6 Denmark

The Danish Data Archive

Nanna Floor Clausen
Hans Jørgen Marker

Introduction

The Danish Data Archive (DDA) has selected to concentrate its collection of data on three main categories--historical, social science and public health. Within these categories DDA offers:

- A central storage and distribution service for Danish computer readable research data;
- A distribution service for international computer readable data between Denmark and abroad.

DDA was established as a pilot project in 1973. Facilities for archiving research data have existed for more than 25 years. In 1993, DDA became an independent part of the Danish State Archives, and is administered by the Ministry of Culture. The

Nanna Floor Clausen is Archivist at the Danish Data Archives (DDA) and Hans Jørgen Marker is the Archive Director. The DDA is an independent, national data bank for researchers and students in Denmark and abroad. Since 1993, it has been affiliated with the Danish State Archives (DSA) which includes the Danish National Archives in Copenhagen; the Provincial Archives in Viborg, Aabenraa, Odense, and Copenhagen; and the Danish National Business History Archives in Aarhus.

For more information about the DDD, see: http://ddd.dda.dk. Web sites for the DDA are: http://www.dda.dk (English) and http://www.sa.dk/dda (Danish).

work is funded by Danish and common market culture- and educational resources.

The Danish Data Archive receives, processes and distributes data from universities, private researchers and research institutes--and from several private Danish institutions, producing Gallup polls. The Danish Data Archive in Odense is among the largest in Europe and contains an extensive collection of public- accessible computer readable data within the mentioned categories. In all, DDA has archived approximately 5000 data sets with a wide range of scholarly uses.

DDA is a national resource center. In addition to data dissemination in Denmark, DDA is a data intermediary for more than 20 countries. This is the result of idea exchange and cooperation with other national data archives and international archive associations. DDA works nationally and internationally to:

- Collect and improve accessibility of computer readable, scientific research results in Denmark within the mentioned categories.
- Ensure a justifiable data storage that eliminates the risk of data loss, decomposition or physical damage.
- Supply data donors with the necessary information for securing the appropriate accompanied documentation.
- Convince users as to the importance of secondary analysis with regards to the economic release of research resources and a higher research quality.
- Develop computer tools, used for searching information and transmission of archive data, better, stronger and more user friendly.

The Source Entry Project

Like other Scandinavian countries, Denmark has excellent demographic sources, which invite computerised treatment. Naturally such treatment requires that the sources exist in a machine-readable edition. In order to achieve this goal, a large data entry effort is needed. Thus, Danish electronic data processing (EDP)-based historical demographic research has made use of small subsets of the available demographically-related source material. Large-scale data entry projects have been planned, but the plans have never been realised. It has not been regarded as feasible to conduct large data entry projects with public funding as it has been done in other countries, especially Sweden. Financing through research grants has not seemed feasible either. Without public funds or research grants, it has not previously seemed possible to complete a large-scale data entry project.

Denmark has, however, many devoted and competent amateur historians, and good relations exist between the amateurs and the professionals. Among other things, this is demonstrated by Danish professional historians regularly publishing books through commercial publishers and selling large quantities. This is specific to Denmark[1].

The organisation *Databehandling i Slægtsforskningen, DIS-Danmark,* is the most important society of amateur historians using EDP. [2] As the name implies the roots of the society lies in genealogy. The abbreviation is *DIS-Danmark* because other Nordic countries also have DIS's. Beside genealogy *DIS-Danmark* also works with registration and systematisation of source entries. *DIS-Danmark* is a large society, and is growing rapidly.

On the basis of these factors it became possible to initiate aa source entry project in which the first step was a standards committee.

SAKI

The Cooperation Committee for Source Entries (SAKI) was formed in 1992 on the initiative of Elsebeth Paikin from *DIS-Danmark*. In the formation of the group a broad representation of archival and historical expertise was emphasised. The members of SAKI are Finn Andersen, *Landsarkivet for Sjælland*, Bjarne Birkbak, *Sammenslutningen af Lokalhistoriske Arkiver*, Svend-Erik Christiansen, *DIS-Danmark*, Ole Degn, *Landsarkivet for Nørrejylland*, Gunner Lind, *Københavns Universitet*, Hans Jørgen

[1] A comparison with Sweden is found in Harald Gustafson's "Dansk historisk forskning set udefra," *Humaniora* 1, 1994, p. 20.

[2] Data processing in Genealogy, *DIS-Danmark* .

