Data Processing of Swedish Parish Records

When the Demographic Data Base was established in 1973, it was one of the very first large-scale incentives to build population databases for research from historical parish records.

It is not a coincidence that Swedish researchers were pioneers in this field. The Swedish records are unique in detail, quality and coverage, and the long, continuous time spans offer almost unparalleled possibilities for longitudinal studies. Church registers including the entire population have been kept since the late 17th century, and until the 1990s they also served as the official system of national registration. This was possible as a consequence of a remarkable religious conformity expressed in a national church system, which for a long time included almost the entire population. Although the church and state were formally separated by the year 2000, the majority of the population (64,6 %) are still members of the Lutheran Church of Sweden.

The sources

The Swedish parish registers were not only kept for ecclesiastical purposes. They were also intended to control and keep track of the population. From 1749 and onwards they also provided the basis for population statistics, since they included the whole de jure population and were kept in a consistent form by the local minister. The parish registration system consists of five major sources:

- Birth and baptism registers
- Marriage registers
- Death and burial registers
- Migration registers
- Longitudinal catechetical registers

What makes the Swedish system particularly distinctive is the presence of the last two registers. The 19th century migration registers accounts for people's movements between different parishes, which is rather unique in an international perspective. Even more valuable are the longitudinal catechetical registers, which is only found in Sweden and Finland. It is a dynamic source where all individuals and households in a parish can be followed over time. It accounts for demographic events, along with detailed information about for example occupation, health, household position and movements in and out of different households. People were registered in household groups and each volume covers a period of 5-10 years. Over that time the registers were continuously updated. When an event took place, the local minister entered it into the register. This way we can get a good grasp of demographic events, including dates, also when it is missing in the vital registers. The registers were also updated at the annual catechetical examination meetings, when the households were examined in Christian knowledge and reading abilities. The following information is generally found in the longitudinal registers:

- Place of residence, farm, village etc.
- Household composition including servants and lodgers
- Occupation
- Household position
- Date and place of birth, marriage and death

- Marks
- Information about migration and movements
- Disciplinary notes
- Notes about delinquency and disabilities
- Vaccination against smallpox
- Reference to military conscription

These longitudinal registers are, for evident reasons, fundamental for the possibility to follow individuals and families through their life span and over generations.

The data entry process

All DDB databases are constructed according to established rules and international standards to ensure the demands of a high quality research infrastructure. Data is entered according to consistent principles and stored in relational databases according to the IBM DB2® system. The database structure including all tables and variables is well documented and described in detail, in order to ensure longevity, consistency, security and quality.

Documentation, quality and consistency are key features of the different stages involved in database building at the DDB. Building research databases from parish records using both static and dynamic sources is a complex process with numerous pitfalls to consider. If we are aiming for datasets characterised by high quality, consistency and comparability, these features have to be established and controlled already during the process of data collection.

At the DDB, the first part of the database building process consists of a thorough documentation of the sources that are about to be processed. Usually we are working with a complete digitisation of all sources in a parish, one parish at the time. The preparatory work starts with an examination and overview of the sources, with particular attention to the quality of the source and the occurrence of the different variables. Missing pages are noted and the time scope of each volume is controlled. Another aim is to find and address potential problems in advance, to avoid disturbances and delays during the digitisation process.

Data entry is carried out with supporting software developed by the DDB, according to strictly controlled rules. Some variables are coded already during the process, while others, such as occupation and cause of death are coded and standardised at a later stage. The digitisation is surveyed by built-in quality controls and carried out by experienced staff. At set intervals 10 percent samples are scrutinized, to ensure quality and consistency. The software offers standard templates for birth- and baptism registers, migration registers, catechetical registers, death registers, and marriage registers.

In the context of a database-building process, access to supporting software plays a central role. If correctly designed, it provides users with a technological platform that supports and forms an integral part of the database-building process, making it systematically organised as well as accurately documented. To be truly useful, such a platform should also be easy to access for researchers and research infrastructures, adaptable to new technology as well as assigned for long-term maintenance. At present a new data production system, PERSONA, is being developed by the DDB, in collaboration with other organisations using the same type of sources. This new software platform will be ready to use in 2016, and will be freely available for download.

The linkage process

The aim of the linkage process is to create complete life biographies and reliable genealogies out of digitized individual records.

The linkage process at DDB consists of three different linkage steps: *Record linkage*, i.e. linking together all separate records relating to an individual; *Relation linkage*, i.e. linking together parents and children as well as partners; and *Region linkage*, i.e linking larger regions together making it possible to trace migrants in time and space. These procedures are performed in a continuous series of steps, each of which is dependent on the previous one. This also involves consecutive quality controls.

In order to build life biographies for individuals, all records corresponding to the same individual have to be interconnected. Furthermore, to define kinship and subsequently extract genealogies, family relationships have to be defined according to carefully applied rules. Finally, to be able to follow migrating individuals, the process of linking individuals who have presence in several adjacent parishes has to be performed. Linkage is comparatively straightforward if data contain civil registration numbers, which is the case after 1947. For older data, linkage has to be based on other matching variables. This process is considerably more complex and to achieve high accuracy, several aspects have to be considered.

During record linkage, individual records are merged and each individual is assigned a unique identifying number. Individuals are linked together on matching variables such as name, date of birth, gender and parish of birth. Matching information about residence and relatives can also be used in the verification process. Most records are linked together by automated linkage, using Core Link, a custom-made software system developed at the DDB. This is followed by a semi-automated linkage, which takes care of the information that for several reasons could not be processed by the software. At present 97 percent of the records are linked together by automated linkage and with the subsequent semi-automated linkage the linkage rate reaches almost 100 percent.

When record linkage is complete, the procedure continues with the linkage of relationships. This step involves a semi-automatic linkage of children to their parents, and the pairing of partners. For each relationship, a link stating the kind of relationship is saved in the database.

The third and final step in the linkage process is a region linkage, which involves the merging of individuals who have migrated between different parishes. This is done through an automated linkage executed by a database administrator, followed by a semi- automatic process to fulfil the requirements of linkage rate and accuracy. The region linkage module uses information about residence, migration and relatives in addition to the usual matching variables such as name and date of birth.

Making data available for research

Before data can be made available for research some final operations are required. Linked data is organized into the database model and all variables have to be verified and controlled. Two databases are created, one source-oriented and one user-oriented database. In the source-oriented database the digitized data is stored in a well-structured way, facilitating updates and ensuring data consistency. The user-oriented database is a read-only version of the source-oriented database, designed to fulfil the particular demands of the users of the

database (researchers and system developers). In addition to complete information from the source database, it also includes tables, compiled from source information, which describe events and life histories concerning individuals and families. This significantly increases the accessibility of the data and facilitates data retrievals, without compromising the unconditional needs for adherence to the source, order and structure in the database. The principles applied for data consistency are surveyed by carefully applied maintenance rules in which regularly validation, monitoring and quality control are important features.