НОРВЕЖСКИЙ ИСТОРИЧЕСКИЙ РЕГИСТР НАСЕЛЕНИЯ, 1801–1815 гг.

THE NORWEGIAN HISTORICAL POPULATION REGISTER — PARTICULARLY 1801 TO 1815

Lars Holden
Dr. Scient.
Norwegian Computing Center
Director
Oslo
Lars.Holden@nr.no

Torkel Rønold Bråthen
Cand. Philol.
The National Archives of Norway,
Project leader for «The Norwegian People in 1814».
Oslo
torbra@arkivverket.no

Gunnar Thorvaldsen
Dr. Philos.
Norwegian Historical Data Centre, University of Tromsø.
Professor
Tromsø
Gunnar.Thorvaldsen@uit.no

Исторический регистр населения систематизирует данные, используемые при проведении медицинских исследований, в социальных науках, истории, географии и информатике. В настоящее время регистры населения, охватывающие большие периоды времени, создаются в нескольких странах. Так, недавно состоялась встреча исследователей, давшая начало Сети исторической выборки. В Норвегии исторический регистр населения будет дополнять существующий с 1964 г. Центральный регистр населения. Наша задача состоит в охвате максимально возможного числа из 9,7 миллионов человек, проживавших в Норвегии в 1801–1964 гг. Национальный архив, Норвежский вычислительный центр и Центр исторических данных Норвегии в настоящее время формируют часть регистра с 1801 по 1815 г., которая будет запущена в годовщину конституции 1814 г. В настоящей статье, в частности, обсуждаются методы связывания записей.

Ключевые слова: регистр населения, база лонгитюдных данных, микроданные, перепись, церковные записи, междисциплинарный подход, связывание данных, медиавики.

The Historical Population Register (HPR) organizes data that are used in health research, social sciences, history, geography and information science. Several countries are presently building longitudinal population registers. Thus, the Historical Samples Network recently had its launch meeting. In Norway, the HPR will complement the current Central Population Register from 1964. Our aim is to include as many as possible of the 9.7 million persons who lived in Norway between 1801 and 1964. The National Archive, Norwegian Computing Center and the Norwegian Historical Data Centre currently build a register part from 1801 to 1815 to be launched at the Constitution Anniversary in 1814. This article explains in particular the record linkage methods.

Keywords: Population register, longitudinal database, microdata, census, church records, interdisciplinary, record linkage, mediawiki.

INTRODUCTION

The long-term goal of the Norwegian Historical Population Register (HPR) is to cover the period since 1800 and where possible even earlier. The short-term goal is an expanding population register covering an increasing number of representative regions during this period by linking information in currently transcribed national and additional computerized sources. We have planned how to bring the transcribed and linked pieces of demographic information into a coherent national database, which is now being built on the basis of some transcribed sources which already provide national coverage. This constitutes valuable parts of the longitudinal database and document procedures for the future coordination of a complete national population register.

In the 21st century we need to access demographic data digitally and longitudinally where possible in order to remain at the forefront of population oriented studies. As the forefront of population research also historically moves from the use of cross-sectional sources to longitudinal registers, the data structure of our source material must be reorganized in order to serve us and our international partners in comparative research. The current stand-alone data bases with the sources organized cross-sectionally and separately form a straightjacket. There, individuals and groups making up the population can only to a very limited degree be followed over time. However, the potential inherent in the combination of contemporary Scandinavian register data and our rich historical data is huge. With national long-term coverage such a database will be an internationally unique resource for many types of population oriented research purposes. For a host of research questions in the humanities, social sciences and medicine the new database structure opens up possibilities in three new ways: 1) The periods before 1960 will be opened up for longitudinal research on a national scale 2) the analysis of recent social and other population phenomena can be based on data with a longer time horizon and 3) it is crucial to be able to follow individuals, kinship and genetic networks over extended periods. Two existing integrated community databases have been expanded in order to test out relevant procedures, especially the developing international standard for the exchange of longitudinal data.

THE NORWEGIAN PEOPLE IN 1814 (DNF1814)

Construction of the Historical Population Register (HPR) is now in full swing, initially for the period 1801 to 1815. Our immediate goals are to reconstruct the population in the days around 17 May 1814 when the Constitutional Assembly finished its work, and to construct a nominative census to supplement the statistical census of 30 April 1815. Upon completion of the constitution jubilee in 2014 this will be the world's first public and national population register and will cover the period 1801-1815[1]. This project has been named “The Norwegian People in 1814” (DNF1814), and the initiative was taken by the Director General as part of the celebration of the anniversary of the Constitution. It is implemented by representatives of the National Archives of Norway, The Norwegian Historical Data Centre at the University of Tromsø, Norwegian Computing Center and other volunteers. The main strategy in making the population register in 1814 is as follows: for each parish we make a list from the census in 1801 and add/remove persons using parish records of baptisms and funerals in the period 1801-1814. The critical part is to link the persons that are buried or migrated. The DNF1814 project represents only a small portion of the exposed part of HPR.[2] In fact, HPR aims to cover the entire Norwegian population until around 1930, and in this larger project several partners will participate. It is realistic in Norway to eventually achieve this because we, along with the other Nordic countries, have complete parish registers that fill the gaps between the censuses[3]. A closed section that ties into The Central Population Register, which was created in 1964, is also planned[4].

The HPR will connect all open historical personal and location information about individuals in Norway. It will be built mainly from parish registers and censuses, and will eventually include other information such as emigration records, probate records, prison records and gravestones. Information about individual persons who appear in multiple sources and information on residences will be linked together. This is an enormous task that only our descendents can finish. The period from 1735 to 1964 contains 9.7 million people and 37.5 million entries in the census records, parish registers and other sources[5]. HPR can be perceived as the repository of all these sources because it provides an overview of the same person and family from several sources and various settlements.

NEW OPPORTUNITIES WITH THE HPR

DNF1814 and subsequent extensions of HPR will be used in a variety of local, regional and national studies and provide a basis for international comparisons. The HPR will, among other things, take part in an international collaboration through the NAPP project[6]. It will give us new historical and social science insights into the relevant periods. With longitudinal microdata, we can study how family structure and social and geographical mobility changed continuously unlike the snapshots given in the censuses. From a medical perspective, HPR will be an important

source for studies of such things as genetic diseases[7], genes in the Norwegian population[8] and intermarriage[9]. The bibliographies at the Demographic Data Base in Umeå[10] and the Minnesota Population Center in Minneapolis[11] contain many research topics which have not yet been investigated in Norway, so the possibilities are virtually endless.

A key demographic issue concerns the population development from 1801 to 1825. The *Norwegian Immigration History*[12] claims an emigration of two percent until 1815 and then a similar immigration until the census of 1825. Previously, Michael Drake agreed with both Eilert Sundt and Gunnar Jahn that this is due to the 1815 census' low quality with an undercount of about two percent[13]. It will be interesting to see to what extent the DNF1814 project can support the *Immigration History's* attempt to create confidence in the 1815-census — without discussing the aforementioned historiography.

Regarding local history and genealogy it will be easier to place families and community histories into a larger context. In practical terms, it will be easier to identify the sources and more rational to connect with the work of others. It will always be possible to find more information, more sources and comparable stories. There will be less need to find the same individuals and duplicate the same links over and over again, and we can instead build on the work of others and complement the sources for each person and family. The result is a database where we can follow individuals, families, farms, places and other locations over time. The challenges throughout the whole period until 1964 are the same. It is difficult to establish such a population register, and this article describes some of the challenges we face and how they are being solved. This particularly applies to the record linkage that will happen through a major effort on the Internet. The goal is realistic if we work together through crowd sourcing. The HPR represents new technology for collaborating on the linking of personal data. This is described in the section on record linkage and builds on MediaWiki software, that also is used by Wikipedia[14]. The presentation contains a number of methodological and technical details that are not necessary to know for users of the system. First, we discuss the main principles behind the construction of the HPR.

MAIN PRINCIPLES

1. The HPR will first and foremost concentrate on the input of data from the sources, initially censuses and parish registers. We must be sure that we include all person records[15] in these sources once, and only once. The challenge is to connect people and places from different sources. This will be accomplished partly automatically and partly manually. The main principles of the HPR are as follows:The HPR shall be based on as many sources of good quality as possible. There will be two-way links, i.e. from the database to the sources and vice versa.
2. We want the greatest possible openness and transparency. This makes it possible to see who has made changes and the reasons for doing so.
3. We will encourage the greatest possible use, and as many contributions as possible to the development of good quality. All users will be given the opportunity to comment on the quality of the database.

These principles secure uniqueness and complete references to sources. It will ensure a high amount of good quality data and ensure that quality increases over time. The HPR will function as an index or register to the sources instead of replacing them. Both with regard to citations and transparency, the HPR will be different from the historical population register in Iceland, *deCODE*. This system is closed so that the public only have access to their own ancestors, and it does not have systematic references to the sources or explanations of how the person records are linked[16]. The historical longitudinal databases in Sweden and the Netherlands have explicit links to the original sources, but cover only parts of these countries. The *European Science Foundation* supports a network aiming to create historical population registers in other European countries[17].

USE OF SOURCES

In Norway, parish registers have been kept since 1623 with entries for baptisms, marriages and deaths. This has been compulsory since 1680. We expect, however, an under-count of 10% to 20% in the 1700s[18]. This is one of the reasons why the national HPR starts in 1801 with the nominative census that year as the starting point, but another culprit is that the earliest sources are usually missing information variables. The persons' age may be missing, the names of married women can be excluded in several records and names of residences are not given.

Unique identities (IDs) are now being established in the Digital Archive[19] for each person and property record for the sources that are transcribed, respectively PKID and EKID. These IDs are used as references in the HPR. The somewhat complex format of these machine references should not concern users. When the reference is available with a click, it becomes easy to determine whether there is additional or conflicting information available from the same source. Combining many possible sources of good quality will improve the HPR both by providing more information about people and places and by making linking more secure. Com-

plementary sources can be anything from probate protocols to newspaper reports. Initially the entire source is imported into HPR and a page is established for each person and property instance in the source with two-way links between the different people and places mentioned in the source and in the HPR. This ensures that you can easily move between one person or a place mentioned in the HPR and the source where the person or place is mentioned. Some sources are not suitable to be scanned in their entirety, such as lists of Ellis Island immigrants to the United States and Yearbook of Danish Gentry where only a small portion of the person instances are relevant for Norway. In these cases, a source table in the HPR with links to both person pages in the HPR and the place where the person is referred to in the source will be established instead. Thus it will be easy to keep track of the persons who are registered in the HPR, and to add more people to HPR who are mentioned in the source. A source table in the HPR will ensure uniqueness.

PROCESSING OF DATA

The censuses contain information about family and household position which provides the basis for grouping individuals in families. The HPR establishes families automatically when importing census records based on information about family position, order in the residence registers, gender, age and name. The automation is based on computer program output from our partners at the Minnesota Population Center as part of the *North Atlantic Population Project* (NAPP).

All references to the sources use one name, year of birth, etc reproduced verbatim. The same person may thus have variant spellings of names, different age and such from different sources. For the overall presentation of the person we will however choose the core data we have the most confidence in. These are selected based on the following order of precedence: 1) Manual override 2) parish registers where older (contemporary) takes precedence over newer 3) censuses where newer takes precedence over older and 4) other sources where newer takes precedence over older.

RECORD LINKAGE AND RELATIONSHIPS

Record linkage can be defined as the act of identifying and determining that the same person is mentioned in at least two different sources. Relationships describe the ties between different family members and are based upon sources that mention several people in the same event. Different record linkage systems have been described elsewhere[20]. Here we repeat briefly that the family reconstitution method can be traced back to Henry Louis in the 1950s, and was further developed manually for the Norwegian sources by Ståle Dyrvik, Sølvi Sogner and Lajos Juhasz in the 1960s and 1970s. The method was adapted for computers in the Etne Parish and Rendalen Parish databases, covering major parts of the 18th to 20th centuries. Lars Nygaard and Eli Fure used a more general software to create a population register for Asker and Bærum. The software "Busetnadssoge" by Arnfinn Kielland and Ole Martin Sørumgård uses residence as an important linkage criterion and the database is used to print genealogical farm and community history books. Automatic detection and linking was programmed to link parts of the population of Troms in the 1800s.

Another asset we can build on is that several genealogical databases now use the MediaWiki software on the Internet[21]. *WeRelate.com* is the largest with over 2.1 million people in the database. *Geni* is a corresponding database that doesn't use MediaWiki[22]. These are impressive systems with a lot of graphics that are designed to handle many users. They are largely based on importing complete family trees via GEDCOM files from genealogists. Since they are not based directly on the original source material or verbatim transcriptions, they have problems with duplicates and database quality[23].

The HPR represents a new technique for creating population registers and the size of the project makes this necessary. In the HPR all the events are read directly from the main sources and stored as separate pages. It is extremely important that the system is clear, simple, transparent and designed so that the quality will improve over time instead of degenerating. A key element in the linkage strategy is thus how the MediaWiki system can motivate many users to offer high quality contributions and how these contributions are monitored and quality-assessed. We have chosen the wiki technology, which solves these problems in a good way. To keep track of all the different groups of people in the database we will actively use the *category* concept in MediaWiki. This will make it easy to find individuals featured in a specific *source,* all *families in a municipality,* all *priests in Norway* or other groups of individuals, families and places.

The national perspective makes it possible to capture migration between different areas as well as immigration and emigration. We know that persons who move between areas are possible to find in both areas, and special lists for migrated persons that have not been rediscovered will be established. Similarly, it will be possible to rediscover the roots of persons of Norwegian descent living abroad in the HPR. From a research point of view, it would be particularly interesting to follow the persons who move and whom we have to a lesser degree found in other databases. Many will enjoy the challenge of linking people who have moved. Experience from other projects, such as the Local History Wiki,[24] shows that different contributors are ea-

ger to share knowledge from their particular area of expertise. The themes may be own family and residence, geographic areas, specific occupations, ethnic groups, known/mentioned persons from other sources and secondary literature, standardization of names and correction of spelling errors. It is possible to describe the work within a particular theme in terms of project pages. We expect that it will provide many users with different approaches to the same data which will improve the quality. The different contributors will use different techniques for searching and linking.

For all links an indicator of the quality of the link on a scale of 0-10 will be established. As a guideline, the following scale will be used:

10: completely secure link, for example when you know the person through family or as a person of some fame

8: the same birth date and same or similar name and geographic location

6: same or similar name of both spouses and geographic location

4: same name, appropriate age and linked to the same place

1: questionable, but probable link, provides grounds for linkage

0: established as a candidate for linkage, provides potential grounds for linkage

If a user is unsure whether two pages refer to the same person, the two person instances can be registered as candidates for linkage. In this case the pages are not merged, but links between the pages are established.

AUTOMATIC OR MANUAL LINKAGE?

In the DNF1814 part of the HPR-project alone information on approximately one and a quarter million inhabitants will be linked, i.e. about 888,000 in the 1801 census, the 350,000 births up to 1815 plus an unknown but relatively small number of immigrants to Norway. There are also 334,000 deaths and 102000 weddings. This requires significant resources and makes it interesting to consider how much of the linkage process can be automated. It is well worth noting what the pioneer Hans Chr. Johansen has written about this — the Danish sources correspond well with the Norwegian during this time[25].

With reference to Figure 1 there are two goals: 1) to link correctly as many individual items as possible and 2) to avoid linking entries that do not belong together. The ideal is at the origo (bottom left) where all linked entries really belong together and no links are omitted. The problem is that when we impose stricter linkage criteria to reduce the risk of linking records from different individuals (horizontal axis), we will also omit several links that really should have been included (vertical axis). Conversely, if we introduce more liberal linkage criteria to ensure that all potential links are realized (far down the vertical axis), we will come to include more links that really should be discarded (when we go to the right on the horizontal axis).

Hans Chr Johansen believes that acceptable linkage results are located between points A and B on curve I in the graph, either with few accepted erroneous links (A) or few omitted correct links (B). The linkage results along curve II he thought were unacceptable because it both introduced many erroneous links in the database and omitted many links for person instances that should be merged. And thus his pivotal point for our context: As of the latter half of the 1800s we have source materials with a precision that allows automatic linking along curve I, while sources from earlier times are so imprecise that they must be linked manually, otherwise we end up with curve II. Fortunately, we have over the past decade developed techniques and methods that allow us to conclude somewhat differently, at least for some of the material.

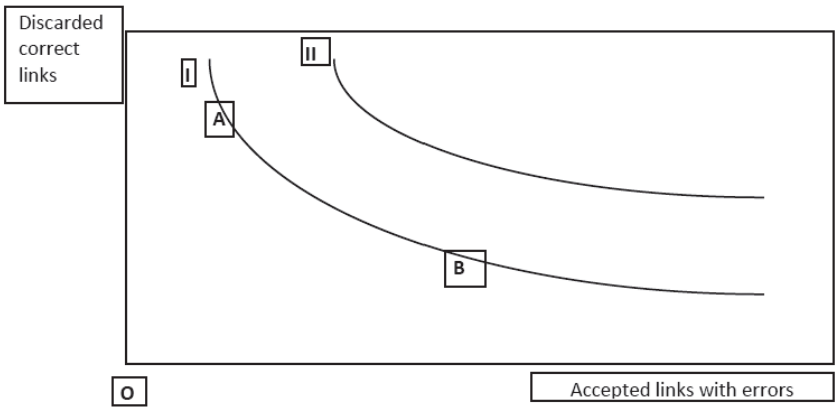


Figure 1. The ratio of accepted links with errors and correct links that were discarded in sources from the early 1800s (curve II) and late 1800s (curve I). Cf Johansen 2002

Firstly, we can flexibly bring together variant spellings of the same name, both using standardized lists of names and algorithms for comparison of strings without making the precision of the names significantly lower. Second, we have developed software that not only links individuals, but also takes into account (married) couples or entire families in linkage. Both are particularly valuable for a period when information other than name (age, place of birth) is imprecise or missing in the sources. By using this we can move the part of the population that can be identified in several sources, from curve II to curve I in Figure 1. This is a considerable proportion. A special processing of the 1801 census shows that about 80% of women and 85% of men were living with at least one other family member[26].

The fundamental problem of the trade-off between the number of links and the percentage of incorrect links is also reduced in the HPR by the possibility for manual linking. The same requirements apply for linking in the entire database and provide a quality indicator and a justification for each link so that each user can make an independent assessment and choose their trade-off. In addition, there is the possibility to establish candidates for linkage which can be tested by other users.

AUTOMATIC LINKAGE IN LENVIK

IT consultant Trygve Andersen at the NHDC has developed software for automatic linkage of married couples and other family members that utilizes special algorithms for comparing names[27]. During the DNF1814 project we are now running these programs against the 1801 census and the parish registers of Troms Province that are transcribed for the relevant period. The starting point is the men who were fathers in one or more of the 696 baptisms in the period 1799 to 1815, from 16 (1811) to 63 (1806) annually. The histogram in Figure 2 shows how many other source records it was possible to link to in terms of the mean proportion of fathers linked each year. Automatic linkage provides the best results within the baptism lists where both father and mother are brought together year after year. Around two thirds of parents are linked to the marriage, a stable number over time. Both spouses' names are in both lists and parish priests generally had the same spelling. The biggest problem is linking to the 1801 census, because here most couples are recorded at different places of residence before they were married and had children together. Thus we get good linkage results early in the period and weaker results towards the end because a bigger proportion of couples have married after 1801. The funerals that were linked to the father at baptism are mostly children, the results with respect to the linkage of baptism and burial must be understood on the background of the high fertility and child mortality during this period.

If the residence is stated in the parish register, and it is the same as in 1801, it is also possible to link individuals with no stated age in the parish register. During the period there are 834 funerals. Of these, 349 burials are not linked to any record about the deceased. Especially for older men, it is common that they are recorded in the parish register without reference to relatives. Also, later in the century many are left unlinked. This applies to at least a third of the persons that were attempted to be linked between the censuses for 1866 and 1876 for part of Troms Province[28]. In other words we see that there is considerable room for manual linking with the detailed knowledge about families and

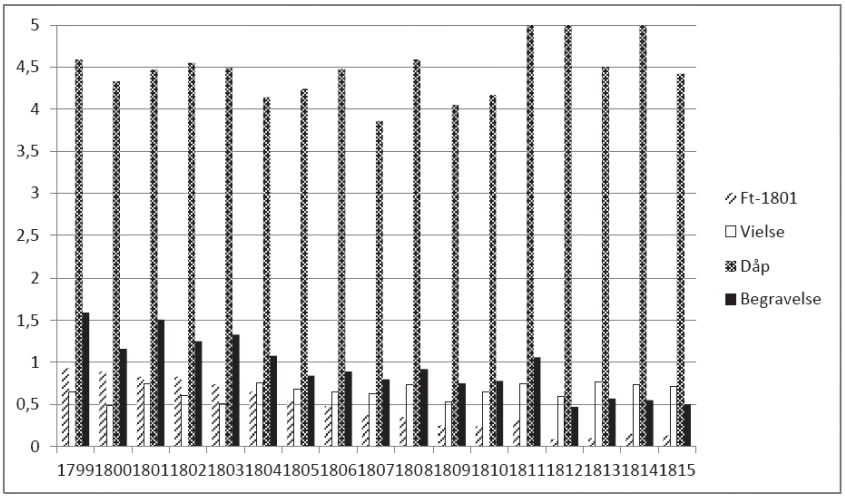


Figure 2. Average number of links from the father at baptism to the baptism, marriage and burial lists for the period 1799-1815 and the 1801 census for Lenvik. Ft-1801 = census 1801, Vielse = marriage, Dåp = baptism, Begravelse = Funeral

