EXPERT SEMINAR

OPEN ACCESS TO DATA:
ANONYMISATION, DATA PROTECTION & CONFIDENTIALITY

OCTOBER 12- 13, 2006

Place: Hotel Golden Age, Room: Sapfo,
57 Michalakopoulou st., ATHENS, GREECE

Data Protection Act and Anonymisation of Research Data

Arja Kuula, Research Officer
Finnish Social Science Data Archive

Privacy is a crucial issue both in legislation and in research ethics. Protecting the privacy of research subjects and ensuring data confidentiality belong to the core principles of research ethics. Both aspects are covered in my presentation.

My focus is on research data collected from research subjects one way or another. My presentation is based on Finnish practices and does not cover register data. FSD disseminates archived data for scientific research and teaching only, not for any other purpose. This means that in our case open access applies only to the scientific community.

I begin by explaining briefly how we take the Data Protection Act into account in the archiving process. My next topic is anonymisation of both quantitative and qualitative data. I finish by discussing the possibility to archive data without editing identifiers. This discussion is based on a brief analysis of the concept of privacy.

Data Protection Act and Archiving

According to the Finnish implementation of the Data Protection Act, the starting point when assessing the possibility to archive a dataset is to check whether the prerequisites for processing personal data are fulfilled. The primary prerequisite is that the data subject has unambiguously consented to the processing. Therefore consent – written or verbal – is of primary importance.

At the time of data collection, research participants are usually given information about the purpose and content of the research. As regards the future use of data, participants are generally merely informed that confidentiality will be safeguarded and personal identifiers will not be published. As a rule, archiving is not mentioned at all. This is the case for most datasets deposited at the FSD, which leaves us with two alternatives: 1) the researcher gives us a mandate to ask each participant separately whether s(he) agrees to archiving, or 2) we have to apply Section 14 of the Data Protection Act. The first alternative is very labour-intensive and time-consuming but we have done so for four qualitative datasets. In most cases we resort to the second alternative.

Section 14 states that personal data may be processed for historical or scientific research also without consent if the research cannot be carried out without data identifying the person, and if the consent of data subjects cannot be obtained because of the age or quantity of the data, or another comparable reason. In this case, use of personal data files must be based on an appropriate research plan, and a person or a group of persons must be nominated as responsible for the research project. In addition, the data pertaining to a given individual should not be disclosed to any outsiders. Usually the option provided by Section 14 is used only for large register-based research projects. It is not a good option for an archive since the application of these clauses must be done on a case by case basis.

The last paragraph of the section 14 states that personal data may be processed for historical or scientific research even after the research project has ended if personal data files are destroyed or transferred to an archive, or the data are altered so that data

subjects can no longer be identified. In practice, transferring the original data to the archive would mean a long bureaucratic route, with many official procedurals including evaluation by the Data Protection Board. We have not tried that route yet.

When a dataset containing personal data is deposited at the archive, we usually opt for anonymisation. I now present our anonymisation practices for quantitative and qualitative data. At the moment, we have about 50 qualitative datasets which is rather a lot considering our limited resources, and also considering that the Finnish research culture sees qualitative data as extremely personal and best kept as much a secret as possible.

Anonymising quantitative data

We start by reviewing the dataset as a whole, concentrating on four key elements: information given to participants, background variables, variables based on open-ended responses, and subject matter of the data. All these elements are taken account, also in relation to each other, and it is only after reviewing all four that we decide which variables to remove, and which to alter etc. As regards background variables, their number and degree of specificity are of particular importance.

Anonymisation methods for quantitative data include:

- Removal – eliminating the variable from dataset entirely
- Bracketing – combining the categories of a variable
- Removing identifiers from open-ended questions
- Top-coding – grouping the upper range of a variable to eliminate outliers
- Using samples instead of total original study
- Swapping
- Disturbing

Removal –eliminating the variable entirely from the dataset. This is the most radical way to anonymise data, and we use it for direct identifiers. Sometimes, after careful consideration, we also remove indirect identifiers, especially if there are many of them.

For example, we may remove the school name variable from a survey on youth crime if another variable gives the school level. Otherwise, a researcher familiar with the school or the area might be able to recognize a respondent from the information the respondent has given about his/her criminal activities. Removing the variable identifying school name considerably decreases the threat of disclosure, without diminishing the scientific value of the data.

Bracketing – combining the categories of a variable – always a good option. I personally consider bracketing to be a better alternative than removing a variable. Bracketing is typically used for variables including indirect identifiers. For example, instead of using the school name we may recode the variable into categories like lower secondary school, upper secondary school, vocational school etc. In case of variables like age, municipality of residence and occupation, recoding values into categories also diminishes the risk of indirect identification. We frequently divide the variable identifying the respondent's municipality of residence into two regional variables (i.e. province and statistical grouping of municipalities). We have special syntaxes for this. Thus, the risk of identification is reduced without loss of essential information.

Removing identifiers from open-ended questions. Identifiers may refer to respondents themselves or to other persons. Removing direct identifiers (names, telephone numbers, e-mail addresses etc.) from the data does not lead to any loss of essential information. Otherwise the risk of identification is assessed for each dataset separately, taking the subject matter and background variables into consideration.

Top-coding – grouping the upper range of a variable to eliminate outliers. Income variable is a typical example. Highest incomes are collapsed into a single code but other income responses are kept as actual quantities (i.e. the actual income in euros). This prevents identification of highly paid individuals.

Using samples instead of total original study. This method is used, for example, by Statistics Finland. There are also other methods which FSD does not currently use. ICPSR's Guide to Social Science Data Preparation and Archiving (2005, 22) presents:
Swapping — Matching unique cases on the indirect identifier, then exchanging the values of key variables between the cases.

Disturbing — Adding random variation or stochastic error to the variable. This retains the statistical properties between the variable and its covariates, while preventing someone from using the variable as a means for linking records.

When making decisions about the level of anonymisation, we take into consideration the subject matter and the degree of sensitivity of the dataset. Anonymisation must be planned carefully if the survey carries many questions pertaining to the respondents themselves, and contains sensitive information in the sense defined in the Data Protection Act. Surveys covering respondents' state of health or social security benefits tend to be much more sensitive than surveys charting opinions on, for example, neighbourhood services. Surveys focusing on attitudes and opinions are generally less problematic, and require a lower level of anonymisation.

Anonymising qualitative material

When we start to anonymise a qualitative dataset, we do it the same way we do with quantitative data. First, we review the material as a whole. The main elements taken into consideration are what kind of information has been given to research participants, how easily participants could be identified from the background information (i.e. how detailed it is), and how sensitive the subject matter is.

Anonymisation methods for qualitative data include:

- Removing direct identifiers
- Altering names and other proper names
- Removing or editing sensitive information
- Editing background information into categories

If research participants have not consented to having their names and other personal information left as such in the data, their names, addresses, dates of birth, e-mail addresses, telephone numbers etc. are removed as soon as the material has been checked technically. This will prevent re-users of data from contacting a particular participant later for more information - even if they would like to.

In most cases we use pseudonyms for proper names. This is always a better solution than removing the name altogether, or replacing the name with a letter (e.g. X) or a short character string. Using pseudonyms maintains the internal coherence of the material. If several different persons are repeatedly mentioned in the data, a lot of information is lost if the names of these persons are just replaced with the marking [male] or [female]. However, this marking can be used if a person is mentioned only once and is not relevant to the data. Using a pseudonym for both the first name and the surname may be justified to make the transcription resemble natural speech or to keep a large number of participants separate from one another. Usually, however, we replace first names with pseudonyms and remove surnames.

It is not necessary to invent pseudonyms for all proper names. If the data unit (an interview transcript, a lifestory, a letter etc.) talks/writes about only one school, workplace or place of residence, we usually replace the name with a more general term like [lower secondary school], [accounting firm], or [home town]. Information technology allows for quick anonymisation processes but we tend to use search and replace techniques with great care, only replacing one item at a time, and not using the 'Replace all' command at all.

Names of persons are not replaced with pseudonyms when the person in question is well-known and the participant is not talking about this person's personal affairs. For example, a participant working for Nokia may mention the name Jorma Ollila (long-time Chairman and CEO of Nokia). If the participant says things like "If only Ollila came and led this project from the beginning to the end, he would…" using a pseudonym for Ollila would totally distort the meaning. Another special case is when the name of a person is used to mean something else than a person with that name, as sometimes happens in subcultures. Hence people doing the anonymisation must know the data.