Marker, *DDA*, Poul Olsen, *Rigsarkivet*, and Elsebeth Paikin, *DIS-Danmark*.

SAKI worked swiftly. During three meetings (the first in June 1992, the last in March 1993) we completed a set of recommendations for the creation of machine-readable editions of structured sources. These recommendations are called the SAKI-model. The SAKI-model is published in the SAKI-manual.[3] Based on the SAKI-model, data structures have been defined for censuses, parish registers, land registers and conscription registers. These structures are also published in the SAKI-manual.[4] In the SAKI-manual, the important distinction between source reading and source interpretation is emphasised. Consequently, the data models are given in two versions--namely, the basic model and the expanded model.

The basic model contains all the structural elements needed to mirror the contents of the source. Usually the elements in the basic model have only one field to hold the transcription. In the expanded model the elements contain further fields, which make room for interpretation. It is stressed in the manual that the fields of the basic model are to be used for transcriptions only. Thus if the source, for instance, gives a person's name as Niells Veffuer, this exact spelling will be found in the field "source name." In the expanded model, the interpretation Niels can be entered into the field "given name" and in the same way Væver can be entered into "remaining name."[5]

In the expanded model person names are understood as having three parts: "given name," "patronymic" and "remaining name." Beside these three parts of the name, the name field in the expanded model makes room for person identification, which is intended to hold a unique identifier, created when the data material is analysed. This understanding of the element name was reached after some discussion. It could be argued that "remaining name" could be many different things, in the example given for instance it is possible that it is an occupation. It could also be a family name or a mere nickname. Similarly it is

[3] The SAKI-manual has been published as *DDA-Nyt nr. 65*, 1993, pp. 7-137. The description of the SAKI-model is found on pp. 57-76.

[4] *DDA-Nyt, no. 65*, 1993, pp. 89-123

[5] Translation: Weaver

often problematic to decide whether a name that looks like a patronymic (i.e., ends on *-sen* or *-datter*) is in fact a patronymic and not a family name or even a noble name (Steensen).

The SAKI-manual gives definitions for the elements name (of persons), date, age (or birth date), occupation, sex, marital status, birth legitimacy, event, place, connection (relation or tie to another person), citation (of a place in a source) and comments. Naturally these elements can occur more than once in a structure. For example, in a child baptism, name is found at least three times as name of child, name of mother and name of father--probably also, names of godparents. Used in this way the elements describe the major part of most demographic sources.

In both models we allow the marking of uncertainty with a double question mark (??) and of gross improbability with a double exclamation mark (!!). The marks should be accompanied by an explanation in the comment field. For instance: "A boy cannot be a mother."

The basic model does not allow any normalisation or coding. These activities are only allowed in the expanded model. When normalisation or coding is performed, a separate code sheet must follow the data material.

SAKI continued its work of co-ordination and evaluation of the source entry project until 1998, when it was decided to replace it with a more formally advisory board. This is expected to be completed within the year 2000.

The Source Entry Program

The structures given in the SAKI manual can be used in any data base program. DDA only accepts semicolon-separated files in SAKI-conformant format or files from the various KIP-programs, which have been developed through the years. The first program for data entry persons who did not have a data base program or who felt insecure about how to make a data definition in the program they did possess, was KIP 1.0.[6] KIP 1.0 was based on a Paradox 4.0 for DOS run time. KIP 1.0 was only created for censuses. The next program, KIP 2.0, which appeared in 1996, included support for parish registers. KIP 2.0 was based

[6] *Kildeindtastningsprogrammet*

on Paradox 4.5 for DOS runtime. As time went by, the DOS-based programmes became increasingly problematic and thus WinKIP was made available in 1999. WinKIP so far is available only for censuses but a version for parish registers is imminent.

The DOS-based KIP programs were developed by Elsebeth Paikin from *DIS-Danmark*, while Otto Thygesen is making the Windows version. KIP provides a form-based data entry, which reflects the structure of the source being entered. Furthermore KIP supports (and actually requires) entry of the necessary documentation which must follow any data material, such as original source, location of the original source, data entry person, proof reader, general comments on the encoding etc. Finally, KIP includes tools for making searches and reports on the basis of the data material. Thus KIP is not only useful as a data entry tool; it can also be used as a search tool on KIP files, which the user has acquired from the DDA. Everyone who receives KIP is required to deposit all data entered in the DDA for free redistribution.

