Tabulations on the DDA study description

by Karsten Boye Rasmussen[1]
Danish Data Archives
Odense University

The object of this paper is to give a brief introduction to the Standard Study Description and to add a few remarks on its recent history and development. For those already familiar with it, the first section will also answer the question: 'Whatever became of the ACCESS project?' The main part of the paper concentrates on a presentation of the holdings at the Danish Data Archives (DDA) in the form of cross-tabulations based on a data file compiled from the contents of the study descriptions at the DDA.

The Study Description

For those not familiar with the term 'Standard Study Description', allow me to clarify: the phrase is meant to be read backwards.

First of all, the Standard Study Description is a 'description'. It is a machine–readable document written in a specific format. It is like a library catalogue card, except that the object is not a book but a machine–readable data file, or, to take another step backwards, a study. The data file or study is, for the purposes of this paper, within the broad area of the social sciences.

The format of the Standard Study Description (SSD) is somewhat complex. The SSD contains a large number of items, which can be compared to variables. Each item consists of a numeric identification code and an entry containing specific information. What makes the format complex is that the different entries may contain different types of information. At the IFDO/IASSIST conference in Grenoble in 1981, I presented a working paper[2] which extensively expounded the format of the SSD. In this paper, a few examples should suffice to illustrate the complexity of the SSD. For example:

1. • Item 101 contains the title of the study. It is an unstructured text item.

2. • Item 212:04 contains the unweighted number of cases in the data file. A numeric subitem (:04) inside a structured item.

3. • Item 222 describes the target population using predefined codes. A precoded item.

---

[2]Karsten Boye Rasmussen: "Proposed Standard Study Description". Working paper presented at the IFDO/IASSIST conference, Grenoble, 1981.

Thus the study description contains text items, numeric items and precoded items; of these, the precoded item is the most complex.

In precoded item 222, an entry "01" will signify that the target population is restricted by "age limits". There may be further restrictions specified by further codes. When a precoded item is viewed as a variable, what we have is indeed complicated (like a multipunch column on an old punch card).

To complicate things further, the SSD format allows a text explanation of unlimited size to follow any type of item, whether it be a text, numeric, or precoded item.

The SSD is like a data file: it is incomprehensible without the proper documentation describing the significance of the item numbers as well as a 'codebook' describing the codes used in the SSD. This documentation is, of course, machine-readable as well. With a computer program, one can merge the SSD data file and the codebook information to produce a human-readable printout describing the study.

Working backwards we have now reached the last word in the term 'Standard Study Description'. The SSD has not yet been accepted as an international standard. Given that this year is the 12th anniversary of the SSD, I shall not insist on the word "standard". Rather, in the remainder of this paper, I shall refer to the 'Study Description' or the SD. On the other hand, it is a standard of sorts, in so far that some of the European social science data archives use the SD extensively, with only minor differences.

Those interested in the true story of the Standard Description are advised to read the paper "Standard Study Description as a meta research data base"[3] given by Per Nielsen at the

1983 IASSIST conference. That paper outlines the development of the SSD and includes references to historical papers on the subject. This paper is an updated version of Per Nielsen's paper.

The Function of the SD

From the beginning, the object of the SD has been to fulfil several functions as a tool for the data archives and the social science community. Some of the functions have changed, primarily because new technology has facilitated the achievement of new goals.

1. Data abstracting and catalogue production

The main function of the SD is (as described above) to produce human-readable printouts of the study descriptions. This function is what was originally[4] termed "data abstracting and catalogue production". Since 1978, the DDA catalogues have been produced using computer programs to generate phototype setting instructions from the SDs and a 'skeleton file'. At the same time, a considerable amount of indexing of the descriptions is done automatically. The same procedure has been used at the Zentralarchiv in Cologne, at the Steinmetzarchief in Amsterdam, and is presently being used at the ESRC Data Archive in Essex.[5] (It should be noted that other data organizations use similar techniques, they are

---

[3]Per Nielsen: "Standard Study Description as a meta research data base". Paper presented at

[3](cont'd) the IASSIST conference, Philadephia, 1983. (Reprinted in DDA-nyt nr. 26, 1983, pp. 5–23)

[4]Per Nielsen: "Study Description Guide and Scheme". Copenhagen, DDA, 1975.

