# Project report: A Historical Population Register for Norway



## Executive summary

Several countries are building and extending contemporary and historical longitudinal population registers, for administrative, statistical and research purposes, to keep track of their populations and to research empirical and theoretical queries using continuous collective biographies. Increasingly, contemporary databases and methodologies with longitudinal population registers are replacing censuses and vital event registers. This report details such developments for Norway, where as a complement to the Central Population Register (CPR) we are currently building a Historical Population Register (HPR). Except for Iceland, this will be the world's first national HPR, and it will be open to all kinds of bona fide research. The aim of the project is to include as many of the 9.7 million residents of Norway who lived here between 1801 and 1964 as possible. This report first reviews population data collection methods in Norway and then describes plans for the construction of the HPR for Norway. It concludes with an examination of some research possibilities with longitudinal population data.

The national Central Population Register covers the period from 1964. The National Archivist will over the next three years build a register from 1801 to 1815 with resources present in the National and Regional Archives. There is ongoing work in the universities, among local historians and among genealogists to cover the remaining one and a half century between 1815

and 1964. Thus, it is only a matter of time before Norway has a working HPR, but this time period may become exceedingly long if the work on this is not coordinated and conducted in a systematic and efficient way. The group behind this report suggests a data structure and methods that would succeed in building it efficiently during a limited period, given funding of the infrastructure during the next decade. The alternative is a project funded on a shoestring which would waste resources in several project areas, maybe over the entire 21st century.

Inefficient use of resources is especially the case in the area of digital transcription from the source material. The traditional method still in use is data entry character by character. Because the sources now are available as scanned images, the HPR pre-project has investigated graphical methods which will rationalize transcription significantly – one of these methods has already been implemented and tested successfully. A full HPR project will thus double or triple the productivity of the professional and voluntary transcription work undertaken every year.

The other main challenge when building a population register is linking the records relevant to the same individual from different sources. Before the era of modern ID numbers, identifiers such as names and age are fuzzy because they are neither unique nor consistent over time. By using SQL string functions that can dynamically compare the similarity of names while linking, and employing data about couples and groups of interrelated persons, we have increased the rate of correct links in the sources significantly. In cooperation with our partners at the University of Minnesota national record linkage has been undertaken and the result is available to researchers. Our prototype HPR-wiki will give local historians and genealogists the opportunity to help us link remaining records via the Internet, based on their specific knowledge. Our experiments indicate that record linkage rates from 80 to 90 % will be the norm.

The third main challenge is the handling of migration. Missing the migrants has rendered the local longitudinal studies undertaken during the previous half century unrepresentative of the population. A national HPR will enable the inclusion in longitudinal study of most domestic migrants. Most out-migrants from Norway were emigrants who can be found in either American or Norwegian registers, most of which have now been transcribed. Immigrants to Norway can usually be identified in the censuses. We are presently discussing with our European and US partners how to treat some problematic minorities among international migrants, especially those moving repeatedly across the Atlantic.

The HPR will be organized as three separate databases that are linked on the individual level via special primary keys. A system of historical source entry and consistent personal id numbers is being implemented in the National Archives. The oldest part of the register, from 1801 to November 1920, will be open to the public, the second part (December 1920 through 1964) will also be kept in the National Archives, but access will only be given to bona fide researchers upon application. The third part will be the Central Population Register as kept by Statistics Norway, and researcher will normally be given access to anonymized data only and upon application.

The need for the longitudinal HPR extending over more than two centuries has been documented in a series of research papers and in support letters from researchers in a wide range of scholarly fields: medicine, demography, history, economics, social sciences. Research topics range from the tracing of hereditary diseases to ethnic differentials with respect to cohabitation.

# Introduction

A population register is longitudinal in the sense that it maintains a continuously updated overview of the population in an administrative geographic area. The register may be national, regional, or local. In the latter case, records about migrants are linked together inside a municipality, but do not necessarily link when records cross municipalities. The population register is often based on censuses and lists of vital events, linking events together at the individual level. In addition, the register will record migration, so that the composition and whereabouts of the population can be documented more or less continuously. Church registers, on the other hand, are limited to recording vital events (births, marriages, deaths) and censuses only record cross-sectional overviews of the population, usually at decennial intervals.

Modern population registers are *in principle* updated in real time so that migration and vital events are mirrored in the database as soon as they are reported, which may in practice be months after the event happened. Even more so, the continuous nature of historical population registers must not be interpreted literally, since they are based on non-continuous sources such as censuses or ministerial records. In regions where a significant part of the population migrates to work or routinely lives in different localities during the year, the very definition of permanent address or domicile becomes problematic (Thorvaldsen 2006). The core variables in the register are name, address, sex, birth year or birth date, and marital status; birth place is frequently recorded, as well. A population register may or may not contain information about other characteristics, such as occupation, poor relief, education, income, and household relationship. These optional variables and especially sensitive information, such as criminal records, are typically kept in separate registers which can be linked to the population registry via stable and unique id numbers.

Nordic population registers have their origin in the seventeenth century. For Sweden and Finland, a church law of 1686 required priests to keep catechetical protocols with topographic overviews of the population in their respective parishes. From 1749 on, when Tabellverket (the Swedish Table Works) demanded updated aggregates for a host of demographic variables, the longitudinal system of catechismal records has been updated annually in the whole country (Nilsdotter Jeub 1993; Sköld 2001). The Norwegian Central Population Register was introduced in 1964, superseding local registers with their longer historical reach. In 1906, the director of the capital's statistical office, Gustaf Amneus, had launched the Population Register for Oslo. He based the register on older European examples, especially on the City Registrar Archive that was operated in Stockholm from 1876 to 1928. He drew also on the Italian, Belgian and Dutch registers from the middle of the nineteenth century (Thorvaldsen 1998; Janssens 1993). His rationale for constructing the register was that censuses and ministerial records could not keep track of the rapidly increasing and geographically mobile populations in big urban centers. Liberal population policies were a thing of the past; new legislation regulating paupers, aliens and vagrants presupposed a population registry system for tracing individuals across time and space. When the population of the Norwegian capital had tripled since 1875, it became clear that many in-migrants continued to pay taxes to their municipalities of origin, rather than to Oslo.

The system Amneus planned in 1905 was based on cards rather than protocols. Because of this innovation, Amneus has been called the "father of the modern population registry." The registry system included four card collections: a main system for families, and three secondary indexes for addresses, in-migrants, and out-migrants. The cards system simplified tracking demographic changes. For instance, when someone married, the two individual cards could easily be brought together (Thorvaldsen 2008). Following Amneus's example in Oslo, officials in Bergen based their population register on the 1912 municipal

census, with retrospective questions on how long immigrants had lived in the city. Municipal population registers spread to most urban areas between the First and Second World Wars.

After Germany occupied Norway in April 1940, the Nazis attempted to control population mobility more closely than the native government had. A population register was a necessary first step in control. Accordingly, in February 1942 they launched an initiative to create a central national population register, written into law beginning on 1 March 1943. Simultaneously, vital registration was transferred from the churches to the Nazi-controlled central authority. Conflicts between the clergy and the occupying forces, along with "go-slow" resistance within Statistics Norway undermined the Nazi *centralized* register. However, in 1946 the reinstated Norwegian authorities made *municipal* population registers obligatory by law.

The current unified Norwegian national register was based on the 1960 census. It took four years to build this computerized register and required the introduction of unique social security-like national identification numbers. Thus, the current Central Population Register has full coverage back to 1964. Statistics Norway, which was in charge of the register until 1991, still use it to produce more continuous population statistics than could be done with censuses and vital registration.

The last questionnaire-based Norwegian national census was taken in 2001. Denmark had stopped taking censuses in 1970, Finland in 1980. The 2001 census was done for two reasons: First, while all persons had had unique identification numbers since 1964, domiciles could not be identified with the same precision. The main objective of the forms Statistics Norway sent out or made available on the Internet in 2001 was to provide each flat or housing unit with a unique address. Second, some information about the individuals, such as specific occupation data about their employers, was not adequately defined in auxiliary registers that could be linked to the Central Population Register. Another important piece of information that is difficult to gather in population registers concerns educational attainment, which must take into account the increasing popularity of studying abroad.

## The Need for Longitudinal Data in Research

While hundreds of interesting papers, articles, and dissertations have used cross-sectional or linked census data from the United States and other NAPP countries, longitudinal data enables the creation of more dynamic pictures of household structures, migration, mortality, and fertility by augmenting the cross-sectional overviews of census data with continuity details from population registers about vital events, geographic mobility, and property transfers. Combining the various sources in one database makes it possible to see how these phenomena are interrelated in time. For example, we can see whether or not a woman became pregnant again soon after her baby died, or whether a mother moved into her daughter's household before or after becoming a widow. New studies using the Demographic Data Base at the University of Umeå have found demographic characteristics running in families across generations, while the new field of epigenetics has shown how children were affected by crises that their parents experienced a generation earlier because of defects in their gene systems. This, together with more classical questions about hereditary diseases, makes the HPR a valuable instrument also for contemporary medical research.

The timing of events shown in several sources related to the same persons can highlight complex cultural behaviors such as marriage patterns. A study based on the database from Rendalen parish in southeastern Norway used censuses, burial, marriage, and property records to examine the extent to which the timing of marriages, and the choice of marriage partner was made by the bride and groom themselves, rather than by their parents. While in the early

eighteenth century most marriages and farm property transfers were contracted before the parents passed away, the late eighteenth and early nineteenth centuries saw a development in which sons who married after the death of their fathers chose their own spouse. It appears that the liberal ideas of the French Revolution extended into the private sphere of matchmaking, easing the pressure on the young to marry a partner of the parents' choosing (Bull 2005; 2006).

In recent years numerous national and regional studies of family composition in the late nineteenth century were based on population samples, but few incorporated community-level economic measures (Sogner 1990; Gunnlaugsson and Garðarsdóttir 1996; Ruggles 2000). In family and aging studies, the main research question is identifying factors that motivate parents and their grown children to live in the same household. With a cross-sectional source such as a census, we can study differential characteristics—such as gender, age, and marital status—of *aged people* who do or do not live with their children. But differential characteristics of the relevant adult *offspring* cannot be analyzed, unless all children are linked to their aged parents. With links to other sources, we can study differences between children who did or did not co-reside with their aged parents. Thus, research based on longitudinal data will contribute more complete answers to the long-lasting debate about the question of the extended family structure (Jåstad 2009). Other examples of the pivotal role kinship plays in social processes include migration, chain migration, and infant and child care outlined later in this report.

## Existing Retrospective Population Registers

Historians build longitudinal population registers by combining other types of nominative source material which survive from earlier periods. The administrative history of the Nordic countries up to the start of the nineteenth century produced two different types of nominative source material, one in the eastern and one in the western part of the region. Norway kept the Danish system of census-taking and nominative registration of vital events even after the union was split in 1814. Likewise, priests in Sweden and Finland continued to build their catechetical registers even after Finland was ceded to Russia in 1809 (Thorvaldsen 1998, 2007). These differences mean that historians in the eastern parts of Fenno-Scandinavia can base their computerized, nominative databases on longitudinal church records which were linked by the priests in each parish. In Denmark and Norway, records belonging to the same persons must be brought together through automatic or manual record linkage after transcription.

### Family Reconstitution-Type Databases

About twenty Norwegian parishes were linked over two generations through manual family reconstitution. Because nearly everyone belonged to the established church, complete families could be reconstituted for the non-migrant part of the population. Even though families moving across parish borders during the fertile period could not be used in the analysis of fertility, and the mortality of out-migrants could not be satisfactorily computed, it was still possible to follow parishioners from the fjord community in Western Norway to prairie settlements in the mid-western United States and to compare their demographic behavior on both sides of the Atlantic (Sunde 2001).

In one case the family reconstitution has been entered into a relational database and augmented with complementary source material. Sogner manually reconstituted the parish of Rendalen northeast of Oslo for the period 1733 to 1828 (Sogner 1979). In the 1990s the information on these cards was turned into a relational structure, allowing information from censuses to be imported from the NAPP-related databases. A person table was constructed, with pointers linking information from individual sources. Information on family structure was transferred into the database by linking to wedding records. The addition of records from

emigration and taxation lists, as well as property protocols and local community histories, expanded the research uses of the database beyond pure demographic research. More than 90 percent of persons in the census were identified in more than one source. These high linkage rates are due both to the low level of migration in this peripheral valley and to diligent record linkage work. In addition to research on marital customs mentioned previously, the database also has been used to study infant mortality and mental illnesses. The research potential of this database is currently being expanded anew by extending it with the 1910 census and ministerial records for the period to 1928. A similar database exists for the parish of Etne east of Bergen.

Longitudinal data is available for one additional municipality in south-eastern Norway, the parish of Asker (1801-1875) bordering the southwest side of Oslo. This parish was more populous than Rendalen with 4,600 inhabitants in 1801 and 13,700 people by 1900. Census records from 1815, 1825 and 1835 could be linked with baptismal, marital, and burial records, as well as farm taxation lists from 1826, 1838, 1866 and 1886. Name standardization and the use of an interactive record linkage program increased the speed of record linkage over the old manual system (Nygaard 1985 & 1992, Fure 2000). By using keyboard commands, Fure greatly increased the efficiency of linking the sorted records. Once married couples and their children were linked, the likelihood increased that any remaining single individuals could also be identified. The value of the nominative censuses became clear when 95 percent of the population of the entire period was identified in either the 1825 or 1835 censuses.

The dataset has been used to study naming traditions and infant mortality. Research using these longitudinal data has shown that infants born to mothers who were themselves born during the crises years of the Napoleonic wars tended to experience higher infant mortality. It appears that these mothers were affected by early life experiences in ways that also influenced the survival chances of the next generation (Fure 2002). This multi-generational finding invites interpretation in the expanding field of epigenetics.

Norwegian municipalities have a long tradition of publishing community histories. In rural areas, these municipal histories often contain genealogies and histories of individual farms extending as far back as the source materials allow, usually back into the sixteenth or seventeenth century. The current state of the art is the program package *Busetnadssoge [Community history],* which contains modules for linking nominative sources with longitudinal genealogies (cf Kjelland below). The geographic information is pivotal, and it is organized in hierarchical levels from the municipality to farms. By importing these community databases, researchers can access longitudinal data for six more municipalities over a period of more than three centuries.

## The Population Register of Norway in 1814

Inspired by the plans for the Historical Population Register and the upcoming constitutional anniversary, the National Archivist has launched a project to establish a register of Norway's population by 1814. At this turning-point in our history an oath to the constitution was sworn in all local communities, and 112 representatives were elected or appointed to the constitutional assembly which presented our current constitution on 17th May 1814. As part of the celebration of the 200th anniversary in 2014 a longitudinal database following the development of the population who performed these revolutionary acts will be presented. This database will form the oldest part of the Historical Population Register for Norway, from 1801 to 1815.

Since the census taken in 1815 was mostly statistical and with only a few nominative lists, the Population-in-1814 register must be based on the full count and nominative 1801 census. Its 879 020 records will be complemented by the virtually complete set of national

church records with 374602 baptisms, 102435 weddings and 334618 burials for the period 1801 to 1815. (Including both census years, according to *Historical statistics* from Statistics Norway.) The 1801 census was transcribed and encoded at the University of Bergen around 1970 and is now in the process of being integrated into standardized formats in the North Atlantic Population Project (NAPP) at the Minnesota Population Center. The constructed variables in the NAPP version contain pointer variables which make specific the relations between spouses and between children and their parents in the census. By adding the weddings for the next 15 years and the children born to the parents present in the 1801 census or who later married, virtually all Norwegians will be represented in the database. Subtracting the persons listed in the burial registers will be a challenge because of the lack of relational information in these lists. Since non-conformism was virtually non-existing, however, nearly all vital events are listed, and there are few lacunae in the ministerial protocols. Thus, there is reason to believe that one result of the project is a better aggregate estimate of Norway's population at the time than what can be found in the 1815 census, especially for some localities and regions. This is considered the most inaccurate full count census taken in Norway, with an undercount of at least two percent.