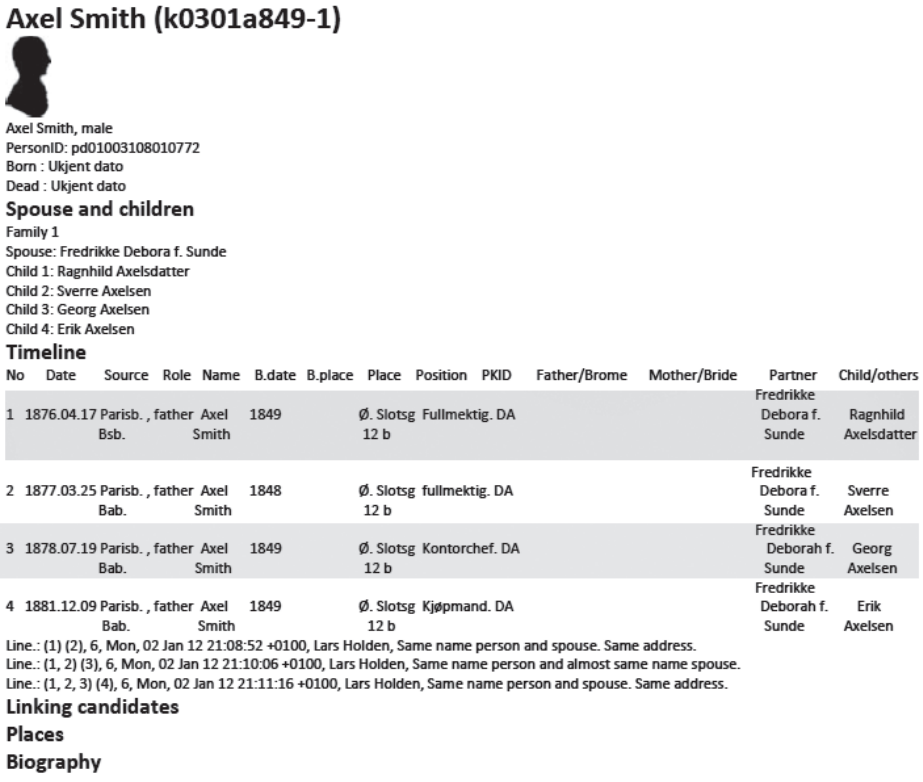
communities that are fortunately abundant in Norway. With HPRs wiki system on the Internet, we can thus gather this knowledge in a great collaborative effort and solve many of the puzzles we as individual scholars are struggling with.

STRUCTURE OF THE WIKI

HPR is being built as a wiki in line with Wikipedia. It uses the software MediaWiki. The wiki is based on a MySQL database and programmed in the PHP and C languages. In a wiki database all changes are saved and stamped with the time of change and who performed it. It is easy to reverse the change. For each page in a wiki there is a page where you can discuss the content of the page, e.g whether or not a link is correct.

Pages are established for each person and each family. A person page has a title of the form "Ole Hansen (k0301a849-21)" which indicates that it is the 21st instance of a person named Ole Hansen, registered in the HPR from the municipality with number 0301 and estimated born in 1849. Figure 3 shows an example of a *person page*. Here it is possible to gather all the source data for the person and links to family members, and have the opportunity to add a picture and write a biography. A *family page* has a title such as "Ole Hansen (k0301-a849-21) and Johanne Thorsdatter (k03010a855-8)". Residences are created from census records and have links to the residents from each census. Their titles contain place names of the form "Møre og Romsdal, Surendalen, Bakken (3)". The number is used only for residences where it is necessary to provide uniqueness. Residences are not established from the parish records because their descriptions are not accurate enough. Person, family and place pages are created by importing from each source (initially censuses and parish registers). The pages are merged when we have sufficient assurance that it is the same person, family or residence mentioned in various sources.

The person page in figure 3 shows a father who is linked to four person instances in baptism. Notice that the age and name of the spouse vary somewhat between different source entries. In the second line in the life-cycle table it says "Line no.: (1,2) (3), 6". This shows that the merge is done by merging lines 1 and 2 with line 3 in the table and this link is given quality 6. The categories at the bottom show some of the sources. The page can be tested interactively at slekt.nr.no.



Axel Smith (k0301a849-1)

Axel Smith, male
PersonID: pd01003108010772
Born : Ukjent dato
Dead : Ukjent dato

Spouse and children

Family 1
Spouse: Fredrikke Debora f. Sunde
Child 1: Ragnhild Axelsdatter
Child 2: Sverre Axelsen
Child 3: Georg Axelsen
Child 4: Erik Axelsen

Timeline

No	Date	Source	Role	Name	B.date	B.place	Place	Position	PKID	Father/Brome	Mother/Bride	Partner	Child/others
1	1876.04.17	Parisb. Bsb.	father	Axel Smith	1849		Ø. Slotsg 12 b	Fullmektig. DA				Fredrikke Debora f. Sunde	Ragnhild Axelsdatter
2	1877.03.25	Parisb. Bab.	father	Axel Smith	1848		Ø. Slotsg 12 b	fullmektig. DA				Fredrikke Debora f. Sunde	Sverre Axelsen
3	1878.07.19	Parisb. Bab.	father	Axel Smith	1849		Ø. Slotsg 12 b	Kontorchef. DA				Fredrikke Deborah f. Sunde	Georg Axelsen
4	1881.12.09	Parisb. Bab.	father	Axel Smith	1849		Ø. Slotsg 12 b	Kjøpmand. DA				Fredrikke Deborah f. Sunde	Erik Axelsen

Line.: (1) (2), 6, Mon, 02 Jan 12 21:08:52 +0100, Lars Holden, Same name person and spouse. Same address.
Line.: (1, 2) (3), 6, Mon, 02 Jan 12 21:10:06 +0100, Lars Holden, Same name person and almost same name spouse.
Line.: (1, 2, 3) (4), 6, Mon, 02 Jan 12 21:11:16 +0100, Lars Holden, Same name person and spouse. Same address.

Linking candidates

Places

Biography

Figure 3. A person page in the HPR where there are links between four baptisms in the family. All names at the top under the heading *Spouse and children* have a link to person and family pages. DA in each row in the timeline table has a link to the corresponding source. If any of the sources had been a census, there had also been given a name for the place and a link to the place at the bottom of the page

Figure 4 shows an overview of the pages in the HPR. A person page is created for each person instance, and these are merged when we are sure that the linking is correct. A family page is created for each family with at least two persons. Family pages are merged when the parents are the same. A page is created for each residence from the censuses with a hierarchy of place pages for parish, municipality and province from the subdivisions in each census. We use the same province pages for all censuses and merge place pages for residence, parish and municipality if they are the same in different censuses. A place page for a farm can thus point to different municipalities in the various censuses because of municipal border changes over time. A place page will then have a list of the inhabitants for each census. This may be important in the linking. Additionally, there are project pages where one can describe the work in a community history association or other topics. The project page will provide links to external sites about the project and links to relevant pages in the HPR, such as specific residences or persons. There are also pages documenting each source that is used in the HPR.

A population register can provide a number of challenges related to the protection of personal privacy. As a point of departure, the rules and policies which the National Archives use for publishing their sources will be followed. The Norwegian Data Protection Authority has allowed for a more liberal practice so that name and family relationships can be published, but this should be removed if the relevant persons request it[29]. Information that may be considered sensitive for living persons will not be made public.

INPUT OF LINKED COLLECTIONS

The automatic linkage can take place outside the HPR database, since the HPR can import lists of groups of PKIDs (described above) to be linked or related. Technically speaking, from the HPR's perspective it does not matter how this list is assembled, but the format is important. For HPR to be able to use the information it is necessary that we can retrieve references to the sources and identify the persons in the HPR. The PKIDs fulfill both requirements. If the list contains information about the quality of the link, this can be translated into the 0-10 scale for the quality of links, as above. In the HPR we try to get contributions from everyone who has linked large collections of the Norwegian population. There will also be automatic and semi-automatic linkage methods provided in the HPR. This ensures that as soon as a new source is read into HPR, the most obvious links can be established.

A person page in the HPR cannot have multiple fathers or mothers. It is, therefore, not possible to merge personal pages where there is a conflict between the parents' records on the two pages. Therefore, if you have two whole family trees that represent the same people, one must merge the family trees by starting with the oldest generation and processing one generation at a time. When HPR imports a linked collection, one group of person instances to be linked is processed at a time. If any person instances cannot be linked because of a conflict between the links to the father or mother, this group of person instances is appended to the end of the list. There is always the hope that the

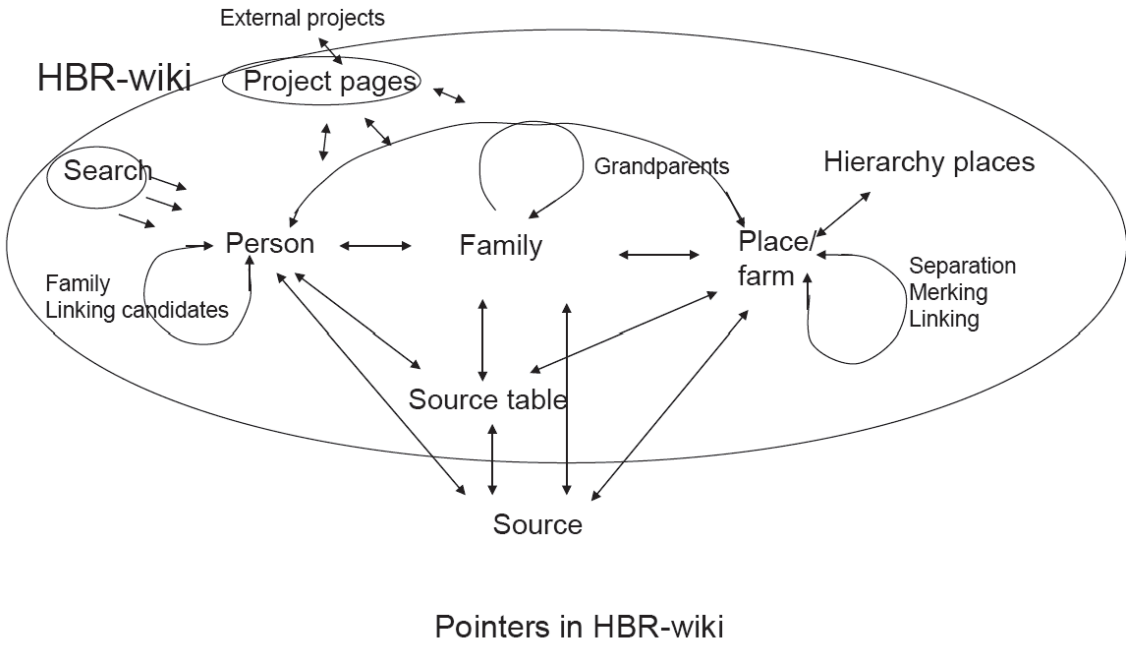


Figure 4 shows one- and two-way links in the HPR represented with single and double arrows. The most important pages are the person, family and place pages. These pages have links to the sources and the project pages

parents' different pages can be linked together at a later time as new information becomes available. If the parents' different person pages cannot be combined as in this instance, the children's person pages are established as candidates for linkage to each other.