[5]ESRC Data Archive Bulletin, January 1987, no. 33, p.1.

not mentioned here because they do not use the SD as the basis for their catalogue production.) At the DDA we are now completing the 1985 catalogue of holdings. Only a subset of items from the study descriptions is being printed, but a number of COMfiche containing the complete descriptions will be supplied with each copy of the catalogue.

The production of catalogues has many drawbacks. First of all, it is very expensive. Of course all the SDs are available, as they are produced as part of the documentation process when data are deposited in the archive. Nonetheless, a lot of proof reading is necessary before the catalogue is ready for printing. Another drawback is the problem that the catalogue is outdated before it even reaches the market. Thus, online access to a computerized catalogue is necessary for up-to-date information.

2. Mapping and Methodological Research Base

The collection of SDs has long been regarded as the perfect object for methodological research or presentation of holdings. But the perfection resides in the very detailed information they contain, and not the process of compiling the information from the SD format to a rectangular data matrix ready for analysis. Because of the complex format of the SDs, such compilation requires some computing, to produce a rectangular numeric file without text information. This paper includes a chapter presenting the holdings at DDA with tables based on such a rectangular file. All the programming, including the extraction of information, was done with the software package SAS.

3. Data (Re)Analysis Prerequisite

At the DDA, the documentation of studies deposited in the archive includes a machine-readable codebook with complete questionnaire text as well as coding instructions plus one-way distributions of each coded variable. The codebook describes the data at the variable level. It also includes the total Study Description describing the background, objectives and outcome (publications) of the data collection. Thus the SD is a prerequisite for the process of secondary analysis.

4. Intra-Archival Loggin

This function, which was to supply the archives with a tool for keeping track of the processing of their data, has completely lost significance in the technological race of the last decade. The development of interactive data bases with immediate updating facilities, as opposed to a sequential and both time- and costconsuming method, has, at least at the DDA, led the archive to implement data base applications which have the power to keep the most important information ready at hand ("just a PF -key away").

5. Inter-Archival Exchange

However the SD is still a standard for inter-archival exchange. It is an exchange format which is simple enough to be read into any machine (including microcomputers). Special software is needed, however, to process the SDs.

It is therefore my impression that the SD will remain a standard format for exchange purposes, and that its list of items will be a check-list of the kinds of information which should be provided for each study. The actual storage mode of the SDs at individual archives may differ, depending on what data base facilities are available, but the data base application should be able to both 'export' and 'import' SDs in the standard format.

Information Retrieval Data Base

The archives in Cologne, Amsterdam and Odense have developed retrieval systems for searching their own SDs. The most comprehensive system is the ZAR system at ZA in Cologne, which has been described earlier in IASSIST surroundings. The data archive at Essex has recently announced[6] that they are setting up an online information retrieval system. Other archives have information retrieval systems as well, but the archives mentioned above are all using the SDs.

The ACCESS Project

The idea of using the SD-format as an exchange format as well as the possibility of setting up retrieval systems on the basis of the SDs, led to a project amongst the European archives within CESSDA (Committee of European Social Science Data Archives). Under the project heading "ACCESS: Integrated European Archive Inventory" a catalogue was to be published, SDs to be exchanged, and a data base retrieval system to be set up with common

access (available through EURONET). The EEC was to finance the project, but the demands of the bureaucrats in Brussels made the project much less attractive, and it was finally abandoned.

The project has instead developed into an ongoing effort at the four archives (mainly the DDA). But due to a lack of funding, this project competes with regular activities of the archives and has therefore often been postponed. The tables in this paper are based on the collections of a single archive, the DDA. It is my hope that within a reasonable time period I shall be able to present similar tables comparing the holdings of the four European archives. At present the DDA has received SDs from the Steinmetzarchief; with the completion of the ESRC Data Archive catalogue, the DDA will receive a new batch of SDs, and finally the SDs from ZA.

Setting up a retrieval data base as described in the ACCESS project will demand a considerable amount of work. At present, network facilities are still not sufficiently effective to supply online access to other computers. When these techniques have been improved, the online Integrated European Archive Inventory will become a reality.

