Curating qualitative data

2nd SEEDS workshop
Between 9th and 11th February 2016
Ljubljana, Slovenia

Arja Kuula-Luumi, FSD

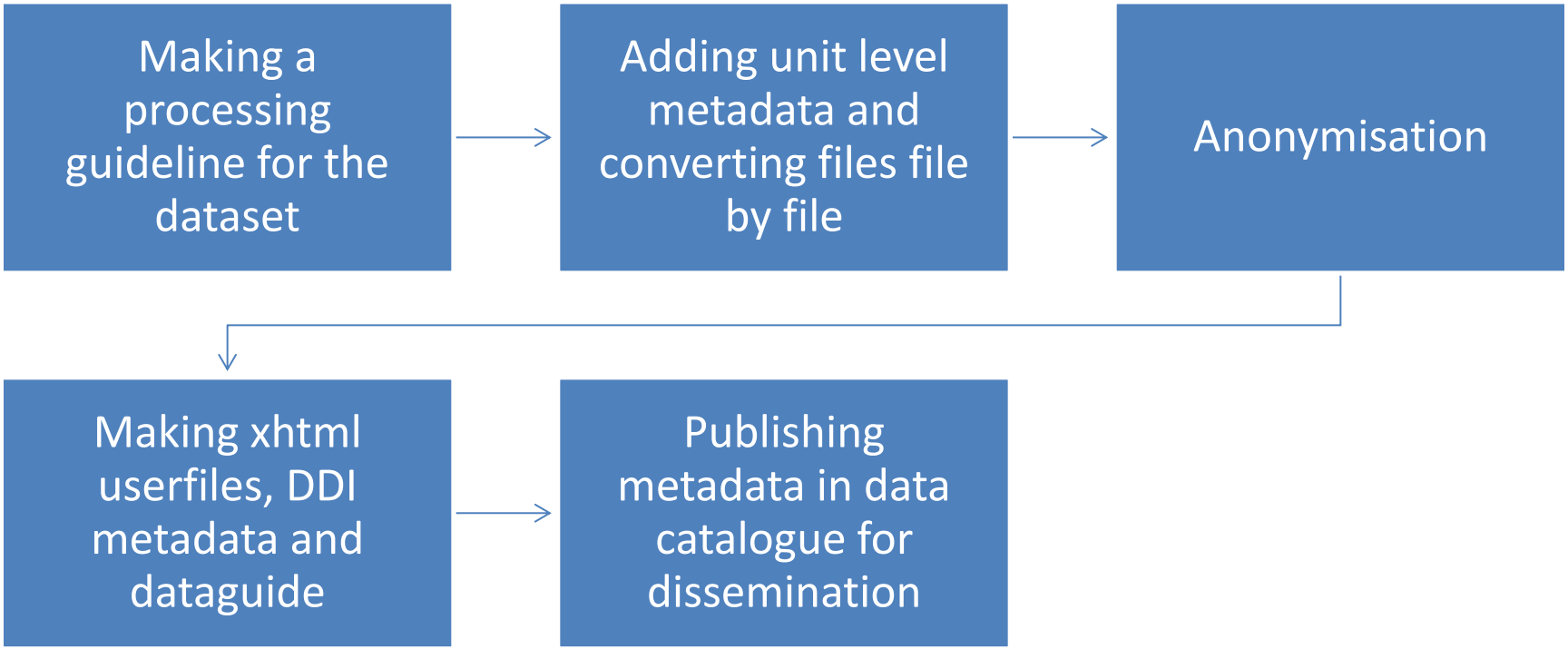
Contents

- The archiving process of qualitative data in FSD
- Data unit level metadata
- Anonymisation

The process of archiving qualitative data

- Check the amount of data files
- Check the contextual material is included, ask researcher if needed
  - Interview questions, observation guidelines, informed consent, information sheet etc.
- Check the data files open correctly
- Check there are speaker tags indicating the question/answer sequence (transcriptions)
- Check there is no entered content or structural information using formatting (i.e. using bold, italics, underline, colours, indent etc.)
- Check there is a document describing datafiles (naming procedures) or that each file contains basic data unit information

Process in FSD



A processing guideline

- Qualitative ”syntax”
- Based on investigation on all the deposited material
- Can be done only by studying/investigating 2-3 data units
- Includes
  - A suggestion of the name of a dataset
  - A suggestion of a data unit metadata
  - A suggestion of an anonymisation procedure
- Needs to be approved by the researcher
- An anonymisation procedure usually tightens during the processing

Data unit

- Data unit = an interview; a diary; a life story or other written text; an article (if analysed as data); a field note on one day; a picture etc.

- One data file is often also a data unit, but not always
  - If all the data are in one file, FSD usually separates them

Unit level metadata

- Metadata elements of data units vary depending on the collection
- Usual elements are
  - Background information of the participants
    - Gender, age, occupation, education etc.
  - Situational information (time, place, the presence of other persons)
  - Information concerning the content of data
    - For instance: if interview is about recovering from divorce, the date/year of divorce

Example 1: an interview with only one interviewee

- Interview date: 08.02.2013 [=8 February 2013]
  Interviewer: Matt Miller
  Pseudonym of interviewee: Ian
  Occupation of interviewee: Journalist
  Age of interviewee: 32
  Gender of interviewee: Male

- I: First I would like to ask you about your choice of profession. How did it come about that you decided to become a teacher?
  Ian: Well, you know, when I was a kid we had this really great guy teaching history....

