Preservation of Knowledge- Data processing in the Danish Data Archives

*by Anne Sofie Fink Kjeldgaard, Søren Priisholm, Birgitte Grønlund Jensen **

Abstract

High quality secondary analysis of sample surveys depends on the quality of the primary data sets and their preservation. The researchers performing the secondary analysis must be able to access as much information about the data sets as possible. In the Danish Data Archives (DDA) great effort is taken to preserve the data sets in a way that meets the needs of the secondary researcher. For this reason data processing is a core operation in the DDA and great importance is attached to producing reliable and useful documentation of the preserved data files.

Data processing is a core activity in the Danish Data Archives[1] (DDA). It is the performance of data processing that makes DDA unique in comparison with alternative data preservation efforts in the Danish research world. Despite the importance attached to the activity of data processing, it is an activity that is invisible to outsiders. In this article we set out to discuss the advantages of data processing for the depositor of the data set, for the end-users when performing secondary analysis on the data set and for the data archive. In this light we will then describe our preservation strategy in detail. Finally, we will guide readers through the data processing process step by step and point to future development.

The Advantages of Data Processing

Data processing has advantages for the depositors, the end-users and the data archive. These advantages can be traced back to the effort of collecting and integrating all available information about a data set. The physical product of the data processing process in the DDA is a Data Documentation Publication (DDP) consisting of a study description and a codebook with frequency tables for all variables in the data set. The study description holds all information about the creation of the data set and facts about its preservation in the DDA, as well as restrictions for access to the data set. The codebook consists of all available documentation about the data file and a copy of the original questionnaire.

Data Deposition

Although it ought to be a straightforward task to add together the information for the study description and the codebook, most often it is a time-consuming job. It becomes especially difficult if it has been a while since the actual survey was carried out. In order to make it as easy as possible to gather information about the study, the DDA urges researchers to deposit their data as early as possible in the research process.

After a data set has gone through data processing a Data Documentation Publication (DDP) is created. The depositor gets the message that her study has a DDP, i.e., a study description and a codebook. She can then be sure that her data set is preserved for the future and she will have no need for storing the data elsewhere.

Data Location

As soon as the data material has a study description it becomes searchable in the DDA's data catalogue on the Internet. A future challenge in this respect is to allow users not only to search the study description, but also to search the whole DDP. At the moment, the DDA is taking part in the MADIERA project[2], which, among other things, will offer users this opportunity.

Secondary Analysis

As regards data analysis, data processing offers essential advantages for end-users. First and foremost, it becomes straightforward to get access to all information about the origin of the data set and its contents because all information is at hand in the DDP. However the DDP is not just a collection of information provided by the depositor. During the data processing process, several additions, standardisations and checks are made. For example, if divergence between the data and the questionnaire is found, the person in charge of the data processing will make a comment about this in the codebook, thereby eliminating the need for future users of this particular data set to spend time finding "wild codes" themselves.

Data Archiving

Internally, data processing has the advantage that users of the data sets seldom need guidance upon having received a data set. As a consequence, the time spent on providing user services is reduced in spite of increasing use of our data sets over the years.

Preservation Strategy

If the DDA is to promise depositors long-term preservation of their data sets, data and documentation must be preserved in a simple format in order to ensure that all content can be read and analysed in the future. Obviously this promise must be made without knowledge of the composition of future computers and their statistical software. Therefore, any kind of software and hardware dependence has to be removed and the data set has to be documented as completely as possible.

The majority of the data sets in the DDA's collection are one-off cross-section sample surveys. In these cases there is only one data file and one documentation file to be matched. Although in practice many data sets relate to each other, e.g., surveys replicated in time, these are handled as separate data sets for the time being. The DDA is part of the Metadater Project[3], which is working on a solution to reflect relationships between studies, by both study naming conventions and methods of storage.

In the DDA, the data material is said to consist of two parts: A data file and a documentation file. This means that the DDA has one job concerned with preservation of the data file and another job concerned with preservation of the documentation file.

The data file is a data matrix that constitutes the substance of the quantitative data material. This data matrix contains codes for respondents/cases, questions, categories for answers, and additional codes the researcher might have added to the data set. The documentation file is the information that describes the origin and contents of the data. In other words, documentation is the information users need to make sense of the data matrix. An example is that in the data matrix you can see that respondents in column no. 47 have answered either 1 or 2. In the documentation you can see column 47 contains information about the respondent's sex (male = 1 or female = 2) as answers to question no. 30: "Are you male or female?" in the questionnaire. All this documentation is gathered together in the codebook.