Automatic record linkage will be undertaken with the fuzzy sql functions developed during the HPR pre-project and described in the paragraphs on linkage in this report. In addition manual linking will corroborate and extend the links using the HPR-wiki over the Internet (cf below). The Norwegian Computing Center will hopefully be in a position to adapt this software especially for the 1814 population project. Main responsibility for this partial HPR project will be with the National Regional Archives. The transcription of the church books will be undertaken by its Digital Archive transcription groups and the Norwegian Historical Data Centre at the University of Tromsø. Other partners are the DIS-Norway (Association for Computers in Genealogy), the Norwegian Computing Centre (NR) and the Norwegian Institute of Local History**.**

While Norwegian social and population history is relatively well researched for the later part of the 19$^{th}$ century, the first couple of decades are clearly under-studied. These were years of dramatic developments and hardships, both very rich years of timber exports, hunger crises during blockades and active participation in the Napoleonic Wars on the side of France and Denmark. Three of the years were with zero or negative population development. Among important research themes in demography and social history can be mentioned the effects of starvation viewed from the new perspective of epigenetics and the recruitment of a new class of businessmen and civil servants.

# Inventory of source material (Hans Hosar and Lars Nygaard)

The mainstay of the source material for input-data to the HPR are 1) the *public censuses 1801-1950,* and 2) the *parish registers* (kirkebøker), in principle from the date the oldest registered person in the census of 1801 was born (about 1700) up to 1960. *Emigration- and passport ledgers* are of vital interest among supplementary sources. Additional supplementary sources are in abundance, but to a limited extent available in transcribed and digitized formats. Some will in the future be added ad hoc by research projects for specific studies. The following is a summed-up inventory of the censuses, parish registers and emigration lists, stating to what extent these sources so far have been digitally transcribed.

**Censuses**

*Nationwide, urban and rural*

- The full count censuses of 1801, 1865, 1900 and 1910 have been transcribed.

- 1875: 1/3 of this census has been available in a transcribed and digitized form for a number of years. The remaining material is currently being transcribed and by May 2011 altogether 2/3 of this census has been processed.

- 1891: approximately 60% has been scanned to image files (May 2011), and the remaining will be finished within a reasonable span of time. However, little of this has been transcribed so far. Altogether 16 local communities are available from the Digital Archive (DA). In addition, the community of Asker is processed at the Norwegian Historical Datacentre (RHD), the University of Tromsø, as a test case for the HPR. Prototype software for partly automating the transcription of this (and the 1920) census has been tested successfully.

- 1920, 1930, 1946, 1950: Practically nothing has so far been scanned nor transcribed. As a test case for the HPR the censuses 1920-1950 for the municipality of Rendalen have been scanned, transcribed and linked.

*Nationwide, urban*

- The urban census of 1870: To some extent microfilmed and scanned, but not on a nationwide basis. Only the city of Lillehammer is available in digitally transcribed form.

- The urban census of 1885: A great deal of the material has been microfilmed and scanned. So far the lists from 43 out of a total of 61 towns and cities (*ladesteder* og *kjøpsteder*) have been transcribed and made available on the DA. The 43 transcribed urban communities are all situated in Southern Norway. Since this census is important for following urban migrants, transcribing the remainder will be prioritized for the HPR.

*Municipal*

Only a fraction of the quite numerous municipal censuses preserved in the local public archives have been transcribed, namely:

- Bergen 1912, Hamar 1920, Trondheim 1925 og Strinda 1934.

- Kristiania and Aker 1923 is currently being processed (50% completed by February 2011).

**Parish registers (*kirkebøker*)**

Practically all volumes of parish registers have now been scanned and made available by the DA up to the point in time (1930) where legal restrictions prohibit public insight on the grounds of privacy.
Altogether, the scanned registers comprise 12 687 volumes. Within these volumes a total of 60 121 separate lists (baptisms, marriages, burials etc.) have been indexed, i.e. an average of 4.5 lists per volume.

By May 2011, 2775 lists have been digitally transcribed. Some of these entities consist of several separate lists joined together, e.g when the lists of baptisms from one particular parish are merged into one from a sequence of register volumes. The number of transcribed *separate* lists is estimated to reach more than 4000. On these grounds we estimate that 6,7 % (4000 out of 60 000) of the parish registers have been transcribed in a manner that is directly applicable for the HPR. This is, however, a minimum figure, since quite a few of the reported 60 121 lists will not be of relevance to the HPR, many because they are duplicates.

*By an educated guesstimate the current percentage of transcribed parish registers may be about 10*. Undoubtedly, a number of parish register lists are transcribed locally on a private basis which have not been authorized and published on the DA or the RHD. The extent, quality and availability of these lists cannot easily be evaluated, and the project group has not yet found resources to make an inventory.

**Emigration lists**
All relevant emigrant- and passport ledgers have been microfilmed and scanned. All emigrant ledgers and practically all passport ledgers have also been digitally transcribed.

| | | | Remaining transcription | |
|---|---|---|---|---|
| Sources | | Total records | 1801-1959 | 1801-1910 |
| Census | 1801 | 883603 | 0 | 0 |
| | 1865 | 1701756 | 0 | 0 |
| | 1875 | 1813424 | 900000 | 900000 |
| | 1890 | 2000917 | 2000917 | 2000917 |
| | 1900 | 2240032 | 0 | 0 |
| | 1910 | 2391782 | 0 | 0 |
| | 1920 | 2649775 | 2649775 | |
| | 1930 | 2814194 | 2814194 | |
| | 1946 | 3156950 | 3156950 | |
| | 1950 | 3278546 | 3278546 | |
| Baptisms | 1801-1959 | 7881121 | 7093009 | 5088464 |
| Burials | 1801-1959 | 4508279 | 4057451 | 2981562 |
| Weddings | 1801-1959 | 2231793 | 2008614 | 909258 |
| Sum | | 37552172 | 27149456 | 11070201 |

Table 1: Number of records in relevant and transcribed source material 1801 to 1950

# Data Entry (Kåre Bævre)

In order to realize the project goals, the most resource demanding task is the data entry of the required sources. Table 1 overviewed the most relevant sources and their volume. A detailed discussion of the various sources is considered beyond the scope of this report. There are a lot of studies discussing various aspects of the sources older than ca 1910 (see Thorvaldsen 1996, Backer 1947-8), while more work will be done on the newer source material as part of the HPR project. The table shows the number of persons in the various sources according to the official population statistics. This provides a good estimate of the volume of person records. All in all there is a total of 37,5 million records (1801-1950). Of these it is estimated that about 30 % is currently transcribed. The bulk of the transcribed records is made up by the censuses 1801-1910. This is the result of various efforts by several people and institutions over the years. Much has been done by the Norwegian Historical Data Centre (RHD) and the National Archives and its associates, but there are also considerable contributions from volunteers. A fair share of the work has been done by people in various public labour market programs (most notable is the Teleslekt project 1993-1998).

The large volume of work remaining dictates three critical points:

1) Investments in solutions for more efficient data entry
2) A strategic plan and time schedule for efficient organization of the task at large
3) Mobilization of voluntary efforts and utilization of cheap labor

The investigations carried out during the pre-project suggest that there is considerable scope for cost-efficiency gains along all three lines. We estimate that these gains will be sufficient to make realization of the project's main goal of a full register 1801 to the present realistic within planned budgets and time-windows. It is also worth stressing that, if the proposed HPR project is not implemented, work on transcription will continue both by public organizations and volunteers, but at a much slower pace and with a minimum of cost-efficiency improvements. To put things into perspective: the National Archives and RHD over a 10-year period used 60-70 man-years to produce the current transcribed version of the 1910 census. The efficiency improvements that will be provided by financing the full scale HPR project will be enough to cut the resources needed for transcribing the 1891 and 1920 censuses by a factor of 2 to 4. We shall also provide solutions that make it legally and technically possible for a major mobilization of volunteers and/or outsourcing to low-cost countries. For the 1920 census alone it should, therefore, be possible to save 30-50 work year equivalents (WYE). These are resources that, were the HPR project not to be implemented, will probably be spent inefficiently by public organizations over the next 10 year period.

**Brief overview of current practices**

Existing transcriptions are being done by verbatim letter by letter copying through the keyboard. Extensive use has been made of programs such as CensIn, BD87 and Augustus, which are simple database applications tailored to standard field structures for the various types of sources. When transcribing one has used to work with paper copies or microfilmed images of the sources. Naturally, the transcriptions and the sources are separate. At best they are linked by indexing records according to page and line number in the source, but there are also cases where not even the protocol number is included in the transcription. A fundamental principle for all transcription carried out by RHD, the National Archives and associates has been to keep transcriptions true to the source (i.e. transcribe the original information even if it is known from other sources to be wrong, only correcting or commenting on errors found with pure

logic). There has also been a strong emphasis on full transcriptions, indexing the sources by only transcribing core fields has generally been discouraged.

**Scanned images**

During the last years there has been a massive effort to publish scanned images of church books (complete) and some censuses on the internet. This has changed the premises for transcription radically, and opens up for new solutions that will make the process far more efficient and convenient without reducing the data quality. These possibilities can be summed up in three main points:

1) Easy access to the sources for a large public of potential transcribers
2) A 1-1 link between image of an entry in the source and the corresponding record in the transcribed database
3) Advanced use of image processing techniques.

**Link between scanned image and transcription**

When transcribing sources from images it has now become standard practice to include links to each relevant image in the transcribed database. Essentially, this makes the scanned image part of the database. Individual records in the database and entries in the sources can then be related 1-1 effectively and securely. Image processing techniques can contribute to making the link even tighter; ideally by splitting images on a page with a table structure into the corresponding data structure of transcribed data fields (i.e. each individual entry in the database can be associated with a sub-image of the corresponding line in the source, and further: the name field can be associated with a sub-image of the name only). During the pre-project we have successfully developed an algorithm that does this for the 1950-census. For more irregular sources like the oldest church books it will be considerably harder to automatically read the table structure, but it might still be useful to use similar techniques to at least identify lines (see Eikvil et al 2010 for more on this).

One obvious advantage of the integration of images and transcriptions into a common database framework is more efficient and reliable proof-reading. Further, when images of the full entry are included it no longer seems worthwhile to transcribe fringe information that is not central to use of the database as such, since it can be gleaned from the image when relevant. This can save considerable time and effort. For example, a major challenge when transcribing is to decipher irregular entries like comment fields. It also becomes more relevant to consider transcriptions that are more of the indexing type. For example, one can concentrate efforts on only transcribing the fields that are most relevant to linking and identifying persons (name, sex, age/birth date, location) as well as their status (marital status, household position and occupation), while leaving fields such as witnesses, religion etc for an optional second round. There should be a minimal loss in dividing the task in two, so the gains of prioritizing the most relevant fields first do not need to be calculated against future costs of filling in the picture. For the HPR project such a strategy is highly relevant. Note on this point that it is fully possible to make a more complete transcription for a sample of the population once individuals have been identified in the core HPR structure. (Sampling in the first round makes identification through record linkage inefficient.)

When images of the source material are divided into its table structure this also makes it possible to allow for anonymized transcription of sources which cannot be made public. Fragments of the images that contain so little information that they do not in any way identify individuals will no longer have restricted access if presented to the public in a way that does not allow them to decipher the original content by combining the bits and pieces. This will make it possible to involve a large (open) audience in the process of transcribing also the sources that have restricted access. Such techniques will also alleviate problems that could otherwise make it difficult to send such material to low cost countries for transcription.

**Use of image processing techniques**

It is still unrealistic to hope for techniques able to read the bulk of the hand-writing in the sources with OCR like methods. The heterogeneity is too large, and state of the art techniques are too crude. Yet, modern image processing techniques still have a lot to offer. A detailed discussion is provided in the report by Eikvil et al (2010) prepared for this project. Here we will only highlight some main possibilities:

1) Fields with a limited set of values can often be recognized automatically (e.g. sex, civil status)
2) Numeric values can often be read successfully, e.g. age and dates. Importantly, many of the fields such as occupation and household position have been encoded and entered on the original scheme by Statistics Norway when working out the statistics for the censuses 1930, 1946 and 1950. The same seems to apply for causes of death in the lists of deceased produced after 1928. It is likely that these codes can be read automatically with good success rates.
3) We have already developed a successful algorithm for reading the underlining of standard words/categories used in the original questionnaire for the 1891 census. Similar procedures are relevant for the 1920 census. This only leaves a handful of fields left to be registered by other means.
4) A technique called 'word spotting' can be used to cluster pictures of words or names that are similar and thought to represent the same word or name. Manual inspection of these clusters can, for example, efficiently assign the name 'Ole' to hundreds of individual records in one single operation after a quick inspection for deselecting exceptions.
5) When context and additional information gives a strong prior or guess for the field to be transcribed one can pose the problem as a "confirmation problem" rather than a "recognition problem" (e.g "does it indeed say Ole Hansen?" replaces the question "What name is entered here?"). Algorithms for solving such problems are far more successful, and automatic techniques can then be used even on the more advanced fields. For the censuses it is usually straightforward to link places of residence across censuses. One can then hope to transcribe a fair share of the records in the one census (e.g. 1950) by initially assuming that the occupants are the same as in the last census which is already transcribed (1960). A similar technique was used manually when transcribing districts in the 1917 and 1918 censuses for Kristiania.

**Efficient organization**

The possibilities of combining manual and automatic techniques make it even more important that these activities are organized in the best ways possible. To illustrate we will use the 1950 census as an example and sketch a strategic plan for the best overall organization:

1) Scan all images
2) Split images into cells and populate the untranscribed database
3) Recognize or Read the birth date
4) Read other numeric fields automatically
5) Link place of residence to the 1960 census and the 1950 tax list (Matrikkel)

6) Automatic reading of as many names as possible using these two existing transcriptions as leads for a confirmation problem and/or use birth dates to link to the full 1960 census for candidates
7) Use cluster techniques on names. Use results of 6) as training data.
8) Manual transcription of remaining heads of household
9) Rerun step 6 using information from step 8 (i.e. using relations to head)
10) Manual transcription of remaining fields

Note that the sequence of these steps is crucial (e.g. that completion of 3 will make step 6 more efficient). Steps 3 and 7 in isolation can be carried out by volunteers even if the 1950 census is not open to the public. It is possible that the National archives will (weakly) enforce such activities on the users of their web sites. The field containing the birth date is unfortunately so small that automatic recognition techniques will only work in part. This single field is, however, very well suited for solutions of the re-capture type, i.e. solutions that require filling in a piece of text to verify manual entry when using a service on a website. The cost will be virtually zero and error rates will be kept low by double registration. The website Digitalarkivet had 155 million pages visited in 2010, so if say 1 out of 20 searches/displays required a cell transcribed a double transcription would be finished well within a year.

**Overall priorities and work flow**

There is a fundamental difference between the sources before and after ca 1920 due to the differences in legal restrictions, but also with respect to current transcription status.