Example 2: a focus group interview, with several interviewees

- Interview date: 08.02.2013 [=8 February 2013]
  Interviewer: Matt Miller
  Pseudonyms of interviewees: Ian (R1), Mary (R2), Ken (R3)
  Occupation of interviewees: Teacher (R1), Headmaster (R2), Janitor (R3)
  Age of interviewees: 31 (R1), 47 (R2), 22 (R3)
  Gender of interviewees: Male (R1), Female (R2), Male (R3)

- I: First I would like to ask you about your choice of profession. Tell me a bit about how you came to have the profession you have now?
  R3: It's not really... er, for me, it's not a profession, I'm just doing this for now and might go back to school later.
  R1: You know, when I was a kid we had this really great guy teaching history....

Example 3

- **Project: “Industrial Citizenship and Migration from Western Balkans: Case studies from Albania and Kosovo migration towards Greece, Germany and Switzerland”.**
-
- **GREECE – RETURN/CIRCULAR**
-
- ***K.***
- ***Place : Greece***
- ***Age : 33***
- ***Status : Return 97- 2000 = 4 years***
- ***Family : Married***
- ***Date 14.12.2014***
- ***Time 14:20- 16:30 = 2 h 20 min***
- ***Place :At my home***
- ***Work experience in Greece – Tailor/ Domestic worker***
- **Reason for returning : husband lost his job / she was pregnant with her second child and could not work in a dangerous workplace (cs the first child was in Albania with her parents)**

What is personal (identifiable) data? (1/2)

- Personal data are any kind of data that may be used to identify a natural person. Identification can be made on the basis of factors specific to the physical, psychological, mental, economic, cultural or social identity of an individual or individuals.

- Information that is **sufficient on its own to identify** an individual includes a person's full name, social security number, email address containing the person name, and biometric identifiers such as fingerprints, facial image, voice patterns, hand geometry, iris scan, or manual signature. This type of data are often called **direct identifiers**.

What is personal (identifiable) data? (2/2)

- FSD also counts as **strong indirect identifiers** the types of codes that can be used to unequivocally identify an individual from among a group of individuals. These include, for instance, student ID number, insurance or bank account number, or IP address of a computer etc.

  - As a rule, the FSD removes both direct and strong indirect identifiers from the archived data.

- Other indirect identifiers are the kind of information which, when linked with other available information sources, could identify someone. For example, age, gender, municipality of residence, or a rare job title may in some cases, when combined with other information, enable identification. Background variables form the most common case of indirect identifiers.

When data is anonymised?

- Data are anonymised if characteristic factors (for instance, indirect identifiers when combined) are the same for several individuals and if any particular individual cannot be identified with reasonable effort.

- The estimate of how identifiable the data of a dataset are and how they can be anonymised are always done on case-by-case basis.

- The level of anonymisation is influenced by the circumstances and purposes of data processing. Contrary to administrative registers, research data may not and must never be used to estimate participating individuals or to make decisions concerning them.

Anonymisation – starting point

- Review the data as a whole

1. Information given to participants, consents
   - If participants are interviewed as experts or in an administrative role and they do not talk about private issues, the data may not need anonymization (consent for archiving needed!)
   - What about Mapping the leaders of Macedonia and Albania?
2. Exactitude of background information of research participants
3. Subject matter of the data

Unit level metadata edited into categories

- Original:
  - Arja Kuula-Luumi: 52-year-old development manager working in the Finnish Social Science data archive, University of Tampere, married, living in Tampere, son aged 16 living at home and daughter 21 living on her own

- Categorized
  - Gender: Female
  - Age: 50-55
  - Occupation: Professional in the field of research services
  - Place of occupation: University (or public sector employer)
  - Household composition: Husband, teenage son
  - Marital status: Married
  - Place of residence: Town in the province of Western Finland

Anonymisation: proper names pseudonymisation

- It is always better to **use pseudonyms** than simply delete the names altogether or replace them by a mere letter or a character string, such as [x] or [---].
- Replacing proper names with pseudonyms enables the researcher to retain the internal coherence of the data.
- In cases where several individuals are frequently referred to, much of the information is lost if the proper names are just removed.
  - Make an excel of pseudonyms used in data files
- **Attention:** A dataset may contain references to persons who are publicly known on account of their activities in politics, business life or other work-related spheres. Their names are not changed for pseudonyms.

Anonymisation: proper names categorized

- No pseudonyms need to be created for persons who are mentioned only once or twice in the data, and who have no essential importance for the understanding of the content.

  - Their names can be replaced by a category (e.g. [woman], [man], [sister], [father], [colleague, female], [neighbor, male])

- Other than personal proper names can be also replaced by a category, for instance, [lower secondary school], [home town] or [residential area], [law firm], [football team], [local café].

Get help from official statistics and their classifications

- The Statistics Finland maintain classification recommendations that are used in FSD when anonymising qualitative data
  - Regional classifications (e.g. division into regions and major regions)
  - Social classifications (e.g. occupation and education)
  - Economic classifications (e.g. industry and sector)
  - Other classifications (e.g. land use and buildings)

Anonymisation – common pitfalls

- Identifiable information of third persons are not anonymised
- Exact locations (eg. place of residence) are anonymised in the text, but unique proper names of restaurants etc. are left in the data
- Using ’replace all’ without checking whether the same persons are referred to by using different names (Thomas=Tom=Tommie)

Anonymisation - ethics

- If parts of the data seem to be too sensitive, they can be removed *[remark of a data archivist: section has been deleted due to its sensitivity]*
  - This pertains mostly to information concerning third persons

- If there are issues that the participant tells to be highly private and cannot be published, they ought to be removed

- If in the end of an interview there are researcher's notes of conversations that have happened after the interview, they ought to be removed

Links

- FSD’s anonymisation guidelines for researchers
- http://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html

- FSD’s guidelines for processing qualitative data files
- http://www.fsd.uta.fi/aineistonhallinta/en/processing-qualitative-data-files.html

Thank you!

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE