Data Processing in FSD: Challenges in a New Archive

*by Hannele Keckman-Koivuniemi & Mari Kleemola**

The Finnish Social Science Data Archive FSD began operation in 1999 as a separate unit of the University of Tampere. It is funded by the Ministry of Education. At present the archive has thirteen full-time employees.

FSD has archived about four hundred studies, of which one third are international in scope. FSD obtains about one hundred new studies annually from a variety of sources: ICPSR, public and private research organisations and individual researchers. Studies from other archives (for example ICPSR) are documented in Finnish but the data are not processed.

The data that FSD archives must be from an area or discipline in the field of social sciences. It should also fulfil certain technical and legal requirements: aspects of copyright and ownership should be clear, there should be no legislative impediments to archiving (e.g. data protection, privacy protection), the original purpose of data collection should not prevent archiving and, last but not least, the information content and technical properties should make the data suitable for archiving.

Data are processed intensively

The archiving process begins when the depositor delivers machine-readable data to FSD. The data are most often received in SPSS or Excel format and occasionally in SAS or ASCII format. Depositors are asked to give detailed information about the collection procedure, resulting publications and the research project in general. FSD prefers to obtain supplementary documentation both in electronic and paper formats. The completed and signed material description and material deposit agreement forms are necessary as well.

All national studies acquired by FSD are processed intensively – a very time-consuming undertaking. There are two reasons for doing this. First, the number of archived studies so far is manageable. Second, researchers do not usually offer their data. Rather, FSD staff identify studies from, for example, scientific journals and then request the data from the researchers. This way all archived studies are by presumption "important" and deserve intensive processing.

A goal of intensive processing is to make the content of the archived data correspond as closely as possible to that of the original questionnaire. Documents received from the depositor, such as the questionnaire, original observation matrix, variable lists, printed books and articles play a key role in this work. All alterations are carefully documented in the syntax.

FSD uses SPSS in data processing and preserves data in portable format. We have chosen SPSS because it provides - at least at present - the best option for preserving data and identification information (that is labels) in the same file. Since SPSS is widely used, converting material to future formats should be manageable. Most of our customers are familiar with SPSS and about 90 % of researchers wish to have data in this format.

Checking the Data

The original data file is preserved without modification as long as necessary for the archiving process. So far we have not deleted any original datasets. A copy of the original file is used to produce a version suitable for secondary use. We focus on the content, not the "look and feel" of the studies.

During the checking process mistakes are corrected and verifications, amendments and additions are made. We confirm that the number of variables and cases match the documentation supplied. We check frequencies to verify variables, valid and invalid values (e.g. missing data, not applicable responses etc.). Variables are renamed corresponding to question numbers. Background variables without their own question numbers are renamed bv1, bv2 etc. FSD has standardised labels of some background variables (e.g. original question: How old are you? --> variable label: Respondent's age). Variable and value labels are constructed based on the questionnaire. Variable labels often become quite long as they may contain the whole question text. We try to keep variable names and labels consistent within studies of the same series.

Questionnaires often include questions that are directed only at respondents who meet certain requirements. The archive checks these filter conditions. If the data include answers from people who do not belong to the specified

target group, the responses are classified as missing data.

In order to make sure that the content of the archived data corresponds as closely as possible to that of the questionnaire some variables may have to be dropped or added. A variable is dropped if it is undefined or data security aspects so require. Constructed variables, such as combined variables and sum variables, are usually dropped. However, those constructed variables that are integral to the usability of the data, especially weight variables, are kept - providing that the documentation provided is explicit enough. New variables are added only if usability so requires.

Data Protection
Confidentiality aspects require that personal data are deleted. It is recommended that depositors remove these types of data (names, addresses, birthdays etc.) before delivering the material to the archive. Under certain circumstances, FSD stores materials which contain personal data. In these cases, the archive anonymizes the data according to its own guidelines and depositors' instructions.

Variables indicating place of residence and business are also problematic. On one hand there is always a risk that a single respondent might be identified, on the other hand the deletion of these types of variables prevents secondary users from conducting regional comparisons, especially if no other regional variables are used. Usually these types of variables are dropped. If necessary, they can be restored. Variables of larger regional units (provinces, districts) are kept.

Version Control
We aim to process the datasets only once. Still, some of them have to be reprocessed. Sometimes additional information about variables is given by depositors, errors are detected or processing procedures updated. For example, datasets processed in the early days of FSD already need repairing.

The first final version of a dataset is called version 1.0. If subsequent changes are more or less cosmetic (typing errors etc.) the new version will be called 1.1. In the case of significant changes (e.g. a variable added) the new version will be named 2.0.

We add these alterations to the end of the same SPSS syntax file used in processing the dataset in the first place. This is not a long-term solution but has worked so far. We track changes and versions as well as all files in our operational database.

Documentation
FSD uses DDI standard for creating data documentation. At present FSD produces study descriptions in Finnish and English and PDF codebooks in Finnish for all our national studies. All documentation is available on the Internet. Datasets are translated into English on request.

The data and documentation are fully compliant with the search engine NESSTAR. FSD also takes part in the MADIERA project (Multilingual Access to Data Infrastructures of the European Research Area) which started in December 2002.

Challenges for the Future
We need to - and will - review and update our data processing instructions and procedures in the near future. It is obvious that FSD will not have the resources to continue processing all datasets intensively, but will have to introduce several different levels for processing and documenting datasets. First we need to define the minimum level of processing required to preserve dataset quality for the long-term.

One problem common to all data archives is how to get principal investigators to provide enough details about the data and documentation. The intensity of our processing requires a large amount of information about each study. Researchers are often unwilling to take the time necessary to dig up and assemble the details and the lack of information slows the archiving process significantly. Using different processing levels would mean a faster archiving process and quicker publication of data.

Another challenge is version control. We cannot continue controlling versions by amending the syntax file originally used to moderate the dataset because the original data might - and probably will - be unreadable in the future. The SPSS syntax should merely be a tool, not documentation needing to be preserved. Also, as our purpose is to preserve the content, not the "look and feel" of the data, we are not planning to preserve the original datasets forever.

The data for a growing number of studies are collected by computer-aided interview systems like CATI and CAPI. At the moment we process these datasets manually. This needs to be automated as the number of archived CATI and CAPI studies grows. Computer aided interviews also mean that there are no traditional-style survey questionnaires - but our current data processing procedures assume that a printed questionnaire exists.

NESSTAR and MADIERA will enable online data downloading in the future. This may create an entirely new set of problems. Another challenge will be data protection. The measures we take today may not be sufficient in the future. The scientific community is increasingly aware of these challenges. Last but not least is the question of long-term preservation of data and other electronic documents.

In this article we have only touched the surface, the list of challenges surely does not end here. FSD has received a lot

of information of good or best practices in data processing from other data archives in Europe and North-America. Co-operation on the international level has been, and will continue to be, crucial for all data archives, especially new archives like FSD.

* Paper presented at the IASSIST Conference, May 2003, in Ottawa, Canada.

Address: FSD, University of Tampere, 33014 Tampere, Finland, Fax: +358 3 215 8520, Internet: www.fsd.uta.fi. Email: hannele.keckman-koivuniemi@uta.fi - Tel: +358 3 215 8530. Email: mari.kleemola@uta.fi - Tel:. +358 3 215 8528.