As mentioned above, the four archives use slightly different formats for the SD. As long as the differences are fully documented, this presents only minor problems. The ZA and DDA formats are very similar, although the ZA does not use as many items as the DDA. At the Steinmetzarchief, the numbering of the items is different, but the mapping of the formats is the same. At the ESRC-DA, depositors, etc. are identified by a number from a special file. At the DDA, the latest change in the SD format has introduced an item (220) pertaining to historical data materials, which shows the time period covered by the data, and an item (225) to show to what regional area or country the data describe.

[6]ESRC Data Archive Bulletin, January 1986, no. 33, p. 1.

Tabulations on the DDA SDs

The tables in the remainder of this paper describe 962 studies. These tables will not be extensively commented upon nor compared with the findings of Per Nielsen in 1983[7] as the purpose of this section is not primarily to comment on the development of the data holdings of the DDA since the 645 datasets were investigated in 1983. The tables refer to the datasets described in the forthcoming DDA catalogue and give an overview of the contents of that catalogue.

There has been no exclusion of datasets with missing data. For some items, missing data should not exist, but for others the existence of missing data is perfectly alright.

Some of the variables are coded with multiple codes. To illustrate this, some of the tables have information on the number of codes. If, in a precoded item, a study has more than one code, the entries are weighted accordingly (e.g. 2 codes, weight=.50, 3 codes, weight=.33 etc.). Therefore even in these multiple response items, all the tables still total 962 studies (apart from rounding error).

Contents of the SD Data Bank

The SDs at the DDA are very extensive in their description of the background of the study. Table 1[8] contains the distribution of SDs by the number of lines they contain. It shows that a typical study has between 51 and 200 lines of information in its study description.

The mean of the 962 studies is approximately 160 lines of information. The comparison between an SD and a library catalogue card is therefore not a very good one. The SDs at the DDA are more like a library catalogue card plus a very elaborate abstract of the study. The magnitude of information shows the potential for making a retrieval data base with the SDs as input data.

In an ongoing archival process, not all studies have optimal documentation; many studies are not yet fully processed. A few remarks about the DDA level of documentation may be useful here. At the DDA, two categories are of especial interest to users. Studies in class "D" contain a complete machine–readable codebook. From this documentation, we generate setups in SPSS– or SAS–format for the user as well as deliver published documentation on the study. Studies placed in class "C" do have some machine–readable documentation, but are not as "polished" as studies in class "D", and do not contain a machine–readable codebook.

The Study Description item '001' contains the status of the study. However, at the DDA, this item is updated by extracting information from our data base which keeps track of the processing of studies. As this process had not taken place at the time that I computed these statistics, I have computed table 2 directly from the processing data base for the purpose of showing the status of the studies.

This total differs from the number of studies drawn from the SDs. This difference is due to the fact that all the other statistics are made on the basis of the SDs to be published in the catalogue of holdings. Since the cut-off of new additions to the catalogue, 141 studies have come to our attention; these studies are typically placed in class L.

---

[7] op.cit.

[8] (tables have been collected together at the end of the article. Ed's note)

Description of the Studies

This section will illustrate the kinds of studies available from the DDA.

Table 3 shows the subject headings or areas covered by the DDA holdings.

The table shows a heavy bias towards the traditional areas of the social sciences. Election studies, general sociology and political science together constitute more than 60 percent of the studies.

A similar picture is displayed when looking at the kind of data on which studies are based. As shown in table 4, 83 percent of the studies consist of survey data.

Earlier, approximately 76 percent of all studies had been generated by the old-fasioned oral interview. Since then, there have been an increasing number of mail surveys. It is possible that the rising cost of conducting the traditional personal interview has led investigators to use mail surveys. Furthermore, it is my impression that many surveys are now carried out by telephone, but this is not yet reflected in the distribution of these data. Typically, the time span between data collection and the deposition of data in the archive is between 3 and 4 years. (Table 5).

The studies stored in the DDA also appear to be conservative with respect to the cases on which the data are based (the "units of observation"). Close to 90 percent of all studies are based on individuals. (Table 6).

In table 7 the definition of the universe is shown. A central variable is age. Most election surveys concentrate on those over 18 (which is the age at which one obtains the right to vote in Denmark). Please note, that this table indicated a frequency total of 1303. Of the 831

studies described in this item, many have more than one code specified. Of interest are also those categories which are lacking in the table below. In the SD 'skeleton' file, provision is made for definition of the universe in terms of 'race' or 'religion'. These do not occur in any of the DDA studies.