Lists of PKIDs may for example be created with an automatic linkage program that compares name, age and residence and calculates the probability that it is the same person as has been done in some parishes and in the province of Troms. However, the list can also be created by scanning a population register from an area, a family or other thematic collection. GEDCOM files can be used, but this requires that the source references exist in order to avoid duplicates. Thus, a quality assessment must be performed before you import the linked collections.

MANUAL LINKAGE

We expect that much of the linkage and establishing of family relationships, especially in DNF1814, will be done manually with contributions from many genealogists and people interested in local history during this great effort. The manual linkage may take place through someone finding people in the HPR and checking references to sources against their personal or family book databases. But it can also happen by searching for person instances in the HPR by chronologically following all events in a parish register, all events on a farm over time or searches using other criteria in the sources. If in doubt, it is possible to "test-link" two person instances as candidates for linkage, and then single-handedly or with the help of others confirm or disprove the link at any time later. Especially in the early parish registers and for emigrants and immigrants it may be necessary to establish many possible candidates for linkage before it is possible to identify one of them with reasonable certainty.

Linking of person pages, family pages and place pages in most cases implies to merge the pages using the name of one of the pages. Links internally in HPR and to outside sources of the merged pages will be updated. In all text that refers to original sources a verbatim reproduction of names, dates and the like will be preserved. If two person instances are to be linked and only one of the person instances is already created in the HPR, the other person instance is recorded on the first page and the necessary links are established. In the HPR family relationships are recorded, i.e. that two people are related, such as parents and children, by establishing links between personal and family pages.

IDS FOR PERSONS

It is necessary to establish IDs for all persons in the HPR so that we can have a stable reference to persons internally in the HPR, and between HPR and external programs. Today's personal ID numbers are based on the reliable identification of date of birth. Date of birth will often not be known for historical persons, and cannot be used as an identifier. As previously mentioned, The National Archives establishes IDs for all person instances (PKID) and properties (EKID) mentioned in the sources. We consider this a reliable identification of a person instance, and this links the person to a well-defined source. If one later discovers that the name or other information about the person is transcribed incorrectly, it will not change the PKID.

The HPR uses a rule-based definition of personID (PID) on the basis of the PKIDs, that is to say we make a priority rule for which PKID to choose as PID. Thus, the PID is defined as the PKID with the highest priority of the links to a particular person. This is a system that can handle changes in the links. As an example, PKIDs A and B are linked, but a third PKID C that we link to B indicates that the link between A and B must be discarded. Then we change only the PIDs for the persons from A(B) and C to A and B(C), where the parentheses show which other PKIDs are associated with this PID (see Figure 5).

Figure 5. Changing personID (PID) when the link between two person instances (PKID) changes. In the example on the left PKID A has the highest priority and PKID B "inherits" PID from this when they are believed to refer to an identical person (thick arrow). When this link is subsequently broken but PKID C may be linked to PKID B, PKID B gets its PID back because PKID C has the lower priority of the two. Otherwise PKID B would receive its PID from C

HPR uses the following priority to create PID from PKIDs:

1. 1910 census
2. lists of funerals and deaths before 1910
3. births after 1910
4. lists of emigrants before 1910
5. lists of immigrants after 1910
6. other censuses where recent censuses have higher priority than the older ones
7. other entries in parish registers where older (contemporary) entries have higher priority than newer ones
8. other sources where newer entries have higher priority than older ones

The first five categories are not overlapping (except possible reimmigration), and the first three are probably of relatively good quality. The 1910 census will later be used to link between HPR and the closed portion of the population register up to the Central Population Register in 1964. The 1910 census is the first census with birth date. This makes it suitable to identification of persons. On the person page in the HPR the persons' PID is displayed. Internally in the HPR however, it will be more natural to use the URL or page name as an identifier.

HISTORICAL PLACES AND ADMINISTRATIVE BOUNDARIES

Changes in the population are naturally linked to the development of settlements, communities and regions. The HPR-project has the ambition to link to the dynamic information about topography when cottars' places were established, farms were split and towns were growing. A residence will be tied to multiple municipalities over time if the municipal boundaries change.

We can import the 1886 and 1950 cadasters and associate place pages from the censuses by using the farm name, place name, street name, title number, farm number, house number and street number[30]. Results from the project *Historical-administrative borders*[31] show that 80 to 90% of the farms can be tied to a unique title number. The remaining residences in the countryside and the northernmost province of Finnmark, which was not assessed for the taxation of farms in the 19th century and the towns will require extensive specialized work, especially the latter because of the transition from serial numbers for the city to serial numbers for each street. Again, we rely on local history expertise to solve the puzzle correctly. Arne Solli at the University of Bergen has done an impressive job with person and location information for cities in the BerGIS project[32] for the period 1696-1906[33].

It is planned to provide coordinates for all places. This provides opportunity to create many different graphics such as the migration pattern of persons or families on the basis of geographical coordinates as has been done for Troms Province[34].

