Data Curation Profile for Interdisciplinary Social Sciences
MANTRA Research Data Management Training, University of Edinburgh
http://datalib.edina.ac.uk/mantra/libtraining.html#profiles

Interviewee: Nick Jenkins
Chancellor's Fellow.
Interdisciplinary Social Sciences in Health, School of Health in Social Science.
http://www.ed.ac.uk/schools-departments/health/

Interviewer: Rowena Stewart, Academic Support Librarian, Information Services.
Interview: Monday 16th December 2013, 2pm Main Library.

For the purposes of the interview, one data-driven research project was used as the primary basis of discussion

1. OVERVIEW

1.1 General overview of the project and its aims?

Qualitative interview study into patients' experiences of genetic testing for familial high cholesterol.

Aim: understand extent receiving genetic test result for patient high cholesterol on health behaviour and/or risk perception.

Data completed May 2011

Publication: http://dx.doi.org/10.1016/j.pec.2011.09.002

[pre-publication paper linked to from http://edinburgh.academia.edu/NicholasJenkins]

1.2 What are the funding sources for your research?

Chief Scientist Office (Scotland) - http://www.cso.scot.nhs.uk/

1.3 Was there a data management plan for the project? If yes, what was the motivation?

Open access data was not as prevalent/visible when the project began as it is now and common protocols were not around. Therefore, data protocol followed the conditions of the funding application and so considered:

- Where electronic copies were to be held.
- Who was to have access.
- What was to be done post-project.
- Data had to be anonymised as part of analysis.
- Specific uses would be subject to informed consent.

The Principal Investigator (PI) for the project would have had to make the decision on whether to make available anonymized data but this was not considered.

[*cf* an ESRC funded project now would have to make available its unanalysed but anonymised data. Coding would not have to be included.]

1.4 Please provide an overview of the data related to the project (include all data and ephemera which had to be managed, not just the raw numbers).

i) Audio files (38 interviews).

Research assistant does the interviews and the voice files stay with the researcher who records them. They can't be shared and are stored as sensitive data according to data protection legislation, ultimately on a secure server. The secure server could be, eg UoEdinburgh personal network space.

When researcher moves on files should be transferred to the PI, who is the overall manager, thereby maintaining one trail. PI ensures the files are destroyed with 5 years of their completion.

It is assumed the original file space, ie research assistant's UoEdinburgh personal network space, is wiped when the research assistant leaves the University's employment. There is nothing in a research assistant's contract to have the files removed.

ii) Anonymized transcripts.

Would expect to share these if request made by other academic researchers with non-commercial interests. They may be printed to make coding easier but then shredded and are, anyway, anonymised.

iii) Keystone database linking anonymised transcripts to interviewees.
Password protected. Held by the researcher only and PI wouldn't necessarily need to link the data.

iv) Anonymised transcripts are uploaded to NVIVO and thematically coded.
The coding is part of the data analysis and considered intellectual property.

N.B. Interviews are not structured questionnaires and so, there is no "interview script" associated with such a project as this.

1.5 Who is the intellectual property owner of this data?

Legally, de facto, IP rests with the PI who designed the project (not Chief Scientist Office/funder) but owner in practice of data produced by its employees is ultimately the employer if the employer is the UoEdinburgh. If the PI moves, negotiations take place over transfer of data and staff.

1.6 Approximately how many data files were generated during the course of the project?

Output transcripts equal the number of participants (38 in this project).

In general, for qualitative interviews "theoretical saturation" is the point when further interviews reap the same responses as already recorded and is the "rule of thumb" used by researchers which means sample sizes vary.

However: 20 is a small sample size. 70 is a large sample size.

1.7 What is the average size of the data files you current have?

The transcripts are Microsoft Word files of approximately 10k for interviews of 1 - 1.5 hours in length.

[Size of the audio files varies from perhaps 10Mb upwards and depends on the length of the interviews and the device used to make the recording.]

2. ORGANISATION

2.1 What format(s) are the data in? (MS Word, MS Excel, MySQL database, etc.)

See 1.7 above.

From other projects there have been other types of data output:

- pictures (although these are data protected) and
- video files.

The video files, actors in vignettes, are made publicly available with locations/links reported in related academic papers. The video files are hosted on the theatre company's YouTube area and deposited in Alzheimers Scotland.

IP for these videos is jointly owned by the theatre company and University of Edinburgh. However, although badged for UoEdinburgh they are not available from the University website or saved in a university space.

2.2 Please describe briefly the way your data is currently organized: for example, file name conventions, any existing metadata or units.

The transcripts are stored in a folder on the researcher's personal University network drive. In some cases, it could be a shared drive.

The transcripts are imported into NVIVO (which can be sole use or shared). For this project, NVIVO was the group's shared format.

Projects are organised into specific folders which are browsed, not searched, and not overly burdensome to scan by eye. The contents do not cross over between projects, ie there is no saved output common to more than one project.