When we consider the applied sampling procedures, we find that most surveys fall into one of two categories. Those coded 'no sampling' are typically studies carried out at a distinct location (e.g. a working environment). The other major category is the many studies (25 percent) which are based on some kind of multi-stage sampling. (Table 8).

Without performing detailed cross-tabulations, it is nonetheless easy to outline the characteristics of a typical study in the DDA collection: it would seem to be an election study, carried out as a multi-stage sample survey, with individuals as participants and as units of analysis.

The DDA Data Bank As Potential For Analysis

It is interesting to note that, according to the frequencies on Item 211, approximately 23 percent of the studies are panel studies. This high percentage is due to an agreement between DDA and the public opinion and marketing bureau Observa A/S to the effect that all their political panels from 1967 to the present are being stored in the DDA[o]. Apart from the Observa studies, the remainder of the panel studies are typically election studies also.

It is also worth noting, in table 9, that the the largest number of studies are in the category of cross-sectional sectional studies with replication

[o]The OBSERVA project was described by Karsten Boye Rasmussen and Lone Borgersen in DDA-nyt nr. 34, 1985

of one form or another. These, together with the large number of panel studies, comprise a majority of the studies at the DDA which are connected as part of a panel study or having variables which are very similar, and which therefore present great potential for secondary analysis.

The European archives have often tried to identify studies carried out within the same period of time in a number of the European countries. At the DDA we have introduced a new Item (225) to indicate area of coverage.

Table 10 shows that the vast majority of the studies at the DDA are national and therefore cover all Denmark. About 40 of the studies, however, are cross-national. These studies are typically the Euro–barometers conducted for the European Economic Commission, but within the last few years the DDA has identified some other cross-national studies. It should be noted here that the DDA does not publish information about studies also stored in other data archives, unless they contain information on Danish matters.

To obtain access to a dataset deposited in the DDA, the user is asked to fill out a requisition form and send a one–page description of how the data are to be analyzed. This information is then sent to the depositor or other person authorized to permit access. Most of the studies (65 percent) are without any access restrictions for the typical user working in the social sciences. Surprisingly, a large number of studies are categorized as being available only by special arrangements with the access–granting authority. At the DDA, we prefer not to store studies that are not available to users. The studies in this category may possibly be those of which the principal investigators have not yet finished analysis, but it might be worth rechecking the access conditions of studies in this category. (Table 11).

Time of Study and the Data Matrix

In this section we investigate the hard facts of some numeric items from the studies placed at the DDA. A discussion of whether this material may in any way be regarded as being representative of social science data in Denmark can be found at the end of this paper. Tables 12–14 are believed to show nothing but the distribution of the studies placed at the DDA.

One of the key variables is the starting year of the time period covered by each study (Item 220:01). This new variable is not be confused with Item 231:01, the data collection date. For surveys, these two dates will of course be identical, but for historical studies based on old documents (e.g. parish registers or census lists) Item 220 is indispensible. Subtracting Item 220:01 from Item 220:02 (end year) shows that 75 percent of the studies start and end in the same year. On the other hand, 15 studies cover a time period of more than 100 years.

In table 12 the start year is cross–tabulated with a grouped variable containing the number of cases in each data file.

Although the DDA was founded in 1973, more than 20 percent of the studies in the archive deal with a time period previous to that year. Of the 62 studies covering the period before 1950, 42 studies are concerned with a time period before 1900. These historical studies have typically a large number of cases, but are problematical in that they are not easily sampled. The cases are inter–related (i.e. family reconstitution data) and therefore all cases must appear in the data material so that the relationships can be determined by computer.

Item 212:01, the number of cases, is missing in approximately 28 percent of the studies. Most of these studies are new. The number of cases is missing because for many of these studies

only a limited description is as yet available; DDA has as yet received neither the data file nor precise information about the file dimensions yet. The typical data file has between 800 and 2500 cases.

These newer, partially documeted studies are also a major portion of the great number of missing cases in table 13 below showing the number of variables in the datasets.

The studies dealing with the time period before 1950 (the 'historical' studies) contain, as expected, a low number of variables. The information concerning the historical cases is not very full, but the number of cases is — as shown in the previous table — often huge. In the course of the last 25 years there seems to have been a rise in the number of variables in a single study. The average number of variables is approximately 160 variables per dataset.