Coordination

In the source entry project, coordination is crucial. Naturally, it is extremely important to avoid wasting resources. The co-ordination is done by the DDA, which has assigned two full-time employees to the task of guidance of the data entry persons and co-ordination of data entry.

Before data entry commences, the data entry person needs to go to KOKI or the DDA and receive a data entry number. The data entry numbers are unique and are used to distinguish the resulting data files.

DDA ensures that there is no duplication of effort in the source entry project. If two people are interested in the same source, DDA establishes contact between the two so that one can make the data entry and the other can function as proofreader.

Data Entry

Data collection began in 1993 and the first thirty data materials were received by the DDA before the end of that year. Over the years, data has been arriving at an increasing pace. As of May 2000 we have received 4789 files containing 2902591 person records.

Since October 1994, the DDA has been giving semi-annual courses for source entry persons to both raise the awareness about the important issues involved in making source editions and to stimulate the interest among the amateur historians involved in the project.

To further accommodate data entry persons, we made an agreement with *Rigsarkivet* that the a complete photocopy of the censuses from 1787 to 1860 would be transferred to the DDA. These are lent, at the discretion of the DDA, to well-deserving data entry persons as base material for their data entry work. Furthermore, DDA has many censuses on microfiche which are lent on similar terms. We have also received a set of censuses for Jutland and Nørrejylland from *Landsarkivet* which are lent on slightly more liberal terms. Finally we have provisions for making paper copies from microfiche at cost.

Distribution

Data materials originating from the source entry project are archived in the DDA in the same way as data materials from other sources. This means that the usual procedure for data delivery is used for the KIP-materials as well. Anyone who wants a KIP-material can fill in a data set requisition form. On that form the user promises to abide by the conditions regulating the secondary use of data from the DDA. Having signed the form, the secondary user mails it to DDA with a brief description of the use they want to make of the material.

Users do not pay for data materials, only for diskettes used to ship the data. Because the data are delivered to users free of charge, it is extremely important that users only ask for data will use. On the other hand, we do not want users to refrain from requesting data that they will actually use.

A more common way to get the KIP data materials is on CD-ROM. CD-ROM’s have been created annually since 1995 and are sold in increasing numbers. In autumn 2000 the range of CD-ROM editions will be expanded to fulfil more user needs. Furthermore, all census data are accessible through the Danish Demographic Database.

Legal Considerations

For KIP materials as for any data materials in the DDA the ownership or copyright rests with the depositor. The copyright is not ceded to the DDA at the deposit. The DDA receives certain limited rights regarding the redistribution of the material. For the KIP-materials these rights are free distribution for non-commercial use.

No one obtains data from the DDA with the right to redistribute it, because the DDA is not entitled to pass that right on to a third party. Furthermore, the DDA guarantees the integrity of the data, which means that the third party can rest assured that the data are the same as the ones archived in the DDA to begin with and have not been corrupted during distribution. Data produced with the aid of KIP always contain information on identifiable persons. This means that the KIP materials always fall into a category governed by the legislation on privacy. In Denmark, a wing of government called *Registertilsynet*, the register supervision, enforces this legislation. Presently KIP does not, however, provide set-ups for sources to which a dispensation is needed to get access in the archives. Thus, the data produced by KIP are not confidential.

If in the future the DDA should obtain data concerning identifiable individuals which contain information younger than the usual age limits, these data will be handled in the same way as that type of information is dealt with in the rest of *Statens Arkiver*. Information older than eighty years will be freely accessible and it will be possible on dispensation to gain access to materials of an age of between fifty and eighty years. The dispensation practice in the rest of *Statens Arkiver* is, by the way, that access is usually not given to complete sources. Under that practice dispensation will not usually be relevant for the KIP data. Obtaining data subsets is not a service that the DDA offers free of charge.

The Danish Demographic Database

In 1996, the Danish Demographic Database (DDD) was initiated as a project coordinated and led by the Danish Data Archive but in close cooperation with the Emigration Archives and the Filming Centre in the Danish National Archives. The intention was to provide on-line access to transcribed sources available and to make it possible to go from a record to the scanned source. In this way users could both see the original source and check the information in the databases. Kulturnet Danmark funded the project in the first year but since then it has been left to the Danish Data Archive to fund.