For the period 1920-1950 the table is almost empty (with the notable exception of death after 1950 and a substantial volume of the early emigration). For the National Archives it is not an attractive option to devote their current transcription resources to work on sources that will not be publicly available for decades. The motivation of voluntary efforts is hindered by the same obstacle. Thus in terms of fresh resources needed this period is by far the most problematic. That being said, the potential for efficiency gains from using automatic or semi-automatic techniques is larger for the more homogeneous modern sources.

The 1920, 1930, 1946 and 1950 censuses thus form a major challenge. It seems clear that the 1946 census will have lower priority given its closeness to the 1950 census. One could argue that for covering the entire period 1801-present as best as possible as early as possible one should prioritize the 1930 census to establish a bridge from 1910 to 1960. However, we have concluded that the 1950 census should receive higher priority because: 1) the 1950 census will be the most valuable addition for the large group of researchers who today use modern register data from 1960 onwards, 2) it is likely that the information found in the 1960 census can be exploited to make the transcription of the 1950 census quite efficient. The format of the censuses 1930, 1946 and 1950 are very similar so the possibility of realizing economies of scale by doing them in tandem should be investigated further. The 1920 census is radically different and closer to the 1891 census. Both use single sheets for each individual and make extensive use of underlining. Since its public release is so much closer in time it seems far more realistic that both project-external public and voluntary resources can be allocated to completing this census. According to business as usually it is planned to be finished by 2020 without any contributions from the HPR project.

For the period before 1920 the censuses are to a large extent covered and the main challenge is to get the church books transcribed. There is already considerable voluntary work going on in that direction (as with the completion of the 1875 census). A premise for this work is that it has to be allowed to largely follow its own priorities and be delocalized. During the pre-project we have been in close cooperation with the organization DIS-Norge that organizes

much of this activity. Both parties see great promise that the HPR project will both substantially increase the efficiency of this work (online registration, use of clustering techniques etc) and mobilize more resources to this task. As already mentioned, a large community makes extensive use of the material provided on the web-pages of the RHD and the National Archives (more than 500,000 unique IPs in 2010), so the potential is there given the right technical solutions and organization. Some element of forced transcription is also considered. The launching of the HPR-wiki (see below) will likely generate large amounts of new traffic and activity that can also be utilized and generate more voluntary transcriptions.

## Name Standardization Procedures for the HPR

The creation of a longitudinal population register is feasible today with freely available record linkage software such as FRIL or FEBRL (Christen 2008) together with extensions developed during the HPR pre-project. Once the static source information (date and parish of birth and death, gender, nationality and names) has been encoded, this software uses these variables to identify record matches. All names in the 1801, 1865, 1875, and 1900 censuses—nearly six million person entries—have been standardized with the support of a grant from the Norwegian Research Council. The standardization was conducted with the assistance of Gulbrand Alhaug, Professor of Nordic Languages. The initial list contained 71,396 different first names, 125,631 different last names and 87,116 different place names; double names such as "Mary Anne" were split into independent names. The size of all name groups was inflated by "decorative" spelling variations and marks to indicate uncertainty about spelling. For instance, the female first name "Gjertrud" was spelled in 51 different ways in transcribed versions of nineteenth century censuses (Alhaug 2008). One problem is that traditional handwriting styles make it difficult to distinguish letters such as "o" and "a". In addition, there is redundancy among the three name groups; place names were increasingly used as last names during the nineteenth century, and most people had patronymics constructed from male first names. Finally, the patronymic suffixes –sen and –datter for men and women, respectively, could be spelled in many ways.

The most frequent name variant was used as the marker name for standardization. This process was facilitated by a database of name fields with frequencies and a list of censuses in which each name occurred. For example, the first name variant "Velhelm" was changed into the marker name "Vilhelm" since these two names occurred 69 and 4,681 times respectively. As illustrated in Table 2, two types of standardization were performed: the first removed orthographic variations represented by letters with the same linguistic value ("Cathrine" became "Katrine"); the second lexicographic standardization eliminated linguistic variation between names, such as "Karolus" and "Karoles." Lexicographic name variants are listed together in the name dictionaries, even if linguistically different. All transformations of name variants have been formalized into fifty name standardization rules; many of these rules specify the elimination or changing of vowels in unstressed syllables.

| A. Graphemic level<br>Orthographic variants | | B. Phonemic level<br>Linguistic variants | | C. Lexicographic level | |
|---|---|---|---|---|---|
| Caroles | 5 | Karoles | 51 | Karolus | 391 |
| Carolus | 107 | Karolus | 340 | | |
| Charolus | 2 | | | | |
| Karoles | 46 | | | | |
| Karolus | 231 | | | | |
| 5 variants | 391 | 2 variants | 391 | 1 variant | 391 |

Table 2: A three level model of name variants

All changes were documented by noting the relevant rule(s) applied to each name in the database. For instance "z>s; nd>nn" means that "Zimmermand" was standardized to "Simmermann." As this example shows, more than one standardization rule could be applied to the same name. After standardization there were 9,067 male first names, 10,058 female first names and 16.392 last names remaining. These totals do not include names occurring only once and female patronymic last names, which are nearly all variants of male patronymics. The variation that remains is still significant and will disrupt record linkage by not matching variations of the same name. Even at the lexicographic level, 53 percent of the 24,885 different first names occurred only once in the sources—a surprisingly high proportion when we remember that individuals could be represented two or three times in the 1865, 1875, and 1900 censuses. Therefore, it is necessary to use automatic name string comparison routines, such as Levenshtein, in a subsequent step in order to bring more name variants together during record linkage (Wikipedia: Levenshtein distance). The use of the Jaro-Winkler string comparison routine is explained in the paragraphs on record linkage below.

There is greater variation among last names than among first names, with more than 100,000 names identified on the graphemic level. Only a minority (21 percent) of these last names was patronymic names, but patronymics were used by a wide majority (76 percent) of the population. After removing farm names and foreign names, a significant group of the non-patronymic last names were found to be of foreign origin. These 8,169 different surnames belonging to some 25,000 persons were classified by country of origin as a proxy for the missing birth place field in the 1801 census (Sogner and Thorvaldsen 2002).

# Record linkage

The HPR project has made both theoretical and practical advances in the field of record linkage. The full implementation of these techniques will be crucial for the building of a population register from nominative records found in the source material. The theoretical advances relate to how fuzzy data elements can be handled. The identification variables in the records are gender, first names, last names, birth year and sometimes birth dates. Birth place is also in principle a stable identifier, while domicile and occupation should not be used as linkage variables since they can lead to skewed data sets. The latter variables have even so been used to control the result of the linkage.

Any linkage variable can be fuzzy in historical source material, maybe with the exception of gender. In probabilistic record linkage theory it has been "good Latin" to combine the similarity of variables with additive or multiplicative matchscoring. We have rather developed a method were we attempt to find what data elements are stable across different types of source material. Thus, some nominative records can be linked because the names are very similar, even if the birthdates are different in the linked sources. In other cases the names

are the fuzzy variables, but they can still be linked because there are stable birthdates bringing source material relating to the same person together. This is especially beneficial when linking records for women, since their changing last names often give them low scores in matchscoring systems.

There will still be persons with fuzziness in all relevant linkage variables, especially in pre 1910 censuses where most persons do not have birthdates. These persons' records can still be linked in many cases through their relations with other family members. Either other family members have enough unfuzzy characteristics that they have already been linked, or couples can be identified (e.g spouses or siblings) who together have enough stable variables that record linkage is possible.

It has been argued that using family relations which in principle are temporary characteristics as linkage criteria will create skewed samples of linked persons because the couples will be overlinked in relative terms (cf Historical Methods 2011/1). For two reasons this is not a serious problem according to our aims. First, since Norway has taken censuses at regular intervals through the two centuries since 1801, it is feasible to construct weights that can be used to compensate for the bias. Second, our aim is to link the vast majority of Norwegians through this period. This will be done with a combination of automatic and manual methods, and all links made will be flagged with the relevant linkage criteria, so that researchers can eliminate any linked records whose basis they do not trust. Local tests show that it is realistic to safely link more than 80 % of the population. When source material from the whole country has been transcribed throughout the whole period, we believe that we can raise this relative figure, because we shall also be able to link the migrants inside Norway. Problems with emigrants and returnee migrants from abroad have been discussed in a separate paper submitted for publication.

**Incremental linkage**

Typically, historical sources contain some records that are unique and with clear identifiers while other records have fuzzy identifiers creating duplicate linkages. In between there will be many degrees of borderline cases. We have, therefore, developed methods to link the most obvious cases first and then gradually proceed with the records that can only be linked with more fuzzy logic. For instance records with identical birthdates and first and last names in the census and baptismal records can first be linked, while those with somewhat different day, month or year of birth or names spelt differently should be dealt with later. This is also good in terms of computing power, since the set of linkable records will be reduced in the first linking rounds, leaving fewer records to be dealt with by the more resource demanding fuzzy algorithms.

By flagging the records according to the linking criteria, the researchers can be told what variables where used to link two or more records in each specific case. On this basis the researcher can decide what parts of the linked sample to employ in different types of study. For instance those who need all links to be as certain as possible, can decide to throw out of the sample records that did not match exactly on birthdate and had a Jaro-Winkler score close to 1 for the name comparison (cf below). Other researchers' priority will be to enlarge the linked sample, and they can choose to include records whose birthdates were off by up to a month and had a reduction in Jaro-Winkler scores of up to 0.2 points. This will be a more instructive basis for including linked subsamples than abstract additive score where it is unclear what caused the variation.

**Fuzzy algorithms**

Three main types of fuzzy algorithms have been used to link the nominative datasets for the HPR project, one for the names, one for birthdates and one for birthplaces.

*Names: Jaro-Winkler distance*: This algorithm computes the similarity between text strings giving the value 1 if the two strings are equal and 0 if there is no similarity. Although the algorithm is weighted to stress the initial characters in the strings, it will give sufficient weight to the ending, so that similarity between the first name in the string from source one and the second name from source two can receive a relatively high score. Thus a link can be created between e.g records with the names 'Andreasd. Østhagen' and 'Østhagen' (score 0.45). Thus, this woman could be linked despite her dropping of the patronymic element in her last name. For further details cf Jaro-Winkler in Wikipedia.

*Birthdates*: All birthdates are converted to an internal date format so that distances between dates can easily be computed as the absolute value of the difference between whole numbers. Imprecise dates in the sources were converted to the nearest or most logical realistic date. Checks have disclosed that both the day, month and year element may be imprecise, in censuses from the early 20th century a tenth of the birthdates were taken inaccurately.

*Birthplaces*: In municipalities with relatively few in-migrants, birthplace does not distinguish well between persons, and may rather be used as a control that the linkage algorithm is not creating false links. In places with much in-migration, typically cities, birthplace is needed to distinguish between persons with the same names and birthdate. For this purpose birthplaces are encoded into a hierarchical four digit system distinguishing between provinces on the highest level and municipalities on the lowest. These numbers may be off the target in cases where municipalities were split, merged or had their borders changed between censuses. Since neighbouring municipalities often have contiguous id numbers, this can be remedied by using only the first three digits of the id number, or more precisely by allowing a small difference between id numbers.

*Family relations:* Especially when some of the variables listed above are missing from the sources, it is important to have access to information about persons who belong together as couples, families, or in households. The 1801 census lacks information about birthplace, and before 1910 censuses contained year rather than date of birth. Also, early 19th century church records often lacked the age of the bridal couple or of the parents at baptism. In these ministerial sources the relations between the baptized child and its parents or between the bride, the groom and their fathers are explicit since these entries were transcribed into one record. In the censuses the family interrelationships are made explicit with location pointers between the spouses and between children and parents created by the IPUMS projects. By using these pointers in own joins in SQL statements, data joined from related persons in a census can be linked to other censuses or church records.

Often linking by relations cannot be performed on records from the burial protocols, since here many persons are not mentioned together with a relative. It is difficult to use such techniques for single persons, typically young people who are lodgers after they left their family of origin and before starting a procreation family. This technique has, therefore, been criticized for creating biased linked samples where single persons are underrepresented. Since our aim is to link a high proportion of the population into the HPR, we shall accept such links while flagging them explicitly. They can usually be corroborated at a later stage, especially by using the data about fathers in the marriage protocols.

*Address / Domicile*: The record linkage literature generally advices against using address (or occupation) as an id variable, because a bias is likely to result from more easily linking non-migrants than migrants. Therefore, we have in most cases avoided using address as a criterion for automatic linkage. Undoubtedly, when nominative sources are linked manually domicile will often be a decisive factor for creating links. It can also be argued that using address will enhance the quality of the links made, especially when resolving duplicate links. When linking early 19th century church records where year of birth may be missing, address will often be necessary for linking the majority of people who have common names.

## Record linkage exercises

The HPR exercises with record linkage have resulted in several longitudinal data sets, some of which are already available for researchers and use has started. Record linkage has been performed by Statistics Norway, the Minnesota Population Center and The Norwegian Historical Data Centre (RHD). The linkage work is centered on the censuses, but to some degree church records (baptisms, marriages and burials) have also been linked to the censuses. While the pre-HPR record linkage to build the BSS, Rendalen and Asker databases was manual or interactive, the linkage work done in Minnesota and by the RHD was done with automatic or semi-automatic algorithms.

### Rendalen 1910 to 1950

The Rendalen database built by professor Sølvi Sogner and her associates at the University of Oslo contained linked censuses, church records and other types of sources from 1735 through 1900. The HPR pre-project added the censuses 1910, 1920, 1930, 1946 and 1950, as well as church records up to the late 1920s. The fuzzy sql method outlined above was used to link these censuses and the baptisms pair wise back to the original database, with first name, last name and date of birth as the linkage variables. Even if most women changed their name at marriage, the birthdates still made record linkage realistic in most cases, even if about ten percent of the dates contained (mostly minor) errors. Birth place and domicile information was not used as linkage variables, but could be used to corroborate the links. Some extra links between two censuses could be made by "correcting" data in one census with linked data from a third source (for instance using birthdates from the baptism file in order to correct the 1946 census so it could be linked to records in the 1950 census). Linkage rates varied from 80 percent (1946 census against baptisms) to 90 percent (1910 against all earlier records including baptisms). Manual checks reveal that missing links seldom had other causes than migration. The linked files are available to researchers upon request.

Table 3: Example of links made with different criteria between the 1950 and 1946 censuses (Jaro 0,9 = Jaro-Winkler score for both first and last names at least 0,9 – fn = first name, ln = last name, <> = not tested:

| | **1950** | **1946** |
|---|---|---|
| Population Rendalen census records | 4072 | 3933 |
| Jaro 0,9; identical dates | 2038 | |
| Children born after 3.12.1946 | 306 | |
| Preliminarily present (not resident) | 218 | |
| Jaro 0,8; identical dates | 428 | |
| Jaro 0,8; date difference < 9 | 92 | |
| Jaro fn 0,7; ln 0,9; date difference < 2 | 254 | |
| Jaro fn 0,9; ln <>; dato <2 | 58 | |
| Jaro 0,8; ln <>; identical dates | 50 | |
| Jaro fn <>; ln 0,8; identical dates | 72 | |
| Net in-migrants | 187 | |
| *Not linked, likely migrants* | *369* | |

Table 4: Links made between individual records from one census to the previous in Rendalen 1950 to 1910 (1910 census linked to Rendalen database) . Absolute and relative number of links made in proportion to people available for linkage. Temporary visitors, in-migrants and people born after the previous census were excluded from the analysis:

| | 1950 | 1946 | 1930 | 1920 | 1910 |
|---|---|---|---|---|---|
| Population | 4072 | 3933 | 3579 | 3599 | 4328 |
| Available | 2875 | 2426 | 2507 | 2158 | 3415 |
| Linked | 2506 | 2069 | 2077 | 1869 | 3063 |
| *Percent* | *87 %* | *85 %* | *83 %* | *87 %* | *90 %* |

**Rendalen linking in the Oracle database with PL/SQ**

We have implemented a prototype PL/SQL code package to aid in the linking of church records (baptisms, marriages and burials) from 1901 to the late 1920s to the Rendalen database and internally in the church records. The package runs on an Oracle version of the database with Oracle SQL Developer. One of the advantages of using PL/SQL over SQL is that more logic can be built in. Some linking has already been done on this data, and with the aid of these packages many of the existing links can be confirmed, and errors and new links can be found.

So far a first version of a package for linking marriage records to baptism records has been implemented, and a first version for linking burial records to baptism records is almost finished. Code needed by more than one package has been placed in common routines, among this is code for calculating Jaro-Winkler based scoring while taking the varying nature of the data into account.

The linking is done by starting out with strict matching criteria (same birth date, same first name, same father's first name) then if no match is found progressing to less and less strict matching criteria (permitting some differences in birthdates and using Jaro-Winkler based scoring for names). Potential new links and potential errors are saved to new tables for verification. Verification is partly done automatically and partly manually. The Jaro-Winkler based scoring code has been made to facilitate handling of variations in names (such as typos, abbreviations and varying number of words in names). This PL/SQL Jaro-winkler function returns a score between 0 (no match) and 100 (perfect match).

Some numbers for linking of marriage records to baptism records:
- Total number of recorded marriages from 1901 to 1931: 506
- 819 persons (412 men and 407 women) were already linked.
- 55 new links were found (22 men and 33 women)
- 524 existing links were confirmed (263 men and 261 women)
- 11 errors were found (3 men and 8 women)

The information in burial records varies depending on the civil status of the buried person, so for the linking of burial records it was necessary to differentiate between children, unmarried women, married women and men when linking them to the baptism records. Often information about the person’s father is missing, and for married women it is their husband that is given as next of kin, and not their father. In cases like this the buried persons surname was checked to see if it contained a patronymic, and the root of the patronymic was compared to father's first name in the baptism records.

Some numbers for linking of burial records to baptism records (as of now, work is still being done on these, and some of them need checking):
- Total number of recorded burials from 1901 to 1928: 1070 (572 men and 498 women)
- 394 persons (209 men and 185 women) were already linked
- Numbers from linking children:
    - Total: 257 children (132 male and 125 female)
    - 115 children already linked (58 male and 57 female)

    - 74 potential new links were found (34 male and 40 female)
    - 109 existing links were confirmed (57 male and 52 female)
    - 2 potential errors were found (1 male and 1 female)
- Numbers from linking unmarried women:
    - Total:  88 unmarried women
    - 25 unmarried women already linked
    - 42 potential new links were found
    - 18 existing links were confirmed
    - 1 potential error was found
- Numbers from linking married women:
    - Total: 284 married women
    - 103 married women already linked
    - 123 potential new links were found
    - 57 existing links were confirmed
    - 9 potential errors were found
- Numbers from linking men:
    - Total: 440 men
    - 151 men already linked
    - 183 potential new links were found
    - 86 existing links were confirmed
    - 28 potential errors were found
- Totals:
    - 422 potential new links were found (217 men and 205 women)
    - 270 existing links were confirmed (143 men and 127 women)
    - 40 potential errors were found (29 man and 11 women)

**Linking the 1865, 1875 and 1900 national censuses**

These censuses were linked on the individual level as part of the North Atlantic Population (NAPP) Project. In addition to US censuses, only Norwegian material was selected for linkage, while censuses from Sweden, Great Britain, Germany and Canada are still only available cross-sectionally. This record linkage exercise is described in articles in a special issue of *Historical Methods* (2011/1) and datasets with documentation are available online to registered researchers at http://www.nappdata.org/napp/linked_samples.shtml. Because of the female name changes, the linked samples were restricted to males and couples. Two-thirds of the nearly 600 000 links made were from the Norwegian censuses. The variable LINKTYPE specifies the reason why a link was included in the output, while the variable LINKWT is a proportional weight which must be used to generate representative estimates for the whole population. This is both because the 1875 census when linked contained only 1/3 of the Norwegian population, because linkage tends to skew the sampled individuals and because each linked person or couple is joined by their household, which in case several household members are linked may be represented more than once in the output. We shall initially also adopt this weighting scheme to the more comprehensive HPR with ministerial records and other sources in addition to the censuses. Since we shall allow the users to download extracts of linked records which satisfy certain quality criteria (e.g only records with links based on matching birthdates), the weights need to be adjusted to mirror the relationship between the population and the actual sample downloaded. For this reason we aim to develop a dynamic interface which will allow the users to select sample specific weights.

The linking was based on four permanent characteristics or variables: Birth year, first name, surname and municipality of birth. (In the US race was added.) Linkage was not based on characteristics of co-residing household members, address or other domicile information, since this might create biased samples by favouring the linking of persons / couples in large

households or non-migrants. However, once some person in a household was linked (as a primary link), it was attempted to link other members of the same household as represented in the two censuses (as secondary links). In addition the households as represented in the two censuses will contain unlinked persons. We do not yet know how many of these ought to have been linked given more detailed criteria or sensitive linkage algorithms, but future manual linkage with the HPR-wiki will provide test results about this and other types of linkage errors rates.

Both given names and surnames where standardized according to information in our normalized person name database prior to linkage. Together with more detailed birthplace information, this improved linkage rates for the Norwegian censuses significantly compared with the US material. (Cf article by Vick et al in *Historical Methods* 2011/1.)

For the US samples the record linkage program Freely Extensible Biomedical Record Linkage (FEBRL) was run on a supercomputer in order to perform the millions of comparisons of names, ages and birthplaces necessary in order to identify potential links. For the Norwegian censuses the process was made more efficient by constructing stand-alone PERL scripts with the relevant parts of the FEBRL algorithms, most notably the Jaro-Winkler name similarity scores. (This is a development parallel to the SQL routines developed at the RHD with fuzzy logic both for the MS Access and Oracle platforms.) The resulting matrix of similarity scores was in the next round fed into the machine learning program LIBSVM which is a Support Vector Machine system, together with training data; linked records that had been verified manually. Potential links were now classified as true or false. A conservative approach was taken for security, so that records too similar to allow linking to only one or several similar records in the other census were not included in the linked sample.

Special care was taken with the Norwegian records to link together duplicate records for the same individual who was registered both as preliminarily present and absent due to the combined de jure/de facto system in the 1875 and 1900 censuses. Since patronymic suffixes can be written in numerous ways, these were stripped off the last names in order to not affect the Jaro-Winkler comparison. Some age discrepancy between representations in different sources for the same individuals had to be allowed, 3 years for individuals and 5 years for couples.

**Record linkage in the northern part of Troms province**

The censuses from 1865 and 1875 were chosen as a starting point. The main criteria for linkage were names, year of birth, and place of birth. To compensate for the different spelling and typos in the sources several methods were applied.

**Standardization**

The names were initially standardized and place of birth encoded. This improved the linkage rates, but we had to use fuzzy algorithms to achieve high linkage rates. Different algorithms were tested such as Levenshtein distance, Jaro–Winkler distance and Q-grams. The algorithm that is most successful for our purpose turned out to be Jaro-Winkler. Jaro-Winkler is implemented or is easily implemented in Oracle and MS Access (the two database systems we use) and in several programming languages.

**Linking individuals**

Several approaches to linking individuals were tested with SQL in MS Access and with linking software like FRIL and FEBRL. The results from linking on the individual level gave approximately 50% links.

**Linking families**
To get higher linkage rates with source material lacking birth dates we had to link families. This will give the opportunity to accept more variance in the variables and link on fewer variables. To make this possible the encoded family relation we have implemented in the NAPP database in Minneapolis was used. The FEBRL software was used to link previously unlinked families in 1875. A separate file was created for each unlinked family and this was run against the complete 1865 census. By sorting the result we could pick the family in 1865 that gave the highest score. This process was run several times with different linkage variables to overcome problems like changing last names, wrong birth place and more. To run this we work together with NOTUR - The Norwegian Metacenter for Computational Science at the University of Tromsø to implement FEBRL on their supercomputer Stallo.

**Linking church records**
The next step was to link church records to make the picture more complete and to assist in linking between censuses. The time gap between 1875 and 1900 made it difficult to get the same linkage rates as between 1865 and 1875. The main linking criteria were name and birth date. The family relation was easy to extract from the sources so the main linking method was based on family relations. The only exception is with the death certificates, which often contain information about the deceased person only. By linking the couples in the marriage and birth records, families can be created.

**Joining the links with IDS**
When linking many sources the process and presentation can be messy. To avoid this and moving toward the international community we implemented the IDS (The Intermediate Data Structure for Historical Longitudinal Databases) structure for storing our linked data. The IDS provides a uniform way of storing historical data. By moving all the data from the censuses and church records into the IDS we have only one comprehensive data source to link to and the different data will be stored in a uniform way. Another benefit is the potential of indirect linking. If a person is linked from baptism to the 1900 census with the information of his parents and the same person is linked to the 1910 census through the birth date, the link between the 1900 and 1910 censuses is an indirect link.

**Results**

| | 1865 | 1875 | 1900 | 1910 |
|---|---|---|---|---|
| Number of persons | 4614 | 5026 | 6326 | 6597 |
| Found in baptism | 1811 | 2268 | 2400 | 2726 |
| Percent | 39,25 | 45,13 | 37,94 | 41,32 |
| Number of children in 194* | 1875 | 2032 | 2632 | 3043 |
| Found in baptism | 1698 | 1867 | 1654 | 1817 |
| **Percent** | **90,56** | **91,88** | **62,84** | **59,71** |
| Number of persons married or widowed | 1664 | 1844 | 2331 | 2344 |
| Found in marriage | 1092 | 1220 | 1116 | 886 |
| Percent | 65,63 | 66,16 | 47,88 | 37,80 |

The number of children in the North Troms province (municipality numbers starting with 194) are those who are born in this region, also covered by the church registers for this database.

The reason for the lower percentage for the 1900 and 1910 censuses are that these censuses are not included in all linkage efforts at the same level as 1865 and 1875. The 1910 census was not included in the NAPP linking project. Also, the changes in last name practices from patronymic to family names especially for married women makes the linking of the 1900 and 1910 different from the earlier censuses. The routines for linking censuses and baptisms had to be rewritten to overcome the problems with changes in last name.

Gross relative number of links between the censuses (percent):

| | **1865** | **1875** | **1900** |
|---|---|---|---|
| **1910** | 29,12 | 31,02 | 44,06 |
| **1900** | 42,04 | 46,78 | |
| **1875** | 73,29 | | |

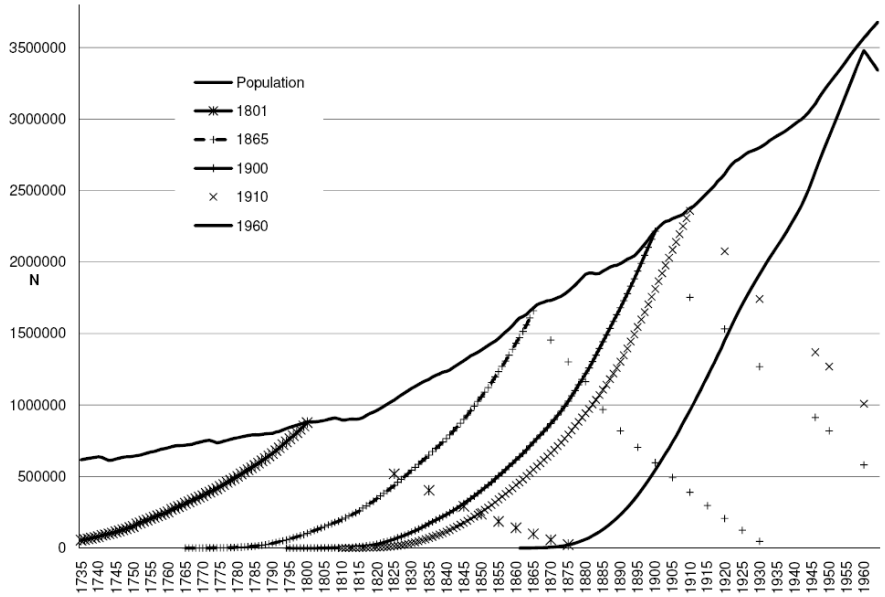
This shows that for instance out of the number of persons in 1910 that are born before 1866 there are almost 30% linked to the 1865 census.

## Central Population Register (CPR) (Helge Brunborg)

The central population register was established in 1964 on the basis of the 1960 census. It should in principle be possible to link everybody who was enumerated in 1960 with persons who were enumerated in previous censuses (or recorded in other sources).

Figure 1 shows the potential for linking transcribed Norwegian censuses with the 1960 census. The curve leading up to each census year shows cumulated birth cohorts as they are represented in our full-count transcribed censuses so far, i.e. for 1801, 1865, 1900, 1910 and 1960. The curves leading up to each census, which are based on the recorded age or year of birth in the censuses, show the cumulative part of the census population that was already born in each year prior to the census year. The curves descending from the census years are based on mortality and emigration data as well as on data from later censuses. These curves show the part of the census population that was still alive and remained in Norway during the post-census periods. Most of the population who lived in Norway during the last half of the nineteenth and early part of the twentieth century is represented in two or more transcribed censuses - which can be joined by record linkage.

**Figure 1. Population size of cumulated cohorts in full count censuses**



According to aggregate statistics 920,000 persons alive in 1910 survived to the 1960 census. Unfortunately, the CPR does not easily lend itself to be linked to the 1910 census with standard procedures. The reasons for this are incompatibilities both with respect to names, birth dates and birth place information. Especially, many married women changed their last name during the long time span in question, and many first names were spelt differently. While the date of birth in the CPR is of high quality, those in the 1910 census, the first census in which these data were recorded, contain errors in 10-20 per cent of the records. Moreover, the many changes in municipality borders make it difficult to use birth place as a straightforward linkage criterion. Together with extensive migration, these census variables explain why only some ten percent of the potential unique links could be made from 1910 to 1964.

An additional problem is that there does not appear to any longer exist any electronic version of the *full* 1960 census, neither in Statistics Norway nor in the National Archives. For example, the existing versions do not include the first and last names of the enumerated persons, which are the basic linkage criteria. Other variables which are useful for verifying uncertain links, such as place of birth, are also missing. Thus, older censuses and other data sources need to be linked to the full CPR (or early versions of the CPR). The disadvantages of this do not affect large numbers of persons, however, but may reduce the proportions of persons that can successfully be linked.

All events, such as births, deaths, internal and external migrations that occurred between 1 November 1960 and 1 October 1964, should in principle have been recorded and entered into the CPR, but this was not always done in practice. In particular, persons who were born and died between 1.11.1960 and 1.10.1964, or who immigrated and then emigrated in the same period, were most likely not given an ID number. This would not affect the linking of the 1960 and 1950 censuses, however.

Moreover, there are about 64 000 persons enumerated in the 1960 census for whom the ID number cannot be found in the CPR.[1] Another problem is name changes between 1960 and 1964. Statistics Norway has an electronic list of such changes, but it probably does not cover the period before the CPR was established. Finally, it is possible that some variables were identical but wrong in *both* the 1910 and 1960 censuses, and that these errors were later corrected in the CPR. For such cases it would have been useful to have the transcribed version of the full 1960 census.

With these shortcomings and problems we would not expect somewhat reduced linkage rate when attempting to link the 1910 censuses with the 1960 census (or the CPR), see table 5. Only about 4 per cent of the 1910 census and 8 per cent of the relevant CPR records were successfully matched with the other source. Due to limited resources there were no attempts at fuzzy matching. As expected, the percentage of successful links was highest for the youngest cohorts in the 1910 census, and significantly lower for women than for men.

[1] Børke, Sindre (1983): Folketellingen 1960. Ny organisering av datafilen. Interne notater 83/18, Statistics Norway. http://www.ssb.no/histstat/in/in_8318.pdf.

**Table 5. Results of first attempts at linking the 1910 Census with the Central Population Register, with exact date of birth and full name as linking criteria**[2]

| | Census 1910 | | | Found in CPR | | | Per cent linked to CPR | | |
|---|---|---|---|---|---|---|---|---|---|
| Year of birth | Men | Women | Both sexes | Men | Women | Both sexes | Men | Women | Both sexes |
| ≤ 1869 | 180 189 | 211 572 | 391 761 | 221 | 332 | 553 | 0.1 | 0.2 | 0.1 |
| 1870-1879 | 122 143 | 139 529 | 261 672 | 3 493 | 3 858 | 7 351 | 2.9 | 2.8 | 2.8 |
| 1880-1889 | 148 457 | 173 115 | 321 572 | 9 284 | 7 590 | 16 874 | 6.3 | 4.4 | 5.2 |
| 1890-1899 | 237 939 | 235 010 | 472 949 | 17 671 | 5 312 | 22 983 | 7.4 | 2.3 | 4.9 |
| 1900-1910 | 297 258 | 286 890 | 584 148 | 27 787 | 5 589 | 33 376 | 9.3 | 1.9 | 5.7 |
| Total | 985 986 | 1 046 116 | 2 032 102 | 58 456 | 22 681 | 81 137 | 5.9 | 2.2 | 4.0 |

A similar attempt was made to link a transcription of the 1950 census for Rendalen (3 839 records) with the CPR. The results are very promising, with an overall linkage rate of 53.1 per cent (table 6). This is probably due to more accurate recording of information in the 1950 than in the 1910 census. Moreover, a relatively low number of persons died or emigrated between 1950 and 1960.

**Table 6. Results of record linkage of the 1950 census for Rendalen with the Central Population Register, with exact date of birth and full name as linking criteria**

| Year of birth | No match | Matches | Total | Per cent linkage |
|---|---|---|---|---|
| 1830-1839 | 1 | 0 | 1 | 0.0 |
| 1840-1849 | 5 | 0 | 5 | 0.0 |
| 1850-1859 | 60 | 1 | 61 | 1.6 |
| 1860-1869 | 177 | 37 | 214 | 17.3 |
| 1870-1879 | 170 | 146 | 316 | 46.2 |
| 1880-1889 | 184 | 191 | 375 | 50.9 |
| 1890-1899 | 195 | 270 | 465 | 58.1 |
| 1910-1919 | 241 | 313 | 554 | 56.5 |
| 1920-1929 | 251 | 308 | 559 | 55.1 |
| 1930-1939 | 219 | 307 | 526 | 58.4 |
| 1940-1949 | 266 | 430 | 696 | 61.8 |
| 1950-1959 | 31 | 36 | 67 | 53.7 |
| Total | 1800 | 2039 | 3839 | 53.1 |

In the full HPR project the problems mentioned in this section may be solved through fuzzy and other forms of creative matching, and by linking the censuses from 1910 through 1950, which for Rendalen was shown to result in linkage rates of up to 90 per cent. In this way the birth dates can be corroborated and we will have information on several name forms and on women's family names both before and after marrying. We can also construct the earliest possible cross-section from the CPR and adjust municipality information according to the information about the many mergers of administrative units during the relevant period.

[2] The linking was made with a 2010 version of the CPR (BREG) and a version of the 1910 census including 2 473 449 records and 37 variables. There were many invalid values in the 1910 census and some of these were recoded. 82.5 per cent had 1850-1910 as year of birth and valid values for date of birth. From these 9 566 duplicates were removed. From the CPR file 1 027 474 records were extracted with valid values for year of birth (before 1911), date of birth, first name and family name. The earliest year of birth in the CPR is 1855.

# Family Reconstitution Type Databases (Arnfinn Kjelland and Ole Martin Sørumgård)

This section deals with databases constructed and used for producing the particular Norwegian genre of local history; the farm- and genealogical history. This is a sub-category of the main genre «*bygdebok*» or community history which is complemented with ordinary local history overviews. There exist a number of such databases in Norway. They contain in principle basic, time-stamped information about all existing or deserted dwellings (farms, cottar's places, modern houses and even apartment buildings) and all people that have lived in the geographic area covered and relations between people grouped into families or ancestries and between these people and their dwellings. However, the database designs and models vary quite a bit.

## The farm- and genealogical 'bygdebok' genre in Norway

In Norway, this genre of community history books has a historiography back to the beginning of the 20th century (Sørumgård and Kjelland 2003; Winge 1995: 244ff). Initially it aimed to tell the history of only some of the farms and families in the community, but during the last 30 years it has become normal that all farm holdings and other dwellings are described, and all persons that have left traces in the sources are listed.

Persons are combined into families, usually nuclear families (not households since co-resident servants and distant relatives are usually omitted). Details on each nuclear family are listed chronologically at one of the dwellings where they lived. Furthermore, each couple and individual has an entry specifying their year or date of birth, marriage and death and cross-references to other dwellings where he or she stayed, place of birth etc. Because of this it is possible to trace families and individuals as they moved from place to place, map the pedigree of a selected person, the degree of kinship between individuals and so on – within the study area.

The research work has usually been done by one author, trained as a historian or self-educated, the overall intention of such projects being to publish the specific genre of community history books. The responsible publisher is normally the local municipality, which covers the expenses for collecting sources and illustrations (old and new photographs and maps) and compiling the final text. Today it is becoming more usual to organize a team of associates who works together on different tasks in order to fulfill the project faster. The author may then be the project leader and main editor.

Such books usually consist of several rather large volumes, and may be several thousand pages long even if the municipality isn't very large. E.g. the three volumes for Lesja municipality (2200 to 3 660 inhabitants) published 1987–1996 consist of 2150 pages. Their content as far as population information is concerned starts around 1600, first only as names of heads of households in tax lists, gradually expanding along with the available sources (church records, probate registers, censuses). From the mid 18th century one can expect the sources for a normal Norwegian municipality to be adequate for family reconstitution. Such projects normally last for a decade or longer, and involves rather high public expenditures (hardly below NOK 5 mill. today). They generate some income from book sales, but hardly more than 20–30 % of the total expenditures. Still local municipalities, local history organizations or others are able to gather enough sponsors to launch a number of such projects annually. While we initially expect the Historical Population Register to benefit from databases constructed for community history purposes, in the long run we expect the existence of the HPR to reduce these public expenses significantly.

## The sources utilized in a typical farm- and genealogical 'bygdebok'

Basically, the author utilizes sources according to the resources assigned to the project. This may differ a great deal, from only the basic church records and censuses to almost all sources that contain information about individuals and families (Sørumgård and Kjelland 2003 p. 3). Most such projects try to give the same kind of information for individuals and families up to the time of printing. Due to legal restrictions on government information about living persons, the 'bygdebok' projects have to collect the same kind of information with special surveys distributed to all households. Other sources used are registrations from local newspapers, graveyards, local history publications etc.

## The methods used in a typical farm- and genealogical 'bygdebok'

The methods and source material heavily utilized in the genre of farm- and genealogical history may justify denoting the system "*the Norwegian extended family reconstitution method"* (Sabean 1998: 4). Two specific characteristics of Norway's history justify this national concept: the settlement structure and the farm and family- or surname system of the country (Sørumgård and Kjelland 2003; Kjelland 1996). Combining these two characteristics with the ordinary extended family reconstitution method and its sources makes it rather easy both to reconstitute the families who have lived in such a study area and to follow their footsteps as far as they have left traces in the sources utilized in such projects.

One more feature should be mentioned: during family reconstitution in historical demography the purpose is mainly statistical; a few wrong links do not matter for the result of the investigation. But when you write a Norwegian 'bygdebok' in this genre it should be stressed that by definition *no wrong link is tolerated*. Many people in the study area have extensive knowledge about their ancestors and pedigree, and if the book presents more than a very small number of «wrong» links there will be complaints sent to the author, published in the newspaper etc together with ensuing discussions. Thus, a Norwegian 'bygdebok' of this genre will contain both a family reconstitution from the 17$^{th}$ century until today, the place(s) of residence for all heads of households and indirectly e.g. place of birth for all children born in the area under investigation.

## 'Bygdebok' Type Database Systems

In Norway the first attempt to develop a relational database system for such projects started in 1987. Authors of such books saw the possibilities for savings especially in the time-consuming manual record linkage process that has to be carried out in order to reconstitute families and decide the place(s) they have lived in order to organize the manuscript properly.

Supported by The Norwegian Institute of Local History (NLI) and a regional computer firm the program *Ættesoge* («Genealogical history») was launched in 1988. Several projects have been fulfilled aided by this program. Almost at the same time another project, *FAMREK*, was developed and released by a group of teachers in mathematics and physics with an interest in local history in the Dovre municipality. The 'bygdebok' author Johan Borgos utilized off-the-shelf programs for the same purpose. (Cf also Kjelland 1990.)

During the 1990s a number of local historians utilized larger hard drives and more sophisticated relational database systems (Dbase II, MS Access) for processing parts of such projects. The project *Busetnadssoge* (BSS) started in 1999 and the software was operational by May 2004. It supports the work process from the point where the primary sources are ready to be imported to a type-set book manuscript. A number of databases have been created also with this program, and a number are under construction in ongoing 'bygdebok' projects.

During the HPR pre-project we have mapped all existing 'bygdebok' databases, both those completed and those under construction (see attachment 1). These 'bygdebok' database systems have been used to create a number of databases, each usually covering a parish or

municipality. Unfortunately, all such projects are locally initiated, financed and completed, and the way the author or municipality archives the computer files after the books are published varies considerably.

## The BSS program package

As explained in Sørumgård and Kjelland 2003 the BSS program package consists of an application and a Paradox database. Appendix 3 in that paper displays a simplified overview of the database model, with its basic entities and relations. The basic database structure remains the same today as in 2003, but the application has been enhanced in several ways recently, especially with better tools for name standardization and more sophisticated searching algorithms. It is now possible to look up all occurrences of a specific couple or individual among events in the database, just by marking one or more events as a starting point and using single keystrokes to perform the actual search. This has proven to be of substantial help by reducing the required time in 'bygdebok' projects, and should also serve the author in detecting more reliable links.

Still, the main principle in the current version of BSS is to use manual record linkage. As stated before, it is crucial to minimize the number of erroneous links in these types of project. A preliminary algorithm for semi-automatic record linking has been implemented and briefly tested, but the result was not promising. Since automated linkage results among the other HPR partners seem promising, there will be synergy effects from exchanging algorithms, so that the time saved with automatic searching and linking is not eaten up by the extra effort needed to investigate all decisions made in the process and eventually correct many of the established links. Further development of the BSS program package will also be directed toward an extension of the user interface, with additional functionality and better reporting tools. But, improvements are also needed when it comes to more general technical solutions, like upgrading of the database and better memory management. This will hopefully speed up some of the most time consuming processes, thus reducing idle time for the researcher or 'bygdebok' author.

## From 'Bygdebok' Type Database to HPR – data exchange format

Figure 2 sketches different kinds of data exchange between BSS and other databases or datasets. First, BSS has its own data format for importing all types of source material. A couple of tools have been developed to convert the original, often arbitrary datasets, into this format. The process includes necessary coding of relationships, and some controls to ensure good data quality.

As shown in the figure, BSS also needs an internal data exchange format for sharing information between BSS databases. Used in the right manner this can be helpful especially in larger projects. Currently BSS is a single-user application, so it is not rational to distribute for instance record linkage among several users. A good algorithm for merging different versions of the same database will probably lead to a better utilization of available resources.

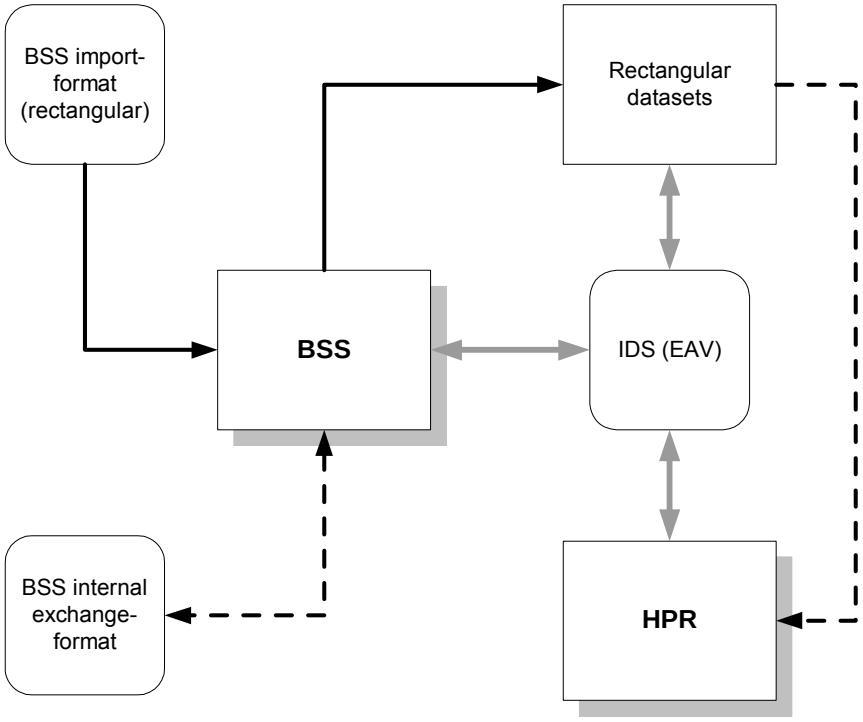



Figure 2: different types of formats for data exchange to and from BSS databases.

During the HPR pre-project a prototype data export module in BSS has been developed generating rectangular datasets suitable for further analysis with other software, for instance MS Access or statistical packages. The export format consists of three main tables representing entities such as persons, the relations between them, and all events each person has participated in (church records, censuses). Two additional tables are needed for information about the places and the inhabitants; the details here are still being expanded. Specific details about the preliminary export format are not included in this report, here it suffices to say that, where appropriate, many of the attributes have a corresponding timestamp. However, some attributes can in fact hold multiple values (for example a person who changes his or her surname several times). It is difficult to transform such attributes to a rectangular dataset in any meaningful way without letting the export tables grow considerably in size.

The BSS lacks many of the explicit variables that are necessary to carry out various analyses, demographic or other. To remedy this it is vital to retrieve relevant data directly from the sources. One obstacle however, is that event data from different sources may have a rather arbitrary format, except for a few properties such as names, birth and death dates. Because of this the event table has an initially unknown dimension and field structure. To ensure complete export of event data it is necessary to first scan all events in the sample to calculate table size and contents, resulting in more extensive use of resources and an event table with no fixed appearance.