Table 13 shows the cross-tabulation of number of cases by number of variables. As has already been mentioned, there seems to be a weak reverse relationship. On the one hand, the datasets with few variables have a large number of cases. On the other hand, the datasets with more than 40 variables are concentrated around the 801 to 2500 case size.

Studying Data Archives or Social Science

The tables in this paper have shown the contents of the data bank at the DDA, the interrelationships between the studies, and their accessibility. At the same time, the tables have served as a test of how the SDs at the DDA are being completed by presenting an overview which is very difficult to obtain when the SDs

are viewed as isolated entries.

The collection of descriptions of Danish social science studies provides an opportunity to examine this collection as a sample of Danish social science empirical research. But is it possible to draw valid conclusions from this sample? I do not intend in this paper to present a solution to this problem. But I should like to discuss some of the problems of bias that must be discussed before any conclusions can be drawn from the collection of SDs. It is my intention, in raising these problems, to stimulate a discussion which will be of benefit to the future comparative analysis of the characteristics of the data holdings of the other European archives.

It is indeed questionable if the collection of studies in the DDA is representative of Danish social science research.

First of all, the studies represented consist of data collections or empirical studies. Secondly the studies must be machine-readable, which will normally mean that computer analysis has been performed on the data. The target of the analysis will therefore at least be limited to Danish machine-readable empirical social science studies.

The most serious threat to the validity of the analysis is whether or not there has been a change in the DDA's criteria for incorporating a study into the data archive. Given the limited resources available at the data archive, it is to be expected that over the years some changes in the basic criteria may have taken place. Furthermore, it is to be expected that such a 'drift' in criteria may have happened unobserved and without being part of explicit archival policy.

One way to prove or disprove this hypothesis would be to compare the holdings of the DDA with a complete inventory of Danish social science research. But the DDA catalogue of

,"start":"iassist quarterly"}]

[]

holdings is the only available source which is close to being a complete inventory. Such a comparison would result in tautological nonsense. Instead, a few limited areas could be compared. The DDA has deposition agreements with some research institutions as well as with the Danish Social Science Research Council. Thus the DDA could check whether these agreements are being fulfilled, or if some studies are for any unknown reason not brought to its attention.

Until there has been a further, more thorough investigation into the representativeness of the studies placed in the data archive, the tables above cannot be construed to be representative of social science research. On the other hand, even if the representativeness of the sample is not tested, we can argue as follows:

Bias is inherent in the selection of social science data files for archival storage. One major source of bias is technical. A simple 'rectangular' survey file is the archetypic data file in data archives. These files are practically ready for storage on receipt by the archive, while a hierarchical study demands more data processing by the archive. Both types of studies, of course, need the production of the proper machine–readable documentation.

The other major source of bias in the selection of social science data for archiving lies in the nationalistic characteristics of the social sciences. Because of the similarity in standards and technical capabilities of the four European archives (in Germany, Great Britain, the Netherlands and Denmark), the technical bias can be isolated.

Based on these assumptions, a table showing the differences in the holdings at the European archives may serve as a guideline for the actual differences in social science research being carried out in the four respective countries.◻

Tables

"Number of lines in SDs"		
100000	frequency	percent
1-50	56	5.8
51-100	395	41.1
101-200	458	47.6
201-300	44	4.6
301 +	9	0.9
Total	962	

Table 1

Class/Status	frequency
D: fully machine readable documentation	285
C: no codebook	56
B: available from primary investigator	72
O: being processesd/ongoing acquistion	456
L: only preliminary donor contracts	234
Total	1103

Table 2

"AREA"		
ITEM 002	FREQUENCY	PERCENT
missing	2.0	0.2
organizational	61.5	6.4
general sociology	134.8	14.0
history, demography	53.1	5.5
law & criminology	25.0	2.6
political science	91.1	9.5
social physics	33.7	3.5
social medicine	75.8	7.9
welfare & leisure	48.7	5.1
socialization	41.7	4.3
election studies	363.2	37.8
macroeconomics	17.0	1.8
microeconomics	13.8	1.3

(Total # of codes = 1358)

Table 3

"Kind of Data"