The DDD comprises four different databases: census records from 1787–1916, Copenhagen Police Records of Emigration, Immigration records, and finally a small database of *dannebrogsmænd*--men who have been given a special decoration. On the English site, only the census records and the emigration records are available. The largest and still expanding database is the one containing census records. As described above the project of transcribing the copies of the original census records is an on-going process as well as the work of archiving and disseminating the data. The database now comprises approximately three million records; when all census records (through 1916) have been transcribed, it will contain approximately 21 million records.

The Census Records 1787–1916

In order to make an online query in the census records, all the records were put into one MSSQL database table. This meant that we constructed a database with as many fields as the questions in each census that was taken. While this decision is debatable, so far we have had few problems with this structure. The only change that has been made so far is to divide the records according to counties. It was reasoned that as the DDD became better known and consequently had more users and the number of records increased, the response time also increased. The disadvantage of splitting the materials by counties is that users can no longer conduct a query across the whole country. The advantage is that response time is now rather low and users get a result! So for the time being, we have no intention of changing this.

As the data originates in the work of volunteers, dissemination must also reflect this work. The fact that the number of data materials in the database is increasing is also reflected in the contents of the DDD site. The DDA cannot force people to transcribe census records so we are very grateful for all the contributions we get. On the other hand we have some rules regarding the quality of the material and rights to them. We only accept data if they cover a whole parish. The DDA promises to preserve the data material and to disseminate it freely as long as the users don't use the data material in a commercial way.

The DDA gives a high priority to providing users and volunteers with as much information about the project as possible. This is done by putting on the Web a survey of all the data materials in the DDA, that has been transcribed or that somebody is transcribing but has not yet finished. In this way it is possible for users to see which data materials they can expect to be added to the database. This survey is now so big that users have to download a program to search it. Another important survey contains information about parishes in the database. This list is extracted from the database and contains information about the parish, district, county, year of census and finally the unique source entry number. The survey is also supplied with information whether the data material is added in the last upload or not. This makes it easier for users to discover whether their own data material is now available on the web. Before DDA introduced the Internet solution, the volunteers were satisfied when DDA recognised having received their data material and had given it a so-called DDA-number. Now the volunteers are not content until their data material is available on-line.

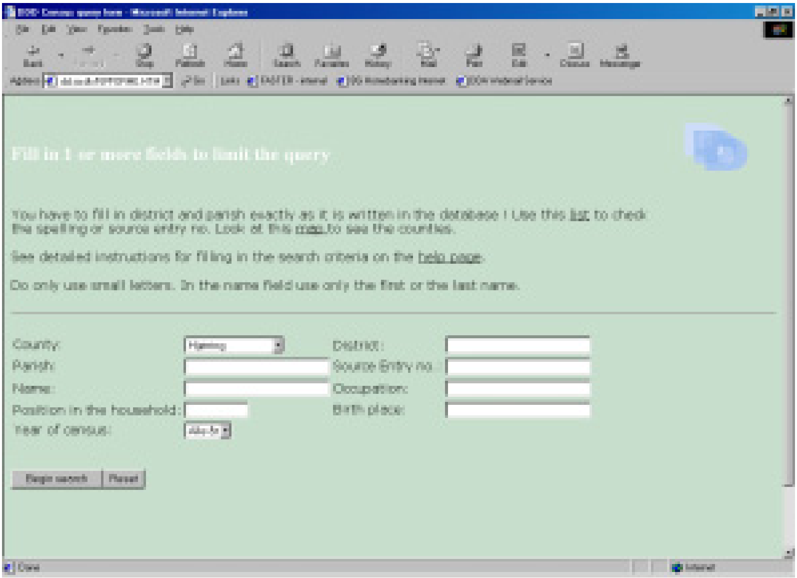
In order to inspire volunteers to work even harder two lists are produced every month. The first one contains a list of all the names of the volunteers ordered by the number of records they have transcribed. This list is read carefully and there is a strong competition regarding rankings on the list. The second list contains the names of the persons who have done the proof reading of the data materials. It is also ordered according to the number of records each person has proofed.

Interest in genealogy varies across the country and is reflected in the areas covered in the DDD. In order to illustrate this problem and to increase the coverage, a map for each of the census years is posted on the Web. This map shows the borders of each parish. Colour indicates whether the census from that parish for the specific year has been transcribed or is being transcribed or if nobody is working with this parish. In this way it becomes evident that few are interested in doing transcription on Zeeland and that some parts of Jutland and Funen are well covered. Volunteers who provide the information also maintain the map. It should be noted that the map is not in a GIS format.