When there is a risk of even partial identification, and the personal/sensitive data are not necessary for the understanding of the content, we either delete or edit it. Even in this case it is better to edit the data than to delete it. Diagnosed severe illness can be changed into another, similar type of illness if doing this does not distort the data too much. Another method would be to change pancreatic cancer, for example, to [incurable

illness] and thereafter refer to as [illness] if the reader can deduce from the context that [illness] refers to the incurable illness mentioned in the beginning. This method is to be recommended when the sensitive information is not very relevant to the subject matter and the respondent mentioned it only incidentally. But if the study focuses on the lives of persons with a severe illness, the threat of disclosure is best reduced by using other anonymisation methods than editing information crucial to the subject matter.

Background characteristics of participants like gender, age, occupation, workplace, school or place of residence are often essential for understanding the data, and constitute important contextual information for secondary analysis. To avoid identification, detailed background information can be edited into categories in the same way as for quantitative data.

Categorisation is always a better solution than deleting background data. If I were interviewed for qualitative research my background information would be: 42-year-old research officer working in a separate unit of the University of Tampere, married, with children aged 7 and 12, and living in Tampere. To avoid identification, my background information could then be categorised in the following manner:
Gender: Female
Age: 41-45
Occupation: Professional in the field of research
Place of occupation: University (or public sector employer)
Household composition: Husband and two school-age children
Place of residence: Town in the province of Western Finland

In the example above, if my place of employment, i.e. a university, were mentioned in the data, it would not need not be generalised into [public sector employer] since other remaining background data would not allow even a partial identification. The province of Western Finland has three universities, and there are also separate university units in the province mentioned.

Usually only part of the background information needs to be categorised, sometimes the only information that needs to be categorised is the place of residence. We decide the degree of categorisation taking into account other anonymisation techniques, plus the

content and subject matter of the data. We aim at a reasonable level of anonymisation. There is no need to anonymise less sensitive data thoroughly because deleting the names and addresses of participants may be enough. On the other hand, in case of sensitive data, the risk of identification can be reduced significantly by categorising background data and using pseudonyms or other editing methods for proper names.

Is Identification Always Harmful?

I myself am not too enthusiastic about anonymisation. Not only because it is time-consuming and needs a separate plan for each dataset but mainly because people presume that anonymisation is always necessary. It need not be. In the research context, we talk about identification without specifying what we mean by it. There is an essential difference between a research publication and research data when it comes to what kind of consequences possible identification might have. When planning anonymisation of research data, (i.e. removal or edition of identifiers), the starting point need not be the level of anonymisation necessary for publishing results. It should be possible for a researcher to study research subjects more profoundly and in more detail, even when he cannot publish the results in such a detail for confidentiality reasons.

I hope that identification can be discussed neutrally, without taking it for granted that identification in itself is an immediate risk and constitutes harm towards research participants. Especially researchers collecting qualitative data seem to presume that research participants would not accept the idea of archiving research data that would be partially identifiable. To check the accuracy of this presumption, we asked a few researchers to let us re-contact their research participants, and did so for four datasets.

It is never possible to locate all research participants afterwards. We were able to find the address for 169 research participants, four of whom did not accept the idea of archiving and 14 did not react to the letter we sent. A data protection official told me that no reaction means silence which in turn implies consent, but we decided not to archive interviews without explicit consent. All four datasets contained unique and personal stories, and some sensitive information about the issues at hand.

When talking with the research participants over the phone I learned that their main reason for giving consent to archiving was a wish to advance science. People had participated in the research because they had thought the subjects of the interviews were worth studying. Giving consent to archiving meant continuing to fulfil this wish. One research participant also said that the original research results had not convinced him, and he warmly welcomed re-analysis by different researchers representing different disciplines. It was really interesting to hear that 25 % of the research participants said they did not want their names removed from the dataset – even thought I explained that removing names is a routine procedure.

It is worthwhile to note that research participants see open access to research data for secondary researchers as self-evident. For them, the research relationship falls to the field of institutional interaction. It means that the interaction is predefined by a research frame where researcher represents the institution of science. They do not necessarily see the relationship as personal and connected to a particular researcher.

I think we as data specialists and researchers especially have to define more exactly what we mean by confidentiality. Instead of secrecy and heavy anonymisation processes, confidentiality should consist of agreements between the researcher and the participants on the future use and preservation of the data. Confidentiality would then entail that when data are collected for research purposes the material could be archived and used for further research unless otherwise agreed with research participants. Confidentiality does not entail total secrecy preventing archiving or enforcing rigorous anonymisation processes. But confidentiality certainly does mean that identifiable personal information gathered for research purposes cannot be delivered or presented as such to the media or, for example, to administrative officials making decisions affecting research participants.

