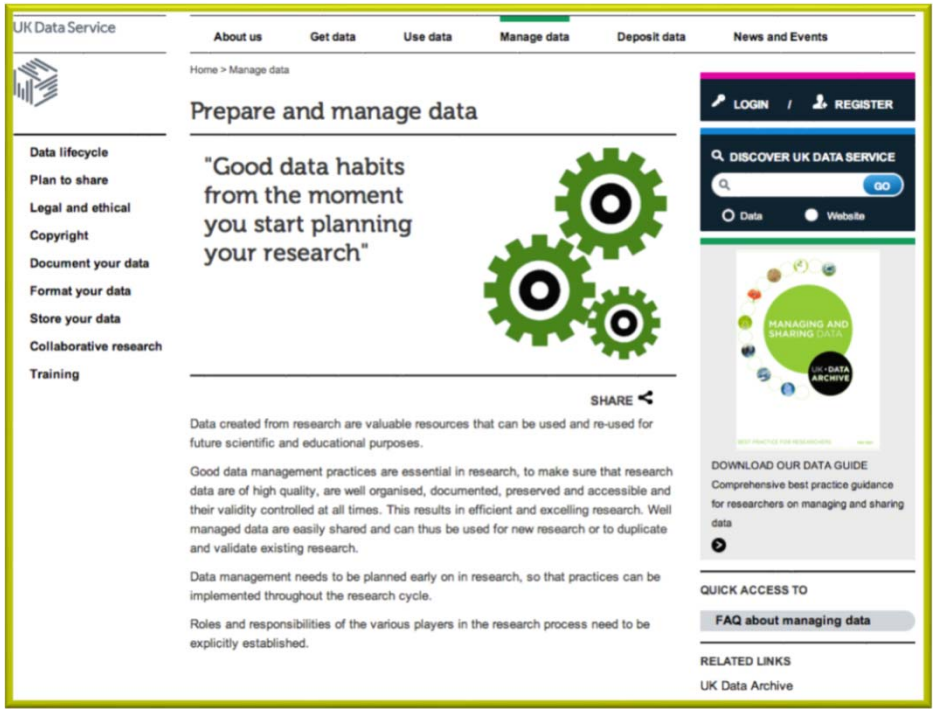
Examples:

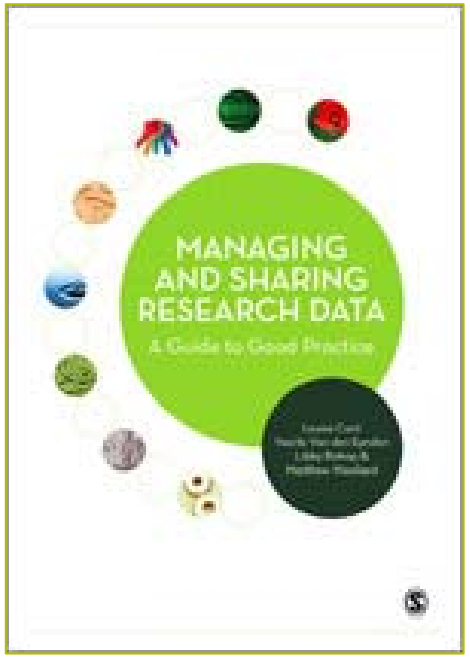
- Group files in folders, e.g. audio, transcripts and annotated transcripts
- Survey data: spreadsheet, SPSS, relational database
- Interview transcripts: individual well named files



Our data management guidance

- 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.ukdataservice.ac.uk/news-and-events/events





UK Data Service

Tools and templates

ukdataservice.ac.uk/manage-data.aspx

- ✓ Model consent form
- ✓ Survey consent statement
- ✓ Transcription template
- ✓ Transcription instructions
- ✓ Transcription confidentiality agreement
- ✓ Data list template
- ✓ Research data management costing tool
- ✓ Encryption tutorials: https://www.youtube.com/watch?v=y4Iosu-Yfsw&list=PLG87Imnep1SmnFGhAjFVHonQSVmMlpHkV

Keep connected

corti@essex.ac.uk
UK Data Service
University of Essex, Colchester, UK
UKdataservice.ac.uk

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

Follow UK Data Service on Twitter:
@LouiseCorti
@UKDataService
Youtube: www.youtube.com/user/UKDATASERVICE