The DDD site also displays a graph that illustrates the development of the Source Entry Project. DDA also produces statistics showing coverage for each census year. Some years are more popular than others. The years 1801, 1845 and 1787 are the most popular. Approximately 60 percent of the 1801 census has now been transcribed, totalling approximately 555,000 records. 1801 is special for two reasons: Copenhagen has been transcribed, which constitutes a large fraction of the total. Additionally, we have launched a campaign for having the whole census transcribed before February 2001, exactly 200 years after it was taken. We expect the percentage of coverage to grow rapidly in the next six months. The volunteers are enthusiastic. For both the researchers and genealogists, it will be a great achievement. If we reach this goal, both Norway and Denmark will have this census on the web for on-line queries. A project has just begun which will transcribe the census records from the southern part of Jutland from 1803. This census had the same questions as the 1801 census. The 1845 census is interesting because it is the first census that asks for place of birth. About 47 percent of the 1845 census has been transcribed, yielding approximately 630,000 records. The 1787 census is especially interesting because it is the oldest and has only a few questions or fields; 44 percent of this census has now been transcribed, equivalent to 370,000 records.

The most visited and most interesting page is the search page (see Figure 6-1). The user must choose the county to search and only one other field. A request of this type yields a large result that can be difficult to deal with. More often, users fill in a name and a few other fields, such as place of birth or occupation. The problem with using the field place of birth arises from the way place names were spelled. Because of spelling variations, the birthplace field is difficult to use in a search. A result from a database only shows the fields relating to the specific census but it does show all the fields in the census. The user is also informed about the number of hits.

Searching a database yields a different result from the census protocol. In the census protocol all the records listed can be seen. One easily discerns relationships among the persons even though their last names are different. In a database, a result is only obtained for exactly the name entered in the search interface. In order to deal with this problem, a link has been made from the person found to the rest of the persons at this address. Sometimes this link fails because the information in the sources is inadequate. The persons are shown in the order they appear in the census and with the most relevant information. Volunteers are instructed to be very careful when they transcribe the census records. It is important when "do" (ditto) is written in the source to add information about what "do" is referring to! Otherwise, "do" would be a very large parish and not of much use in the query.



Fill in 1 or more fields to limit the query

You have to fill in district and parish exactly as it is written in the database ! Use this list to check the spelling or source entry no. Look at this map to see the counties.

See detailed instructions for filling in the search criteria on the help page.

Do only use small letters. In the name field use only the first or the last name.

County: Hjørring
District:
Parish:
Source Entry no.:
Name:
Occupation:
Position in the household:
Birth place:
Year of census:

Begin search | Reset

Figure 6-1 Census query form

From the records found a link is made to the documentation of the data material. Each volunteer is required to fill in meta-documentation about the source used for transcribing and when the transcribing was done. Information should also be given about proof reading. If users find errors in the records notice should be given to the person who did the transcribing and not to DDA. This person can then check the information and if an error is found then the correction is made and sent to DDA, where the entry is replaced.

For several years it has been possible to download data materials as zipped .csv files. This size has been popular, with more than 100,000 downloads per month. But due to abuse and overload of the server, this service will soon be replaced by CD-ROMs.

Copenhagen Police Records of Emigration

Beginning in 1868, the Copenhagen police kept a record of all emigrants who purchased tickets to the New World from an approved Danish emigration agent. These records contain personal data about each emigrant: age, last residence, occupation and, not least, the date and destination of emigration.

This database holds all the records from the period 1868–1903. The database contains about 300,000 records and comprises persons who emigrated directly from Copenhagen or indirectly from Hamburg. We are still hoping that the records for the period after 1903 will be added.

This database is especially popular, as ancestors search this database for their emigrant forebears. The query interface allows the users to search in most of the fields in the database. It is now no longer possible to look in the scanned police records as the query result was a bit confusing, with the same record shown several time. This feature was not much used anyway.

The Immigration Database

The Immigration database is maintained by the Immigration Museum, which continuously upgrades the content of this database. As mentioned above, it is only available in Danish. The intention is to find information about all persons who immigrated to Denmark. The sources used are the laws giving permission to stay in the country, police records for persons who were given the right to work in the country or records of persons who were expelled from the country.

In the query form you can search by name, place and occupation. All fields from the database are displayed in the search result. There is more information about these datasets on the web at:

DDA http://ddd.dda.dk
DDA: http://www.dda.dk (English) and
http://www.sa.dk/dda (Danish).