For DDA, the data preservation strategy has two strands. Both the data matrix and the documentation have to be preserved technically and physically, and the semantic information/documentation in the study description and the codebook has to be preserved in a way that ensures that the data stay 'understandable' to future users.

Technical and Physical Preservation

At a very early stage, the DDA and other data archives decided to use the system-independent format OSIRIS III (symbols 0-9) as the technical preservation format. OSIRIS III is an extremely simple format that was developed to describe existing data materials useful for sample surveys. As soon as DDA receives data material, the data are transferred to a DDA server. In most cases it is possible to make a conversion of the data to OSIRIS format at once, thereby creating a long-term preservation file at this stage. Data are processed in order of priority according to novelty, demand from users, relation to other studies, etc.

From the OSIRIS format, data can easily be converted to up-to-date formats like the current versions of SAS, SPSS, STATA, etc. As stated above, it is an advantage to users to get data material that has been processed; otherwise they must cope with the original data file themselves.

Data are physically stored on a server with a daily backup routine.

Although technical and physical preservation is just as important to data preservation as semantic preservation, it is most often much more straightforward, depending on the complexity of the material, of course.

Semantic Preservation

The semantic preservation is carried out to ensure that all available information about the data set is collected and merged with the data matrix. As mentioned above, the documentation of a data set has two main components: A study description describing information about the depositor, the production of the survey and access conditions, and a codebook with identifying keys to translate the codes in the data matrix into something meaningful. The DDP consists of the study description and the codebook with frequency tables.

The Study Description

The study description consists of a large amount of background information that the user must take into account when drawing conclusions on her analysis of the data set. The study description has information about the following subjects:

It has always been an objective to prepare a study description upon reception of a data set. However, in practice it often happens that the study description is not prepared until the data material is processed. The unfortunate consequence of this is that there is a delay before the study description appears in DDA's search catalogue.

DDA recently introduced a new programme that was developed in-house for the preparation and preservation of study descriptions. From now on, study descriptions will be placed in a database structure, which allows us greater flexibility than we had with the fixed text format we used before. The new programme offers a template for structuring information inputs. This allows everyone in the archive to take part in the preparation of the study description and to gather the information step by step. The programme has benefits to the end-users too, since

Table 1: Subjects in the Study Description
General information:
Title and year of study, Status of the study, Classification of the study in cluster(s), Relevant keywords for the study, Language employed in the present study description
Identifications and references:
Bibliographic reference, Local archive where the study is stored, Archive where the study was originally stored, Depositor (donor), Date of deposit, Primary investigator (research organization), Data collector, Research initiator, Funding agency
Analysis conditions:
Research topic (abstract), Kind of data, Units of observation, Number of units (cases), Dimensions of data set, Completeness of study stored, Time period covered, Time dimensions, Definition of total universe (universe sampled), Sampling procedures, Geographical area covered, Dates of data collection, Method of data collection, Type of research instrument, Actions to minimize losses, Data gathering staff, Characteristics of data collection situation noted, Weighting
Re-analysis conditions:
Current data representation, Applicable analysis packages, Language(s) of written material, Control operations performed by primary investigator, Control operations performed by archive, Accessibility, Access directing authority
References to the relevant publications/results/studies:
Publications/reports from primary investigator
Variables included:
Basic characteristics, Residence, Household characteristics, Characteristics of parental family/household, Occupation, Education

information in the database will be become simultaneously searchable on the web.

The Codebook

The DDA codebook is the main product of the data processing process with its translation of the data file into human understandable information. It consists of a description of variables containing information about the positions of every single variable in the data file, the exact text from the questionnaire and a number of other essential pieces of information. It also explains any discrepancies between the data and the documentation that were found during processing.

An OSIRIS codebook is a text file in punch card format, which means that each line must not exceed 80 characters. It is, however, possible to insert a continuation card (k-card). The first 10 columns are assigned as follows: column 1: Card type; column 2-5: Variables number; column 6-9: Reference number; and column 10: Typically blank. Various types of cards for different kinds of information are available to the data processor.