As indicated in the figure, the next logical step would be to import the rectangular datasets directly into HPR. However, a better approach will be to implement the Intermediate Data Structure (IDS) as described in Alter, Mandemakers and Gutmann 2009. A major advantage of IDS is the principle known as entity attribute data model (EAV), where each record constitutes only one attribute and its corresponding value. This should largely solve some of the challenges described here. Before using IDS-formatted datasets in research, they must be converted one step further, for instance to a rectangular dataset for analysis. This will be accomplished by the applications generated within the IDS project as supported by the European Science Foundation. All problems will not disappear with this solution, but the end

user will have a better chance to control the outcome. Once an export has been made it will be possible to transform the data in several ways for different purposes.

It is recommended to pursue this solution further in order to obtain synergy effects together with the development of other components in the HPR and IDS projects. Some important benefits can be obtained, such as easier maintenance of the export module in BSS by using a standardized format. It should also be considered how to substitute the BSS import format and eventually the internal exchange format with IDS structures. This will require some rewriting of code in the import module of BSS and redesign of existing tools for converting original datasets, but despite a lot of work one will probably achieve some gains by having to deal with only one format for all data exchanges.

# Wiki web techniques (Lars Holden)

For the data from before 1920 we will use an open wiki database, denoted HPR-wiki available via the Internet. In addition to data before 1920, we will consider to include more recent open sources in the database, naturally within legal restraints. The database has the following objectives:

- Establish electronic links to all open information regarding historic persons with focus on the family relationship and places where the persons lived.
- Establish two-ways electronic links to primary sources when these cannot be established with the algorithms outlined above.
- Preferred internet site for cooperation between different organizations, projects, researchers and amateur genealogists in this area with many contributors and users.
- Establish data structure and routines that lead to improved quality and increased volume over time.
- Contribute to increased transcription from primary sources.

The open and closed databases (from before and after 1920) will be merged by copying the open database into the closed database using especially the 1910 census for identification of the persons between the two databases. The open database will be continuously extended by contributions from many persons and we will perform this copying regularly.

## Status

The first version of the Internet site is established at slekt.nr.no. The Norwegian Computing Center (NR) is responsible the administration and development. In a main project the National Archives will take over the responsibility for administration and maintenance. All partners in the project will contribute to the content of the database.

Our HPR-wiki is a PHP extension of MediaWiki. There are a large number of open source extensions and it is only necessary to make new programs that are closely connected to our special application. The main structure of HPR-wiki is established. The code is extended to be bilingual and the most important pages will be available both in Norwegian and English. Currently, May 2011, we are preparing the reading of the first data into the database.

## Management and maintenance of the database

In a main project RA will be responsible for the database. The project board with representatives for the main partners in the project will continue as an advisor for all development of the database.

Each partner of the project will be responsible for different parts of the development of the database: RA for maintenance and support in the main project and links to primary sources and to the closed database, RHD for maintenance and support in the pre-project, linking between different sources and several local databases, NR for software development, Folkehelseinstituttet for medical information, the universities for different research projects (e.g. urban mapping at UiB, emigration at UiS), NLI for links to local history and Busetnadssoge for several local databases. In addition, we will establish local groups of DIS-Norge and Landslaget for lokalhistorie that will supervise changes in the database in their region.

RHD or RA will be responsible for user administration and support and a group of super users that will supervise the development of the database. NLI has long experience with administration of historic information in wiki databases and is an important contributor for building up the user administration.

All changes will be identified with the time and person who made the change and it is easy to identify and remove e.g all changes made by the same person. Whenever possible we will establish two-way links to the (primary) source making it possible for users to verify that the data is correct online. This makes the database transparent. By following both families and places over time and adding more sources, we expect the quality to improve and the value of interest in the database to increase. There are almost unlimited possibilities to contributing to the database. We expect different users will focus on different topics from their own family, locality or region, ethnic group, well-known persons, transcription, quality control and spell checking. All interaction is performed online via the Internet.

## The page structure in the HPR-wiki

The page structure is illustrated in figure 2. The most important pages are:

**Person pages** have links to all sources referring to the person, parents, spouse(s) children and the family pages the person is represented, in addition to links to persons named in sources that may potentially be the same person. Person pages may be tagged to indicate that the person belongs to a particular group, e.g. priest. Then it will be easy to find all person pages for priests. The person pages may also have a text describing the person and the different events in the life of the person.

**Family pages** have links to all sources referring to the family, the persons pages for all persons in the family and the farm/house of the family.

**Place pages** are ordered hierarchically as geographic levels from the nation down to the farm or address level and are linked vertically. We will follow the changing organizations over time. When reading each census, we will establish the hierarchy of places that were used when organizing the census. At the lowest address level there will be a list of all members in the household living at the address for each census and links to the person pages. Hence, it will be possible to follow the persons living at the place during the time span of the censuses. All places will have GPS coordinates. This will help with identification and gives many possibilities in visualisation of the database. At a later stage this may be used in more advanced user interaction.

**Source pages** describe each source and how this source is handled. The source pages will in some cases be used in order to establish two-ways links between the source and the person. When necessary, there will be a table in the source page with one line for each entity in the source in order to ensure unique links. However, we do not believe this is necessary for the

censuses. This will ensure that there is a unique person in the database linked to particular data in the source.

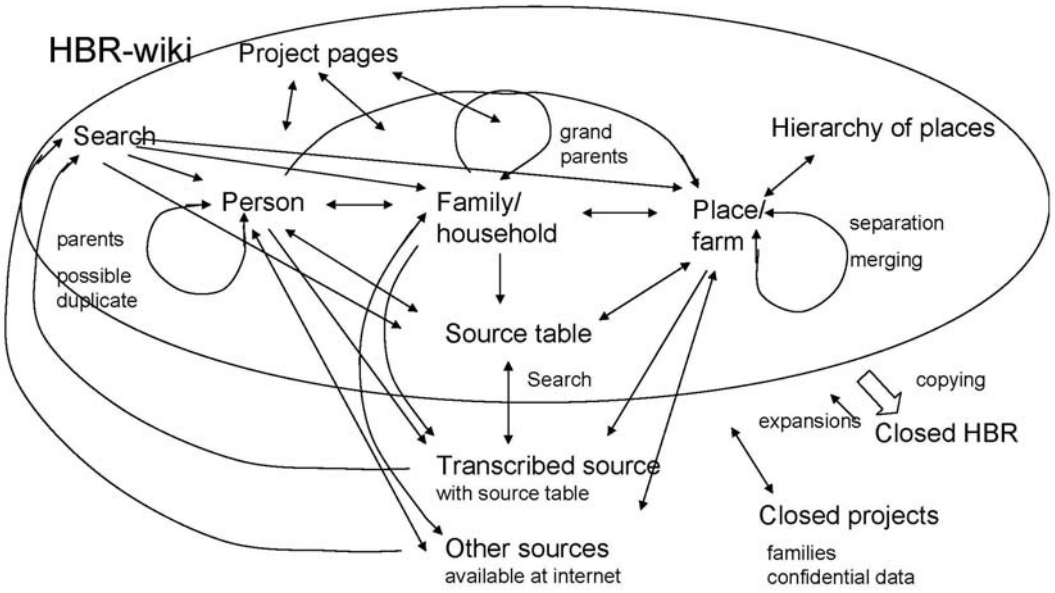



Figure 3: Single and two-way links between the different pages in HPR-wiki represented with single and double arrow. The most important pages are the person, family and place pages. There are links between these pages and to/from these pages to the sources. When pages are merged, links will automatically be moved to the merged page.

**Project pages** describe several different topics as regions, ethnic groups, persons with the same occupation, known persons and families. These pages will encourage a multitude of different users of the database. We will encourage persons working within the field to establish a project page. If the project has its own internet pages, we would like a link to this page from the HPR-project page in order to get more information and to honour the contributor and their work. We expect the project pages to have links to some of the other pages in the database, e.g. an important person in the family. The project page may also have large tables to e.g. the person pages of all members of the parliament in 1814.

Many pages will be established directly from census and church books. When a census is read, it a person page is established for each person, a family page for each family and a place page for each farm/flat with a list of the household and the hierarchy of places used in the organization of the census. Hence, the main challenge in the database is to merge pages between different sources. We will merge pages when we can document that the same person, family or farm/flat is established from different censuses or church books. Instead of merging there is a possibility to suggest a link between pages when we are not sure whether the two pages should be merged or the person making the link is not authorised to merge the pages. Similarly, both family pages and place pages from different censuses may be merged.

### Transcription

We expect that the open database will have many users. Many contributors would like to link "their" persons to sources that are not transcribed yet. We will encourage persons to contribute to transcription and allow transcription of single entities from the sources.

We will establish pages in the open database for each source. On this page it will be possible to write the transcribed text and a link to the scanned pages. We expect most users will only transcribe "their" person, but expect others to want to transcribe entire pages or several pages from the same source. The scanned image will be shown together with the transcribed text, making it easy for users of the database to control the transcription. This transcription is performed online.

We cannot expect to have identical core data (name, date/year of birth etc) in the different sources. When data differ, we use the following priority rules:

1. Manual set data
2. Church book in priority baptism, confirmation, marriage, burial
3. Census, where newer has higher priority
4. Other sources.

This list is based on the expected quality of the data.

## Migration and representativity

A primary rational in building a *national* population database is to improve understanding of migration inside Norway. Internal pointers in the database allow researchers to trace migrants, whether they crossed administrative borders or not, by using the information about names, birth places, age, and even birth dates that are found in most registry sources. By being able to follow migrants inside Norway the representativity problem caused by in- and out-migration in local community or regional databases is minimized. To handle the migration of immigrants and emigrants, other methods must be used to estimate when they moved in and out of Norway.

Since the registration of immigrants was not thorough until World War I, it is methodologically fortuitous that they were relatively few in number. There were only 70,000 immigrants in 1900 and, as a group, they did not exceed 3 percent of the population until the 1980s. Immigrants were also a relatively homogeneous group in Norway until guest workers arrived in the 1960s. The vast majority were Swedes and ethnic Finns who arrived in the late nineteenth century. Immigrants can be identified in censuses and church books whenever they were involved in vital events; however, these records often lack information about when they first crossed the border and about intervals spent back in countries of origin. Return emigration from America brought US-born children to Norway. Returned emigrants born in Norway are listed in special forms in both the 1910 and 1920 censuses.

Quantitatively, non-returning emigrants pose a more serious problem for the historical population register. After Ireland, Norway had the world's most intensive export of people across the Atlantic during the late nineteenth and early twentieth centuries. Traditionally, the official figure has been set at 800,000 emigrants to North America during this period of mass emigration. However, recent research indicates that at least 50,000 more overseas migrants should be added; few emigrants leaving Norway before the American Civil War were registered and many who simply jumped ship while in US harbors were not included in official totals. The Migration Forum at the University of Stavanger is working to collect the relevant source material, also to be used in the HPR. Most of the general emigration protocols kept in main ports from the late 1860s onward have been transcribed. Emigration records often

recorded the latest place of residence rather than birth place, but this identification problem will diminish as we expand the longitudinal data register. Based on his investigation of emigrants from Tinn parish in Telemark Andres Svalestuen stressed the diminishing value of the church records from the end of the 1870's, stating that the emigrant protocols must be considered the more reliable source from that time on. Even so, the church records still have some value as a supplement.

Absolute and relative numbers of emigrants not mentioned in the church records for Tinn:

| Period | Number of emigrants | Not mentioned in church records | In percent (approx.) |
|---|---|---|---|
| 1871-80 | 160 | 23 | 15% |
| 1881-99 | 660 | 152 | 25% |
| 1900-07 | 210 | 109 | 50% |
| Sum 1871-1907 | 1030 | 284 | 30% |

A major source of additional information about immigrants to the United States is the compilation of passenger lists from the period 1820 to 1959 kept in the US National Archives. These lists contain detailed nominative information organized by some one hundred immigration harbors in the U.S.; Canada is included from 1895 onward. Ship captains were required by law to deliver lists of all passengers to the customs officials; over time, this reporting shifted to immigration authorities. These lists have been microfilmed, and there are up to 6,000 rolls of film for each harbor. Transcription from this unwieldy material is kept in databases by Ancestry, Castle Garden Clinton National Monument, the American Family Immigration History Center and others—altogether, tens of millions of records. Among the lists from ships crossing the Atlantic from harbors outside Norway, we expect to find many Norwegian immigrants who were not registered in our domestic emigration protocols. Prior to the Civil War, these passenger lists are the main inventories of Norwegian immigrants. Spot tests reveal that there are more than 100,000 persons with the last name "Andersen" in these lists; since birth place data are not very specific, it will be quite a challenge to cross-check US immigration lists against the Norwegian emigration protocols.

Similar challenging opportunities exist with the muster rolls of crews of Norwegian ships; many contain notes about those who jumped ship in foreign harbors. It, too, is unwieldy material, but it contains detailed information adequate for identifying these illegal emigrants. Thus, while the majority of emigrants will be identified in the emigration protocols of the harbors from which they left Norway, the final 10 percent will require considerably more time and effort to identify than those with more orthodox departures.

The tracing of emigrants and returned emigrants is more fully explained in an article submitted for publication in the journal *Heimen*. This is available upon request and also exists as an English version. Representativity is also treated in the paragraphs on record linkage in this report.

# ID numbers (Lars Nygaard)

The open, public part (1800-1919) of the HPR, the HPR-wiki, will be established in close connection with the website Digitalarkivet (The Digital Archives - DA) of the National Archives of Norway (NA). DA publishes digitised source material to the public, both scanned and transcribed material. The databases and the software of the DA are in 2010-2011 in a process of being redeveloped. In this process every single source (e.g. the 1910 census for a particular municipality or the baptism list 1878-1892 from a particular parish), every scanned page, and every transcribed record is given a unique ID. These IDs are included when the basic source material is transferred to HPR, and the Norwegian Historical Data Centre (RHD) also plans to use these IDs in their source versions specially processed for research purposes.

The IDs are planned to be the discriminating part of standardised web-links (URLs) between different versions or occurrences of the same sources, records or persons on the three websites mentioned. For example, a link from a person mentioned in a census in DA to this person's linked life course in HPR-wiki, a link from an occurrence of a person in HPR-wiki to the scanned or transcribed version of the corresponding page in the source, or a link from a source mentioned in HPR-wiki to the coded version of this source downloadable from NHDC.

To be utilised in URLs the main properties of the IDs are that they are unique and permanent. When a locally unique ID is used in an URL based on the URN standard (Uniform Resource Name), the URL will be globally unique (and permanent) and will work as a stable and reliable reference to the page or record in question. An important condition for keeping the IDs permanent is to avoid putting meaningful and thus possibly erroneous information into them, because erroneous information will require correction (change). Our IDs consist of a two letter prefix, which tells what kind of "object" the ID identifies, a two digit institution code to "guarantee" the uniqueness of the ID, and 12 digits which in practice constitute a serial number. The IDs are supposed to be mainly computer internal.

A "logical" (as opposed to physical) source gets an ID called KID (Norwegian abbreviation for "source ID"). The corresponding URL leads to a source presentation page in DA with links further to different digital versions of the source. Each transcribed record from a particular source gets a record ID with a certain prefix, *pf* in "person-source IDs" (PKID) for person records from a census, *bf* in "property-source IDs" (EKID) for residence records from a census, etc. The corresponding URL leads to a page in DA where the record is displayed in its environment. For example, a person record from a census is displayed together with records about the residence and the other persons in the household.

The *linked persons* (life courses) in HPR-wiki are not given a permanent ID from the beginning. Instead of creating a new ID when the first two records are linked together, one of the two PKIDs is chosen by priority rules as the PID (Norwegian abbreviation for "person ID"). Thus, as long as the record linking is going on, the PID will change when a record with a higher priority is linked in, or if the record with the highest priority is de-linked from the life course. Anyway, URLs to HPR-wiki containing any of the PKIDs linked to the same person, will lead to the same page in the wiki. The arrangement for *linked properties* in HPR-wiki is analogous.

Here is the priority list for choosing a PKID as the PID for a linked person:

1) PKID in the 1910 census (1910-12-01)
2) PKID in burial list recording deaths before 1910-12-01
3) PKID in baptism list recording births after 1910-12-01
4) PKID from list of out-migrants before 1910-12-01
5) PKID from list of in-migrants after 1910-12-01
6) PKID from another census (the newest with highest priority)

7) PKID from another churchbook record (the newest with highest priority)
8) PKID from another source (the newest with highest priority)

The first five sources in the list are non-overlapping. Notice that this list differs from the priority list of quality of core data for each person. In the priority list for PID we have prioritized sources that are suitable for linking between the different parts of HPR and sources that are non-overlapping.

## Variables and registers in the HPR

An important distinction is made between significant demographic variables contained in the historical population register proper, and complementary information found in the transcribed sources which go into a separate historical data register:

***Historical Population Register (HPR)***
Unique PersonID
Timestamp (from-to dates)
Anonymous link to Central Population Register
Anonymous link to closed register
Source type
Source reference
Household number
First name
Last name
Sex
Birth date/ Birth year
Birth place
Birth place code
Baptism date
Death date
Marriage date
Marriage place
Family status encoded
Marital status
Momloc (Mother's location in household)
Poploc (Father's location in household)
Spouseloc (Spouse's location in household)

***Data about domicile:***
Domicile in Parish/Municipality/Province
Domicile name / Street name
Domicile number
Houses lived in
Number of rooms/kitchen/bath
Household number
Rent
Data on agriculture

***Historical data register:***
Unique PersonID
Timestamp (from-to dates)
Source type
Source reference
Age
Confirmation date
Related to husband by ancestry?
Number of children
Number of children alive
Occupations
Provider's occupations
Out of work?
Employer
Employs others?
Able to work?
Education
de jure or de facto status
Usual / presumed temporary address
In- / outmigration
Citizenship
Religion
Disabilities
Sleeps in which building
Ethnicity
Language
Comments

# Legal issues and solutions

The historical population register will be divided into three separate parts: the open part (until November 1920), the closed part (December 1920 to 1964) and the Central Population Register (DSF) from 1964 until now. The three parts will be administered in different institutions and have different restrictions for researchers' access to the contents. The background for splitting the HPR chronologically is mostly juridical. The Law of Statistics protects the disclosure of identifiable information assembled by the Norwegian state (particularly the censuses) for one hundred years. The Law of Publicity similarly protects the church records for 60 years, but sensitive issues such as adoptions are protected for 100 years. Practices are also important since the church protocols are sent to the Regional archives after 80 years, while the censuses are kept in the archives for the open and closed periods (until 1960). Information from the municipal censuses may be disclosed after 60 years, but little of this material has been digitized.

## Open database

*Responsible: The National Archives and the Norwegian Historical Data Centre*
The open part contains records for the period until November 1920, but no data from the census taken in December that year. In 2020 the open part will be extended until 1930, and we expect to extend the database regularly as further censuses are released according to the century rule. In addition lists of deceased persons will be released for later decades, but not containing cause of death.

Stand-alone graphical elements from more recent source material, e.g names, occupations, birthdates from the 1920 to 1950 censuses may be made public after the columns and rows have been cut to pieces with graphical techniques. These isolated elements can be transcribed by the public via the Internet or in low-cost countries. The keys to combine the graphical and transcribed elements will be kept secret by the National Archives.

## Closed database

*Responsible: The National Archives*
The closed part will contain records for the period between the open part and the start of the Central Population Register, i.e from December 1920 to 1964. The significant source material will be the national censuses, church registers and vital records sent to Statistics Norway, all material kept by the National and Regional archives. Access to this database will be given by the National Archivists for researchers who can document restricting the use only to statistical purposes according to the Law of Statistics.

The key register to link the closed and the open parts will be created by the The National Archives and the Norwegian Historical Data Centre as a joint effort. Data is transferred from the closed to the open database every tenth year as described above. Researchers requesting data from the Closed Database or the Central Population Register will need to report their need for a copy to the Data Inspector through their respective ombudsmen.

## The Central Population Register

*Responsible: Statistics Norway*
Statistics Norway will continue to run their version of the Central Population Register (CPR) as usual. The National Archives will receive a copy containing the population as of 1964 according to ordinary rules of archival transfer. Linking the closed database and the Central Population Register will be a joint effort of Statistics Norway and the National Archives. The

key register to link the closed database and the CPR will be kept as protected files by both the abovementioned partners. Researchers can access the as of 1964 version of the CPR by application to the National Archivist, while applications for more recent versions will be handled by Statistics Norway to be approved by the Data Inspector.

## Research possibilities

A longitudinal database encompassing the Norwegian population in the nineteenth and twentieth century will open up new terrain in disciplines such as history, economics, demography, social medicine and sociology. The censuses, parish books and other nominative sources include extensive information on demographic and social structure that can only be fully utilized through the creation of an integrated longitudinal HPR. The nineteenth and twentieth century form a critical period in the study of medical advances, fertility decline, urbanization, international migration, household composition and occupational structure. The database will allow statistical modelling on a wide range of topics that have not been covered by census publications or have been incompletely tabulated. Even more important is the potential for longitudinal and multilevel multivariate analyses opened up by the availability of the database. A longitudinal database will constitute an invaluable resource in its own right by enhancing the value of both previous and current historical microdata samples. Used in combination, these microdata will constitute an invaluable resource for studying the development that led up to our contemporary society. The paragraphs that follow sketch only a few of the most obvious research applications of the HPR.

*Public Health.* For many common diseases it is thought that the causes reside in an interplay between genetic factors and environmental effects. This implies that persons who develop the disease, for instance cancer, heart disease or chronic rheumatic diseases carry certain predisposing genes and that disease occurs when one is additionally exposed to certain environmental stimuli, such as an infection or a component of the nutritional intake. Recent genetic studies suggest that many chronic diseases are genetically heterogeneous. Very many different genes may give rise to the same phenotype. Some of these variants may be regarded as rare mutations only occurring in one or a few families. In order to disentangle the vast genetic variation in today's available DNA data and tomorrows more detailed sequencing DNA data we need to have the study population composed into families, as can be done with the new HPR. The familial disentangling of the genome together with information from linked cohorts and health registries will enable us to understand the genetic basis of diseases, and subsequently understand the biological mechanisms that will lead to better prevention and treatment.

*Fertility transition.* During the late 19$^{th}$ and early 20$^{th}$ centuries Norwegian women and men started deliberate fertility limitation. A historical population register will allow the study of differential fertility patterns in this critical period of demographic transition, to assess the importance of such factors as occupational class, region, literacy, local economy, size of locality and family structure. Study of these shifts in population structure has the potential to enhance our understanding of ongoing demographic change in the contemporary developing world. Aggregates do not allow controlling for individual-level socioeconomic characteristics, and the fertility pattern of families cannot be followed through the reproductive period of the mother. With a longitudinal database we will be able to study the starting, spacing and stopping behaviour of families directly. Thus, the database will allow a new and more sophisticated generation of comparative studies of the first demographic transition.

*Industrialization and economic developments.* By the nineteenth century most of Norway was affected by industrialization. The HPR will allow unprecedented opportunities to explore economic structures within Norway and with other nations during this and later transitional periods. For the first time, we will be able to follow people at the individual level over generations throughout the country, with consistently coded occupational and other variables. This will allow comparative analysis of the careers of persons and families, and investigation of the geographic organization of economic activity. For example, the database will allow a comparative national and possibly international investigation of maritime communities.

*Household and family composition.* Political theorists, sociologists and historians have been debating the relationship between industrialization and the family. Some studies argued that the harsh economic conditions of early industrial capitalism strengthened the interdependence of family members and led to a high frequency of complex households (Anderson 1971; Hareven 1978). In recent years, numerous national and regional studies of family composition in the late nineteenth century were based on population samples, but few incorporated community-level economic measures (Sogner 1990; Gunnlaugsson and Garðarsdóttir 1996; Ruggles 2000). Thus, there is presently little agreement about national similarities and differences in family and household composition in the late nineteenth century. In order to understand the transition between family forms, such investigations need data sets where households can be followed over time. This type of analysis will be particularly conducive if longitudinal databases from several countries can be combined and the context of changing family forms can be related to their setting through multi-level analysis. An example can clarify the need for longitudinally organized source material: In a family and aging oriented study the main research question is to find factors that motivate parents and their grown-up children to live in the same household. With access to a cross-sectional source, we can study the differential characteristics of aged people who live with or without their children, for instance the extent to which this depends on gender and age. But differential characteristics of the relevant grown-up children cannot be analyzed, because the children who are not living with parents are not linked to them. As soon as we have links to persons in at least one earlier census, we can study differences between children who do or do not co-reside with their aged parents. Thus, only research based on longitudinal data will be able to contribute more than half-way answers to one of the longest-lasting debates in social history research, the question of the extended family structure (Jåstad 2009). This paragraph is but one example of why the significance of *family networks* has become so topical for understanding social and demographic history among researchers internationally. In addition, kinship is the pivotal point in many fields of human interaction, from the family firm via infant and child care to chain migration.

*Name studies.* The culture of changing name traditions for men and women over two centuries can be related to different family types, occupations and geographic areas in longitudinal data. In a longitudinal database names can be studied more source-critically and dynamically because onomatologists get access to information about name forms for the same person at different points in time. (Cf concluded NFR project 164186 which standardized all names in the censuses 1801-1900). During the pre-project we have extended the standardization of names from the censuses to the 1910 census and the church registers. We also see a potential for using the HPR in studies of dialects.

*National and international migration.* The late nineteenth and early twentieth century saw geographic population movements on an unprecedented scale. The massive emigration to

America profoundly shaped both the receiving and contributing countries. Many of the emigrants remained only a few years before returning to their homelands, often bringing home money and always bringing new ideas and experiences (Gjerde 1992; Thorvaldsen 1998). The HPR will be a rich resource for the study of migration history, and will open a new window on the implications of national and international population flows. Particularly in combination with the BSS databases the HPR opens up for research following cohorts and families over long periods with manageable work spent on the record linkage process, also taking local gross migration into account.

*Social sciences.* The time window 1960 to the present, for which there is rich access to data on population related issues, cover only the latter phases of several processes involving major changes to our society. By expanding this window we can learn considerably more about the reasons for and consequences of e.g increased female labour participation and higher levels of education. When it comes to aging, it is evident that this process cannot be satisfactorily studied without data covering several decades. An important feature is the ability to follow families and households over several generations. This allows for a multitude of new approaches to the study of family and household organization, social mobility and other inter-generational processes such as transmission of education investments. Even when the research interest lies exclusively in present day phenomena the access to family and generational data from the HPR can be extremely useful. For example we can shed light on the role of family background in connection with studies of labour market outcomes, or social differences in mortality and health outcomes. By combining the national HPR with the local BSS registers such social research questions can be studied in even greater detail over up to three centuries.

*Challenges in information technology.* The database will be very large, possibly 10-50 million pages. The HPR-wiki may well become the largest wiki in the world. As a comparison, the English version of Wikipedia has 3,5 million pages and family-genealogy wiki www.werelate.org has 2 million pages. As the database increases, we must ensure that the database is able to handle the amount of data and that the search engines will work properly.

There is also a challenge in the social aspect with a large number of contributors. How do we encourage many contributors with different competence and at the same time maintain or improve the quality of the database? There is much experience with building wiki databases, but the HPR has some unique features. It will be very large and most of the pages will be produced automatically. The contributors are encouraged to merge pages and add content to each page. We also want contributors to add new sources of high quality and link them to the other pages in the database. We expect there to be a large interest in the database, but from persons with varying competence both in genealogy and information technology. We will gradually increase the number of users gaining knowledge as the number of users increase. We will try to establish a group of super users supervising the database and local groups who monitor changes in their regions.

*Automatic methods for transcription* One of the major challenges in building the database is the large amount of sources that are not transcribed. This challenge will be addressed by both encouraging to online transcription and by automatic recognition methods. The first stage of the automatic recognition methods is to identify each cell in a scanned page. The information in each cell is typically a name, place, occupation, year etc. In the newer sources, it may be easy to identify the cell in a scanned image, but in the older church

books this may be a challenge. This first stage is important for the user interface and control of the manual transcription. This first stage is tested in the pre-project with success. The second stage in automatic transcription is to identify the correct content of the cell, i.e. the name, place etc. If a cell has a limited number of possible values, e.g. married/unmarried or a digit, the automatics method works much better. We expect that automatic methods will be able to identify the content in some of the easy cells, e.g. position in family (father/mother/ son/ daughter), married/unmarried and religion. If we use a list of possible names then also some names may be recognized from the hand written text. But if we can identify a possible candidate for the name from another source, the automatic method will be much more successful. Other sources may be census, church books, list of properties etc. This will be particularly important for the closed database where the volume is larger, quality of sources are better, there are several other sources available and it is more difficult to use volunteers in the transcription. We may use that many persons live at the same place for decades and hence occur repeatedly in several of the censuses taken during the last 100 years. The pre-project has described several different methods on how this may be performed in the report by Eikvil, Holden and Bævre (2010). The project thus has research competence in automatic transcription.

# Bibliography


Ran Abramitzky, Leah P. Boustan, Katherine Eriksson, Univ of California, (Economics), *Measuring selectivity and returns in the age of mass migration.* SSHA 2009 / http://scholar.google.com/scholar?q="Productivity and Migration: New Insights from the 19th Century"

Alhaug, G. 2008. Resultatrapport. Prosjektet. Namnevariantar – eit problem ved bruk av digitaliserte folketeljingar for 1700- og 1800-talet. Memo, University of Tromsø.

Alsvik, Ola: *Local history in Norway*. At the website of The Norwegian Institute of Local History: http://www.localhistory.no/local-his.html (read May 2nd 2011).

Alter, G, K. Mandemakers and M. Gutmann. 2009. Defining and distributing longitudinal historical data in a general way through an intermediate structure." *Historical Social Research / Historische Sozialforschung* 34: 78-114.

Andersen and Hynnekleiv (2007). "Hospital-treated psychosis and suicide in a rural community (1877–2005). Part 2: Genetic founder effects." *Acta Psychiatrica Scandinavica* 116: 20-32.

Anderson, M. *Family structure in nineteenth century Lancashire*. London, 1971.

Anderson, M. 1977. Some problems in the use of census type material for the study of family and kinship systems. *Time, space and man. Essays on microdemography,* ed. S. Søderlund. Umeå 69-80.

Backer J.E., 1947-8, "Population statistics and population registration in Norway. Part I-II. The vital statistics of Norway: an historical perspective, *Population Studies*, 1(2) / 2(3), pp. 212-226; 318-338.

Breivik, T. 2005. Arkivaretikk og brukeretikk. Bruksregler for offentlige arkiver i et etisk perspektiv " *Heimen* 42: 99-109.

Bull, H. H. 2005. Deciding whom to marry in a rural two-class society: Social homogamy and constraints in the marriage market in Rendalen, Norway, 1750–1900. *International Review of Social History* 50: 43-63.