ITEM 202	FREQUENCY	PERCENT
missing	9.0	0.9
survey	797.5	82.9
census data	9.1	1.0
statistics	52.0	5.4
legislative roll	4.0	0.4
clinical data	8.7	0.9
textual data	3.2	0.3
coded textual data	23.2	2.4
coded documents	55.3	5 8

(Total # of codes = 1042)

Table 4

Method of Data Collection"

ITEM 232	FREQUENCY	PERCENT
missing	100.0	10.4
oral intreview	452.7	47.1
telephone survey	13.8	1.4
mail survey	317.5	33.0
pencil & paper	73.5	7.6
psychological test	0.5	0.1
other	3.9	0.4

(Total # of codes = 1019)

table 5

"Units of Observation"

ITEM 211	FREQUENCY	PERCENT
missing	80.0	8.3
individuals	858.0	89.2
families/household	16 0	1.7
groups	5.5	0 6
other	2.5	0 3

(Total # of codes = 972)

Table 6

"Definition of Target Population"

ITEM 222	FREQUENCY	PERCENT
missing	131.0	13.6
age limits	535.5	55.7
sex	5.7	0.6
marital status	1.0	0.1
ethnic group, nationality	1.8	0.2
language characteristics	0.2	0.0
location of unit	69.7	7.2
housing conditions	5.7	0.6
postion in family	0.5	0.1
occupation	47.1	4.9
education	17.0	1.8
physical conditions	8.2	0.8
mental conditions	3.2	0.3
time limits	116.4	12.1
other	19.0	2.0

(Total # of codes = 1303)

Table 7

"Sampling procedures"

ITEM 223	FREQUENCY	PERCENT
missing	329.0	34.2
no sampling	178.4	18.3
quota sample	2.4	0.2
simple random number	91.4	9.5
stratified random sample	36 2	3.8
area-cluster sample	13.0	1.4
multi-stage sample	245.7	25.5
other	66.2	6.9

(Total # of codes = 1019)

Table 8

"Time Dimensions"

ITEM 221	FREQUENCY	PERCENT
missing	59.0	6.1
cross-sectional	322.6	33 5
as above - with partial replication	350.0	36.5
panel study	218.6	22.7
trend study	10.0	1.0
other	1 0	0.1

(Total # of codes = 974)

Table 9

"Geographical Code"

ITEM 225	FREQUENCY	PERCENT
missing	2	0.2
local	133.5	13.9
regional	34.5	3.6
natioanl	750.5	78.0
cross-national	41.5	4.3

(Total # of codes = 985)

Table 10

"Accessiblity"

ITEM 331	FREQUENCY	PERCENT
missing	38	4.0
no access restrictions	169	17.6
no restrictions to scientific usuage	454	47.2
no publication without permission	13	1.4
no use of data without permission	149	15.4
available after special arrangement	136	14.1
other access conditions	3	0.4

(Total # of codes = 965)

Table 11

D212:04
"DIMENSIONS OF DATA number of cases"
D220:01
"TIME PERIOD begin year"

Frqncy	mis-sing	before 1950	1950-1959	1960 1969	1970 1979	1980-	Total
missing	5	3	2	20	151	88	269
1-800	9	6	11	14	70	47	157
801-2500	6	14	17	47	203	59	346
2501-20000	3	30	2	12	57	28	132
20001+	28	9	0	6	14	1	58
Total	51	62	32	99	495	223	962

Table 12

D213:02
(DIMENSIONS OF DATA number of variables)
D220:01
(TIME PERIOD begin year)

Frqncy	mis-sing	before 1950	1950-1959	1960-1969	1970-1979	1980-	Total
missing	4	5	1	43	171	120	344
1-40	34	25	12	12	45	6	134
41-100	11	17	16	33	56	26	159
101-250	1	8	2	4	154	46	215
250+	1	7	1	7	69	25	110
Total	51	62	32	99	495	223	962

Table 13

D212:04
(DIMENSIONS OF DATA number of cases)
D213:02
(DIMENSIONS OF DATA number of variables)

Frqncy	missing	1-40	41-100	101-250	250+	Total
missing	244	8	6	6	5	269
1-800	30	24	28	44	31	157
801-2500	37	39	86	128	56	346
2501-20000	23	35	31	28	15	132
20001+	10	28	8	9	3	58
Total	344	134	159	215	110	962

Table 14