Towards Good Research Practice and Procedures

The basic philosophy behind the Data Protection Act is to protect individuals and social groups from harmful use of their personal information. The law aims to protect people 1) from the power of markets so that their integrity would not be hurt by very focused and

intrusive advertising and 2) from the power of public officials. The law does not strive to hinder scientific research or prevent archiving research data. When asked about the risk of identification people usually say their greatest fear is to become targets for aggressive advertising. In contrast, they generally esteem scientific research. According to the Finnish Science Barometer 2001, Finns trust scientific institutions more than the legal system or the church. According to a UK research, ordinary citizens do not regard researchers as a threat to their privacy. Rather, the people tended to be concerned that already existing research data were not being used sufficiently and in appropriate research (Heeney 2004).

Use of identifiers in scientific research is not harmful by definition. Both the EU directive on the protection of personal data and the ensuing Finnish Personal Data Act allow archiving of data containing personal information. Whenever data are collected directly from participants, the level of anonymisation depends on what kind of information on the use and processing of data has been given to participants. Data can be collected and archived for secondary use if participants have been informed of this. In the ideal case, when planning what to inform to participants, the researcher takes into account both the data protection legislation and the possibility to share the data once the original project has ended. This would permit archiving the data for the use of bona fide secondary users without editing identifiers. After all, this is why data archives ask secondary users to sign legally binding access and use agreements and, in some cases, a pledge of confidentiality.

Even though privacy is protected by the law, there is no unambiguous definition for the concept. The meaning and content of the concept keep changing as the society changes (Saarenpää 2004, 16-17). Culture-spesific factors, age, and sex have an influence on how people define privacy. Even people of the same age with similar backgrounds draw the boundaries of privacy differently. One person may talk about intimate matters in a mobile phone conversation on a public place whereas another might regard a casual question about his family during a break at work as intrusive. Not to mention people willing to reveal all their private problems on a television show whereas other people might not be willing to discuss similar problems even with – or particularly not with – their relatives.

Research participants draw the boundaries of their privacy in two stages. First, when they decide whether they want to participate or not. Secondly, during data collection, when they decide what they want to reveal about themselves and their thoughts to research: they decide what to answer and what not. There is no need for researchers to give automatic promises that data are processed and archived fully anonymised. FSD staff has worked hard to provide guidelines for researchers on how to inform research participants and how to take care of data security and confidentiality when handling research data. Our aim is to reduce the need for anonymisation by the archive staff during the archiving process. Researchers can do the anonymisation themselves or alternatively, research participants are informed in a manner which makes it possible to archive the data with as little editing as possible.

Instead of regarding researchers as potential offenders waiving confidentiality, it would be more sensible to increase the teaching of ethical guidelines and confidentiality issues to students and researchers. If we adopt anonymisation as the unquestionable starting point, we are limiting the freedom of science and reducing the choice of meaningful research questions.

Literature

Guide to Social Science Data Preparation and Archiving. Inter-university Consortium for Political and Social Research (ICPSR) 2005. Accessed 10/31, 2005 (http://www.icpsr.umich.edu/access/dpm.html).

Heeney, Catherine. 2004. "The Role of Privacy and Confidentiality in the Work of National Statistical Institutes." Unpublished Ph.D. thesis, submitted to the University of Manchester.

Kuula, Arja. 2006. "Tutkimusetiikka: aineiston hankinta, käsittely ja säilytys" (Research Ethics: The Acquisition, Processing and Preserving data. 240 pages, available only in Finnish). Vastapaino: Tampere.

Kuula, Arja. (Forthcoming). Ethics reconsidered. In Mathieu Brugidou, Magda Dargentas, Dominique Le-Roux, Annie-Claude Salomon, Gilles Bastin (eds.) SECONDARY ANALYSIS IN QUALITATIVE RESEARCH. Lavoisier: Paris.

Saarenpää, Ahti. 2004. "Yksityisyyden suoja tietämättömyyden yhteiskunnan uteliaisuusympäristössä." Tietosuoja 16(1):12-19. (The protection of Privacy in the Society of Ignorance)