Bull, H.H. 2006. Marriage decisions in a peasant society: The role of the family of origin with regard to adult children's choice of marriage partner and the timing of their marriage in Rendalen, Norway, 1750-1900. Faculty of Humanities, University of Oslo. nr 268.

Christen, P. 2008. Febrl: A freely available record linkage system with a graphical user interface. *Proceedings of the second Australasian workshop on health data and knowledge management*, January 2008, Wollongong, NSW, Australia, cf https://sourceforge.net/projects/febrl/

Coale and Watkins (1986). *The Decline of fertility in Europe: the revised proceedings of a conference on the Princeton European Fertility Project*, Princeton Univ Press.

Drake, M. 1969. *Population and society in Norway 1735-1865*. Cambridge University Press.

Dyrvik, S. 1983. *Historisk demografi. Ei innføring i metodane*, Bergen.

Line Eikvil, Lars Holden, Kåre Bævre, Automatisk metoder som hjelp til transkribering av historiske kilder. SAMBA/44/10, Norsk Regnesentral, 2010.

Engberg, E. Karesuando (2007). 1900-talsmaterial. Rapport från ett utvecklingsprojekt vid Demografiska databasen i samarbete med Landsarkivet Härnösand och Statistiska Centralbyrån. Umeå, Demografiska databasen**:** http://www.ddb.umu.se/Karesuandoprojekt-Rapport2008.pdf

Engelsen, R. 1983. Mortalitetsdebatten og sosiale skilnader i mortalittet. *Historisk tidsskrift* 2: 161-202.

Erikstad, M. and G. Thorvaldsen. 2006. Statistikk basert på individdata fra folketellingene, nye muligheter. *Heimen* 43: 41-54.

Fure, E. 2000. Spedbarndødelighet og sosiale forskjeller i Asker og Bærum 1814-1878. En metode for studier på individnivå. *Heimen* **37**: 293-304.

Fure, E. 2000. Interactive record linkage. The cumulative construction of life courses." *Demographic Research* 3.

Geneology. 2005. The Genealogical Society of Utah as a data resource for historical demography. Retrieved 2 Nov, 2007, from http://www.genuki.org.uk/big/Linking/index.html.

Gjelseth, M. 2000. *Relasjonsdatabaser som verktøy i en historisk-demografisk studie*. Historisk institutt. Oslo: Universitetet i Oslo.

Goeken, R et al. 2011. New Methods of Census Record Linking. *Historical Methods. Vol*. 44, Iss. 1; p. 7

Gunnlaugsson, G.A. and O. Garðarsdóttir. 1996. Naming practices and the importance of kinship networks in early nineteenth-century Iceland. *The History of the Family* 4: 297-314.

Gutmann, M. P., ed. 1992. Historical Methods: Theme issue on record linkage.

Helgheim, J. "Sjeleregistre som kjeldemateriale". *Heimen* 1977-1: 269-274.

Hogan, D. P. K., David I. (1985). " Longitudinal Approaches to Migration in Social History. *Historical Methods*. 20-29.

Janssens, A. 1993. Family and social change. The household as a process in an industrializing community. Cambridge.

Johansen, H. C. 1999. Urban social and demographic reconstitution. The case of eighteenth century Odense." *History and Computing* 11: 115-128.

Jåstad, H. 2009. Northern cohabitation across generations. Paper at Aging Seminar, Umeå University, March 2009. Forthcoming as article 2010.

Kjelland, Arnfinn 1990: EDB og gards- og slektshistorie. In *Heimen* hf. 4: 227–241. Also: http://www2.hivolda.no/ahf/historie/edb_hist/edbstatus90.htm (read May 2nd 2011).

Kjelland, Arnfinn 1987–96: *Bygdebok for Lesja. Gards og slektshistorie for Lesjaskogen* (vol. 1, 1987), *Gards- og slektshistorie for nørdre del av Lesja hovudsokn* (vol. 2, 1992), *Gards- og slektshistorie for søre del av Lesja hovudsokn* (vol. 3, 1996). English introduction, cf http://tilsett.hivolda.no/ak/Lesja/engelsk-intr.html (read April 2011).

Kjelland, Arnfinn 1996: One Tenant, Several Landlords. The Land Tenure System of Norway until ca. 1800. Paper for the pre-conference «Land, Labour and Tenure: The

Institutional Arrangements of Conflict and Cooperation in Comparative P erspective». The University of Leicester, England, August 21–24, 1996. Published at: http://www2.hivolda.no/ahf/historie/tilsette/ak/Tenant_landlord.html (read April 2011).

Kjelland, A. 2008. The Norwegian tradition of farm- and genealogical history as a source for microhistorical research. Paper presented at the Center for Microhistorical Research, Reykjavik Academy. http://tilsett.hivolda.no/ak/KjellandPaperMicrohistoryReykjavik2008.pdf

Lie, E., and S. Boquist. 2001. *Faktisk talt : statistikkens historie i Norge*. Oslo: Universitetsforlaget, 2001.

Jåstad, Hilde. Forthcoming. "Northern coresidence across generations. Northern Norway during the last part of the nineteenth century." Dr. Art. Degree thesis. University of Tromsø.

Leeuwen, Maas, et al. *HISCO : historical international standard classification of ccupations*. Leuven, Leuven University Press, 2002.

Miller, R. and G. Thorvaldsen. 1997. Beyond record linkage: Longitudinal analysis of turn-of-the-century inter-urban Swedish migrants." *History and Computing* 9: 106-121.

Nilsdotter Jeub, U. 1993. *Parish records. 19th century ecclesiastical registers.* Umeå. Extract in Swedish: http://www.ddb.umu.se/digitalAssets/26/26356_kyrkbcker_historik_091214.pdf

Nygaard, L. 1985. *Fra historiske kilder til persondatabase*. Oslo: Universitetet i Oslo.

Nygaard, L. 1992. Name standardization in record linkage: an improved algorithmic strategy. *History and Computing:* 63-74.

Quass, Q. 2010. weRelate : About. Accessed 22 February 2010. http://www.werelate.org/wiki/WeRelate:About

(The) Research Council of Norway**.** (2008). *Evaluering av norsk historiefaglig forskning. Bortenfor nasjonen i tid og rom: fortidens makt og fremtidens muligheter i norsk historieforskning.* http://www.forskningsradet.no/en/News/Call+for+greater+international+focus/1211259699183 2008.

Roberts, E. 2003. The North Atlantic Population Project: An overview." *Historical Methods* 36: 80-88.

Ruggles, S. 2000. Living arrangements and well-being of the elderly in the past." Presented at Population Aging and Living Arrangements of Older Persons: Critical Issues and Policy responses." Population Division, United Nations, New York, February 2000.

Ruggles, S. 2003. Linking historical censuses: A new approach." *IMAG International Microdata Access Group workshop*, http://www.nappdata.org/imagpapers/ruggles.pdf.

Sabean, David Warren 1998: *Kinship in Neckarhausen, 1700–1870*. Cambridge University Press.

Sarkar, S and Patricia Kelly Hall (2011) Mapping NAPP: Examples of Spatial Analysis of Nineteenth-Century Migration in the North Atlantic Countries Using NAPP Data. *Historical Methods***.** Vol. 44, Iss. 1; p. 25

Schurer, K., et al. 2003. *The Victorian Panel Survey – A scoping study for the ESRC.* Colchester: UK Data Service, University of Essex.

Sogner, S. 1979. Folkevekst og flytting : en historisk-demografisk studie i 1700-årenes Øst-Norge. Oslo: Universitetsforlaget.

Sogner, S. *Far sjøl i stua og familien hans: trekk fra norsk familiehistorie før og nå*. Oslo, Universitetsforlaget, 1990.

Sogner, S., et al. 2002. The rural reward. Infant mortaltiy in Norway during the demographic transition. A case study." *Historical Studies in Mortality Decline*. W. Hubbard, et al. (eds). Oslo: Det Norske Videnskaps-Akademi / Novus forlag 3**:** 79-95.
Sogner, S. and G. Thorvaldsen. 2002. Surnames as proxies for place of origin in the 1801 census for Norway. *Scandinavian Population Studies*. J. Carling, (ed). Oslo. 251-265.
Soltvedt, K. 2004. Folkeregistre og personnummersystemer i Norge fra 1905 til 2001. *Folketellinger gjennom 200 år*. K. Soltvedt. Oslo: Statistisk sentralbyrå**:** 159-189.
Sunde, R. 2001. Vikjer ved fjorden - vikjer på prærien. Ein demografisk-komparativ studie med utgangspunkt i Vik i Sogn. *Historisk institutt*. Bergen: Universitetet i Bergen.
Szreter, S. 2007. The right of registration: Development, identity registration, and social security - A historical perspective. *World Development* 35: 67-86.
Sørumgård, Ole Martin and Arnfinn Kjelland 2003: *Putting a main part of Norwegian Local History – the old farm- and genealogical history genre – into the Computer*. Paper presented at the XVth Conference of the International Association for History and Computing, University of Tromsø 8th of August 2003: http://tilsett.hivolda.no/ak/BSS/Paper_AHC_2003.htm (read April 27th 2011).
Thorvaldsen, G. 1995. The encoding of highly structured historical sources." *Computers and the Humanities* 28: 301-305.
Thorvaldsen, G. 1995a. Migrasjon i Troms i annen halvdel av 1800-tallet. En kvantitativ analyse av folketellingene 1865, 1875 og 1900. Registreringssentral for historiske data. Universitetet i Tromsø.
Thorvaldsen, G. 1995b. Longitudinal sources and longitudinal methods - Studying migration at the Stockholm Historical Database. *Swedish urban demography during industrialization*. A. T. Brändström, ed. Umeå: Umeå universitet.
Thorvaldsen, G. 1996. Håndbok i registrering og bruk av historiske persondata. Oslo: Tano Aschehoug.
Thorvaldsen, G. 1998. Historical Databases in Scandinavia. *The History of the Family. An International Quarterly* 3: 371-383.
Thorvaldsen, G. (1998). "Marriage and Names Among Immigrants to Minnesota." *The Journal of the Association for History and Computing* **1**(2).
Thorvaldsen, G. 1999. *Databehandling for historikere*. Oslo, Tano Aschehoug.
Thorvaldsen, G. 2006. Away on census day. Enumerating the temporarily present or absent. *Historical Methods* 39: 82-96.
Thorvaldsen, G. 2007. An international perspective on Scandinavia's historical censuses." *Scandinavian Journal of History* 32: 237 - 257.
Thorvaldsen, G. 2008. Fra folketellinger og kirkebøker til norsk befolkningsregister. *Heimen* 45: 341-359.
Thorvaldsen, G. (2011). "Using NAPP Census Data to Construct The Historical Population Register for Norway" *Historical Methods* 2011/1.
Vick, Rebecca, Lap Huynh, Thomas Lenius. 2009. "The Effects of Name Standardization in Historical Census Record Linkage: Evidence from the United States and Norway." *Historical Methods* 2011/1.
Wikipedia. 2010. deCODE genetics. Checked 22 February 2010. http://en.wikipedia.org/wiki/DeCODE_genetics.
Winge, Harald 1995: Local History. In Hubbard & al.: *Making a Historical Culture. Historiography in Norway*. Scandinavian University Press: pp. 240–60.
Wrigley, E. A. and Roger.Schofield. 1973. Nominal record linkage by computer and the logic of family reconstitution. *Identifying people in thepPast*. E. A. Wrigley, ed. London.
*For further record linkage references, cf http://www.recordlink.org*

| **Family Reconstitution Type Databases ('bygdebok'), fulfilled and ongoing projects** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | **Sum:** | **302082** | **55,4** | **13,69** | |
| Pr.no. | County | Municipality | Type database | Responsible for family reconstitution | Starts (year) | Ends (year) | Number of linked persons by fall 2010 | Used FTEs by 2010 | Remaining FTEs 2011-19 | Notes |
| 1 | Møre og Romsdal | Sula | BSS | Ole Martin Sørumgård | c 1600 | 2008 | 28757 | 2,4 | 0,00 | Completed 2008 |
| 2 | Møre og Romsdal | Volda | BSS | Olav Myklebust | c 1600 | 2010 | 42896 | 6 | 2,00 | Ongoing (completed 2011), may be about 1000 more persons |
| 3 | Møre og Romsdal | Herøy | BSS | Per-Ståle Moltu | 1900/1920 | 2010 | 16785 | | 2,00 | Ongoing until 2012, may be about 9-10.000 more persons |
| 4 | Møre og Romsdal | Haram | BSS | Stein Arne Fauske | c 1600 | 2010 | | | | In initial phase |
| 5 | Hedmark | Stor-Elvdal | BSS | Håvard Kongsrud | c 1600 | 2010 | 4922 | | 1,14 | Ongoing |
| 6 | Hedmark | Vang | BSS | Ole Jacob Tomter | c 1600 | 2010 | 1934 | | 3,70 | Ongoing, special version of program |
| 7 | Hordaland | Ullensvang | BSS | Marta Gjernes | | 2013 | n.a. | | 1,80 | Just starting, less than 1000 linked persons |
| 8 | Troms | Bardu | BSS | Elin Torsetnes | c 1600 | 2010 | 1709 | | 2,05 | Ongoing |
| 9 | Akershus | Høland og Setskog | BSS | Frode Myrheim | c 1600 | 2010 | 35779 | | | Ongoing, special version of program |
| 10 | Nordland | Bindal | FAMREK | Håvard Sylten (deceased) | c 1600 | 2010 | n.a. | | | |
| 11 | Sør-Trøndelag | Oppdal | FAMREK | Odd Magne Mellemseter | c 1600 | 2010 | n.a. | | | Ongoing |
| 12 | Hordaland | Fjell | FAMREK | Halvor Skurtvei | 1728 | 1943 | 15000 | | | Completed |
| 13 | Nordland | Øksnes | dBase/Borgos | Johan Borgos | c 1610 | 1950 | 21800 | 6 | | Completed |
| 14 | Nordland | Sortland | dBase/Borgos | Johan Borgos | c 1610 | 1950 | 31000 | 8 | | Completed |
| 15 | Nordland | Andøy | dBase/Borgos | Johan Borgos | c 1610 | 1950 | 30000 | 8 | | Completed |
| 16 | Nordland | Bø i Vesterålen | dBase/Borgos | Johan Borgos | c 1610 | 1997 | 36300 | 11 | | Completed |
| 17 | Nordland | Saltdal | dBase/Borgos | Guri Solheim Clark | c 1610 | 1940 | 17900 | | | Completed |
| | Overlap persons in database no. 13-17: | | | | | | -15000 | | | |
| 18 | Nordland | Hadsel | dBase/Borgos | Johan Borgos | c 1610 | 1930 | 32300 | 14 | 1,00 | Ongoing, may be about 10.000 more persons |
| 19 | Troms | Astafjord | dBase/Borgos | | | | n.a. | | n.a. | |
| 20 | Troms | Trondenes | dBase/Borgos | | | | n.a. | | n.a. | |
| 21 | Sogn og Fjordane | Sogndal | Ættesoge | Finn Førsund og Anders Timberlid | | | n.a. | | | We have not been able to get information about Ættesoge projects |
| | | | | | | | | | | |
| | This table shows some fulfilled and ongoing bygdebok projekts, period covered and number of linked persons (over 300.000). | | | | | | | | | |
| | The overview is about complete for BSS, FAMREK and dBase/Borgos. | | | | | | | | | |
| | FTE: workyear, i.e. in these projects 55,4 years of work has been designed, and about 13,7 remained by fall 2010. | | | | | | | | | |