It should be mentioned that OSIRIS is a preservation format – not a presentation format. The DDA has developed a separate presentation format for codebook information.

Data Processing Step by Step

The data processing process is described below and illustrated in the accompanying flowchart.

1. Allocation of study number to the data set
When the DDA and the depositor have made a deposition arrangement, the data material gets allocated a unique study number in DDA's journal. This number will tie together all parts of the data material stored in different places on DDA servers.

2. Reception of the data set (data0.org)
The DDA has no required format for data sets. The archive has a wide selection of software programmes to convert the original data to an OSIRIS temporary file. This file is preserved until data processing has been completed.

A check is performed to ensure that the essential documentation (questionnaire, data files and documentation files) has been transferred to the data archive.

The original questionnaire is scanned and stored as a PDF file.

3. Control of data (data1.org, createData.sas)
A check is performed to test if the number of respondents and variables are in accordance with the documentation from the researcher. Frequency tables for all variables are printed, in order to perform a more thorough control of the material.

4. Data recoding and documentation
Recoding and modification is performed in a SAS script. The first step is to insert two standard variables: The DDA study number (V1) and a sequence number (V2).

The value labels are then modified so they do not exceed 24 characters and have expressive names. Variables are recoded for filters and formats, e.g., length and numbers of decimals. Values for missing data are recoded into the DDA standard format. Three categories are used:

- Not answered (9,99 ... etc.) – the respondent has not answered the question
- Inappropriate (10,100 ...etc) – the respondent should not answer (due to a filtering variable)
- No participation (11, 101 ...etc.) – the question

was not presented for the respondent (this typically occurs when surveys with different variables and/or respondents are merged together)

By the use of standard missing codes, the work for the end-users is made easier. This is very important when making comparisons between several data sets.

In parallel to the recoding, a codebook is edited in an editor (KEDIT). The job here is to write the exact text from the questionnaire from top to bottom into the documentation file.

For all variables, the codebook contains frequencies in figures as well as in percentages.

Finally, the variables are renamed to a continuous variable row.

The data processor then makes a profound check of the data to ensure that the data definition is in accordance with the original data and the documentation. DDA does not recode data if there is a discrepancy, but makes a note in the codebook about this. Such a note can save researchers or students using the data set hard work.

The questionnaire in PDF format is converted into a text file.

5. Data processing of data and codebook
By running a script on the data and codebook and by merging the codebook and data file, the Data Documentation Publication (DDP) is produced. Beforehand a preface and a study description are inserted in the codebook. The final files are DATA.OSI, which is a simple data matrix, and DDP.LAS, a codebook file that contains print codes for page shifts and page numbers.

6. Preservation and user services
The processed material is stored in two parts – a data file and a documentation file. The two files and the SAS script file are preserved on the server.

7. Proofreading internally and approval externally
When the DDP is produced, there will be a proofreading of the produced material by a staff member who has not been involved in the data processing process so far. If the proofreader finds mistakes, the material is given back to the data processor, who corrects it accordingly. Afterwards, the proofreader goes through the material a second time to make sure that mistakes are corrected.

The material is now send out to the depositor for her approval of the processed data set.

8. User Services
When a user requests the data material for secondary analysis, the data and the documentation files are merged by use of a program (OSI2SPSS) that creates a SPSS file. If the user requests a format other than SPSS, further conversion must be made. When uploading to the Internet, the two files are merged by the use of a program for XML and HTML files.

The flowchart below shows the data processing step by step.

- All steps marked with italics indicates that the data processing step is done manually.
- All filenames are written in bold characters.
- All scripts/programs are put in quotation marks.

Future Development in Data Processing

From the flowchart (table 2) it should be obvious that data processing is a very time-consuming activity for the DDA. At the moment the average amount of time spent on data processing is about 150 hours per data material. However there is great variation, as data processing of some studies only takes about 50 hours and, obviously, some take much more. In the data processing process it is codebook editing and recoding and renaming of variables that are the great time consumers, whereas running scripts is something that is quickly done.

In order to reduce the time spent on data processing, it is planned to develop a new programme for data processing. The objectives for this development will be to minimise the time spent on data processing and to implement a higher degree of standardisation of the process. For example, at the moment, 10 scripts are used in every data processing process. Hopefully this can be cut down to a single one. Similarly, the number of working files may be cut down from 16 to one.