2.3 Is your current metadata system important for you to keep or would you be willing to adjust to a more universally compatible metadata system?

Metadata is not applied and so no preference.

However, in a repository situation:

Interest lies in finding individual transcripts, or chunks of transcripts, of particular interest from other research and the provenance of the individual transcripts would be necessary. Therefore, application of metadata to each transcript would be anticipated, if part of a "project bucket of transcripts".

Description at this level of data is not so pressing for anonymised stats, eg BMIs.

2.4 What specific software programs or tools were used in the collection and organization of this data?

NVIVO

2.5 What specific software programs or tools are required to utilize this data (proprietary file formats, GIS, etc.)

Word processer software/ file reader.

2.6 Where are your files currently stored? How have you backed up your data?

Day to day storage: The personal (institutional) file space of relevant persons/researcher and transferred to PI on completion.

Back up: running multiple back-ups creates the risk of CDs/USBs/etc getting lost and data protection legislation being broken and is therefore not usual.

However, the transfer of the files from the researcher to the PI is done via secure data transfer e.g. a memory stick taken to the PI by hand and then erased following data transfer.

The researcher has, or has access to, the audio files, transcripts and personal database key. The researcher could negotiate access to these files when they leave.

2.7 What measures are currently being used to control access to your data?

Access control is as the control to employees' personal file spaces on the University of Edinburgh's network.

An additional layer of control is added to particular files. For example, password protection is needed to access the keystone database which holds the information to link anonymised transcripts with their interviewees.

[The audio files are retained by the researcher who recorded them. Nick has never heard of audio files being stored on personal devices. One shared username/password is used to access the audio files.]

2.8 What measures do you require to control future access to your data?

Informed consent of each research participant would have to have been given before anonymised transcripts could be deposited in an open access, searchable database. At the point consent was given by the participants, it was for specifics, *eg* "help drive service delivery" and no specific consent to deposit was given. For this project, permission would have to be sought retrospectively.

Seeking retrospective consent is not usually an ethically robust approach e.g. participants' recollection of the data they have provided may be limited at the point of seeking retrospective consent. In other cases (e.g. qualitative dementia research) participants' capacity to provide informed consent may fluctuate/diminish

However, if participant permission has been given:

- **Would want a contract between user and repository provider to stipulate ethical usage policy**. The contract would vary project to project and would stipulate that the signee would not knowingly use the data to misrepresent the interviewee's story or do "bad work" or use to further commercial interests.

- Acknowledgement or citation agreement – following a stated format laid down by the repository. Also, offer of co-authorship on research papers to those researchers who generated the primary data (see BSA guidelines for authorship for further details)

3. STORAGE & SHARING

3.1 How long would you like your data to be preserved? (if different types of data should be preserved for different time periods, please specify).

Anonymised transcripts have no cut off date.

The interviewer has a duty of care to disclose criminal behaviour/abuse or act if they learn someone else is at risk. Information on the transcripts relevant to proceedings against criminal behaviour would be sub-judice.

[Audio files of the interviews are not made available but, for completeness, UK data protection legislation means they are kept for 5 years or less.]

3.2 Do you intend to publish the results of your research in an academic journal? Do you intend for your data to be linked to this publication?

Yes, project results went for publication in academic journals. The journals to which papers were submitted were chosen on the basis of impact factor. They had no formal policy on data availability and did not require underlying data to be made available.

[Funding proposals include probable costs, if known, which would be charged by publishers to make articles open access. Open access data is not standard therefore costs not applied for in funding applications.]

3.3 Who is the intended audience of this data? (there may be more than one audience expected for different types of data)

- Authorised researchers in the field.
- Journalists.
- Authors for eg character research.

3.4 Looking at the 'data sharing matrix' please tell me about the reasons behind your sharing choices.

Participants are sharing personal insight and the reason they consent to share is overwhelmingly because they trust the person interviewing them. **There is potential for exploitation as the participants are often vulnerable.** They open up due to their rapport with the interviewing researcher. The researcher has a duty of care towards their participants, which includes protection from likely sources of distress and mis-appropriation.

The concern underlying the contract measure given in 2.8 above is of participants' data not being treated with respect, for example, a quote being used out of context in the public domain for sensationalism and/or for profit.

The data is context specific and use like this breaks the level of respect between the participant and the researcher.

3.5 Please describe any conditions or constraints placed on the sharing of this data (mandatory dissemination agreements, confidentiality clauses, etc.)

- UK data protection legislation.

- The interviewer has a duty of care to disclose criminal behaviour/abuse or act if they learn someone else is at risk. Information on the transcripts relevant to proceedings against criminal behaviour would be sub-judice. [Repeated from 3.1].

- There is a desire for research output to be used outwith immediate research communities, eg in policy, re-use by creatives (novelists, drama, arts etc) but interviewees participate on a basis of trust (see 3.4 above) so there is a need for repository users to agree to ethical/appropriate re-use.

A particular concern would be stop insurance companies trawling for information to identify "loopholes" in their policies. That they do this was considered "beyond the pale". Nevertheless, interviewees will have disclosed how they stopped/got around declaring problems, or had stopped getting tests so they didn't have to declare them and have their insurance policies adversely affected. Participants have a right to have this information protected.

3.6 If you were to share your data, would you want to be able to obtain usage statistics for your data? What measurements would be most important to you? (for example, times viewed, downloads, or citations)

- Number of downloads.
- From where the download was initiated.
- The broad purpose for download: research, journalism, personal.
- Contacts form: user contact details.
- Times cited therefore have to be able to track back to deposit.

3.7 If you were to share your data, would you like an embargo time period on access (if different types of data require different embargoes please specify)?

Post-publication.

3.8 What uses would you anticipate your data could be put to in the future?

For qualitative studies, it is time-consuming to work backwards to get familiar with someone else's data. Strength comes from replicating the study in different places/with different populations. However, as stated in 2.3 above there is interest in finding individual transcripts, or chunks of transcript, of particular interest from other research.

Although the primary consumers of this project's data would be other researchers in the same field of research, The School of Health in Social Science at University of Edinburgh wants research output to stimulate real world discussion and to be used to make differences to other people's lives in line with survey participants reasons/agreement for being surveyed in the first place.

3.9 Thinking about provision at the University of Edinburgh, are there additional support services you would like to see?

Having a readily accessible data repository system!

To include:

- Standard set of support material on how to get consent of participants to deposit.

Qualitative research is not standardised and interpretation is idiosyncratic, so there is the fear that extra scrutiny would result in others finding problems with data. However, **if qualitative researchers are not depositing their data** it is not because they are afraid of being pulled up on statistical methods etc **but because of concerns over ethics**, ie the right to deposit in the first place and the risk of mis-use of the data post-deposition.

- Central support staff to operate as Scholarly Communications Team currently do for open access publications support, ie to oversee out/input; advise on not infringing copyright or data protection; ensure compliance with copyright and data protection law.

This extra staff resource would be needed to remove the time project staff would have to spend. Such time, if spent by project staff, would be seen as prohibitive and the result would be deposit would not take place.

- Some input of metadata from repository end as, for qualitative data (cf quantitative) this could be time-consuming (see 2.3 above).

- Assistance on tracking citation/usage.

Repositories work well for statistics but not narrative research output and depositors of **"big data" gain more from open access data than depositors of narrative data**. REF is good in its metrics for use by the non-academic world. Therefore, would want the same for data, especially for qualitative data. Would data re-use count as a separate citation? Is citation the most effective metric for data?

In addition, authorship is attributed to anyone who has made a significant contribution to the publication, not just in the writing of the paper, and includes those involved in data collection only, *eg* interviewers who do not write or review still get author credit on published work. If open repository data is used in a published paper, how do such contributors get the author credit they would have under original circumstances?

This may be more fear than reality but is driven in part by personal knowledge of anthropologists who act accordingly: having lots of field notes from which the quantity deposited are sparse, the high quality data kept by the person who originally collected it.

Data-sharing matrix: data types and levels of sharing anticipated for project forming the primary basis of this discussion

List each type of data here (planning documents, raw data, analysis, etc)	Wouldn't share with anyone	Would share only with my collaborators	Would share with others in my field	Would share with other academics outside my field	Would share with the general public
Voice files		☑			
Transcripts					☑ Caveats: see comments re repository contract with users and citation credit

4. OF INTEREST

Nick is the School of Health in Social Science's open access champion and feels he has thought more about the issues and therefore feels differently about them to some of his colleagues.

Nick is also an Academic Editor for PLOS ONE for which there is a publication criterion asking whether authors are making available their data to other researchers and, if so, how.

PLOS ONE's preferred option is for the anonymised data to be held in a secure repository, not self-stored, and for the repository to be vetted according to data protection and managed by an institution.

Authors who say their data is not available for other researchers, usually do so because of consent issues.

Nick feels his involvement in PLOS has changed his attitude to the ethics of barring access. It has brought into focus that research is public funds spent on behalf of the public good. Advances tend to come from sharing data and this is a good thing as long as authorship is acknowledged. Often, patients themselves take part in research because they want good to come from their contribution.

[PLOS ONE also grades each submission on how likely it is to attract media attention, trying to schedule for those articles considered likely to attract general news industry attention, release times which coincide with greater staff availability.]

5. INTERVIEWEE'S CLOSING STATEMENT

The move towards open access qualitative data is likely to be gradual (discussions around this first began over a decade ago!) and will be met with a certain amount of resistance within the academic community. Key to the effective development of OA to qualitative data is the establishment of effective protocols for obtaining informed consent **at the point of data collection** as well as robust protocols for data management/use which are specific to the nature of qualitative research.