With the introduction of the new programme, we can look forward to a substantial reduction of the time spent on data processing. We will then be able to process more materials than we can today, and thereby be able to provide our users with more high quality materials faster.

* Authors: researcher Anne Sofie Fink Kjeldgaard, data archivist Søren Priisholm and information science officer Birgitte Grønlund Jensen, the Danish Data Archives. Contact: Anne Sofie Fink Kjeldgaard, asf@dda.dk.

Table 2: Data Processing of a Study in the Danish Data Archives[1]

Data	Documentation
Data0.ORG* (original data from the depositor, e.g., SPSS or SAS files)* ↓	
Formal Control (check of machine readability, number of variables and respondents are right according to documentation) ↓	
Data1.ORG* (checks data in *.por or *.sav format)	
DanData.SAS (creates new SAS script file using the editor KEDIT) ↓	
Use **DanData.SAS** to read **Data1.ORG** into SAS = Creates Frequency tables (**SasData.SD2**)	
Print **SasData.SD2** *or the frequency tables.* *At this point it can be estimated how many resources to be put into data processing of the study, e.g., by checking the quality of data documentation.* ↓	
Changing label- and variable names and recoding of variables according to DDA's guidelines (this includes re-coding for filters). This is a "Trial and error"-process. Saving and running the script **DanDataX.SAS***) *using SAS – over and over again. Output file is* **SasDataX.SD2**	**Cdbk.txt** *(create new codebook file – using an editor such as KEDIT). The questionnaire in PDF format is converted to a machine readable text file. Comparison with* **DanDataX.SAS**.
Running "SASOSIR" script on **SasDataX.SD2**	Run "KodeByg" script on **cdbk.txt** (this script moves the codebook text in accordance with the OSIRIS format) **Cdbk.osi** (converted codebook-file)
Converted files (**Data.OSI** and **Dict.OSI**) **Data.OSI*** = DataMatrix **Dict.OSI** = Dictionary (List of labelnames or T-cards) ↓	Run "MERGET" script on **Dict.OSI,Cdbk.OSI** **Dicb.OSI** (codebook with T-cards/labels) Run "SASMARG" script on **Dicb.OSI, SasDataX.SD2**
Manual check of **Dict.OSI** *(All variables are recoded and label names are max. 24 characters long.)*	**DicbM.OSI*** (Codebook with correct margins) *Copy/Paste writing of* **Fortale.osi** *(a preface/short introduction to the study) into* **DicbM.OSI**

<table>
<tr><th>Data</th><th>Documentation</th></tr>
<tr><td>.. From last page...</td><td>From last page..
Run “DICBDOK” on <b>DicbM.OSI</b>
<b>DicbM.DDA*</b> (codebook with ‘Guide to the codebook format’)
<i>Writing of study-description</i> <b>S1234GB.SD</b>
Run “DDASDF” script on <b>S1234GB.SD</b>
<b>S1234GB.DDA</b> (Change item-number and code to standard-text.)
<i>Copy/paste</i> <b>S1234GB.DDA</b> <i>into</i> <b>DicbM.DDA</b>
<b><i>DicbM.DDA</i></b>
Run “DDALIST” on <b>DicbM.DDA</b>
<b>DicbM.LAS</b> (or <b>DDP.LAS</b> – the print file without any internal codes, but with printing information such as ‘new page’, ‘page number’, etc.)
Print <b>DicbM.LAS</b></td></tr>
<tr><td colspan="2"><b>DDA preserves</b>
1. Data.OSI
<b>2.</b>
3. Doku.OSI (=DicbM.DDA)
4. DanData.SAS (consists of all used SAS-statements = documentation of the process)
5. Data1.org/Data0.org</td></tr>
</table>

Footnotes

[1] The DDA was established in 1973 as a national data bank for quantitative research carried out primarily in the social sciences but also in medical science and history in Denmark. In 1993 the DDA became an independent unit in the Danish State Archives. At present the archive has 13 full-time employees. The DDA collects, preserves and disseminates machine readable research data.

[2] Read more about the MADIERA Project here: www.madiera.org.

[3] Read more about the Metadater Project here: www.metadater.org.

[4] Rasmussen, Karsten Boye, 1996, *Oparbejdning med SAS i Dansk Data Arkiv* (*Data Processing using SAS in the Danish Data Archives*), version march 1996, DDA, Odense, p. 21.