TRANSCRIBING

By transcribing we mean computational copying from the sources via the keyboard, unlike scanning where only the picture, which is not searchable, is transferred to a digital format. Traditionally, this has been done by transcribing all records in a source systematically and collectively. Next, the quality is assured via proofreading performed by another person. Transcription is carried out both in state institutions and by volunteers. Many sources have already been transcribed in this way, but there are still large parts of this work left. In the HPR project work is therefore underway to develop more effective techniques for transcription[35]. Parish registers until around 1930 are available in the Digital Archives as scanned images, some records are on microfilm, but a lot is still available only on paper, for example the national censuses after 1910 which are not yet available to the public and the municipal censuses.

The HPR opens opportunities to transcribe individual records and to link records, via a source table, both to person pages in the wiki and to the scanned image of the source. Later, users will be able to check the quality by following the links to the scanned image. This will increase access to transcribed sources and improve the quality of the HPR, but there is also a danger that this could reduce the interest for full transcription, rendering the transcribed material less representative. If fewer volunteers undertake the job of transcribing whole sources, we shall lose important contributors. A transcription where the user based on accumulated experience has knowledge of the name of the person appearing in the source allows for better quality of transcription. However, it can also cause errors by not emphasizing minor spelling deviations.

TRACKING USERS

The HPR is openly accessible via the Internet. To contribute users must register with their full name and identity which is checked via e-mail. All changes are recorded with a time stamp and the full name of the person performing the change. A user can establish alerts so that they get an e-mail if there are changes to pages where the user has requested to be notified. It is also possible to block the accounts of rogue users. There is a large user and developer group that develops the MediaWiki platform which is used in a large number of websites worldwide. It is necessary to have significant IT skills and experience with social media on the Internet for such a site to work well. We must be prepared to adapt the development process

as we get feedback from users. The HPR will be a wiki with significantly more pages than the Wikipedia in English and with a large number of users; the association Computers in Genealogy has over ten thousand members in little Norway.

Within MediaWiki there are many different ways to track users available and Wikipedia uses this to create an encyclopedia of high, although varying, quality. For example, one can establish a group of super users who receive notifications when new users make their first changes. That way one can intervene early and offer guidance. It is also possible to lock pages, such as in cases of disagreements between users, so that they cannot be changed by anyone other than those with administrative privileges. The HPR group aims to establish super-users who follow up different parts of the site. The most natural procedure would be to establish connections with local historical societies that will contribute to the development and monitor changes within their area.

The HPR wants to encourage as many users as possible and facilitate that the HPR is used in various ways and with different approaches. Users can create projects that cultivate work within a theme. Projects may for example be run by a local historical society or individuals that focus on a family or people with certain occupations. The project page can link to sources or project sites external to the HPR, and link to key pages in the HPR that concern the project. We believe this will increase the interest in the HPR, disseminate information on relevant projects and give credit to contributors. The primary language of the HPR is Norwegian, but the most important pages will also be available in English. There is considerable interest abroad among genealogists and other researchers to follow emigrants or handle other issues on the basis of a better organization of the Norwegian sources.

We do not intend to compete with or replace other work or documentation. HPR will link together different types of information and encourage building on the work of others. This especially concerns the relationship with the Local History Wiki[36]. Longer texts and biographies are preferably added there or to other external sites. The HPR will primarily link information, refer to sources and provide short descriptions of life-cycles.

The HPR-wiki follows the same copyrights that are common to Wikipedia[37]. This means that content can be copied, modified and redistributed as long as users give *their* users the same rights, and as long as HPR is credited as the source. However, restrictions are imposed on the use of images. They can only be reused if the person publishing them gives permission.

A LOCAL EXAMPLE

In the Rendal database where most available nominative sources have been linked for the period 1735 to 1950, we can get an idea of the challenges we face and what results can be expected at the local level[38]. Of all the people that have been observed in the period 1815-1824, half of them are identified in baptism, marriage and burial registers, while an additional 20% are identified among baptisms and funerals, but not in the marriage lists. In view of the high mortality during the period, we know that a large part of the latter died unmarried. Rendalen was a parish with little migration and the database is the result of thorough work on a small geographic area. We cannot expect equal coverage in municipalities with greater migration.

Even after linking a majority of records for individuals, as has been done in Rendalen, we face challenges when we attempt to reconstruct the population at a given time. These challenges are not diminished by the fact that mortality was high and that part of the population could be geographically mobile during the turbulent times in the early 1800s. The high mortality meant that one hundred marriages were entered into where at least one partner was a widow or widower in Rendalen from 1801 to 1815. It is not straightforward to determine where they settled and which children from previous marriages moved with the surviving parent. In Rendalen we can determine this by exploiting information from sources that cover later periods, such as residence at baptism after 1815. Sources such as these will only to a limited extent be available in transcribed format for the DNF1814 during the next few years. In these cases access to local and family history expertise through the HPR will be especially helpful. The same applies to following individuals with common names. If we fail to link Ole Olsen, 15 years old in the 1801 census to the Ole Olsen who married in the neighboring parish in 1814, we risk counting the same person repeatedly. This may lead to an overestimation of the population if we do not manage to adequately take into account missing links to the census. Since we have censuses for most decades since 1800, it will be possible to compute weights based on how the linked sample differs from the whole population with respect to key characteristics such as age or occupation. Such weights are used in the NAPP project to adjust the representativity of the linked samples[39].

It is hard to tell what coverage we can achieve in the HPR. It also depends on the extent to which one includes “probable” links and not just “secure” links. In areas and periods with high migration the linkage rate will be lower. By having a national database we will be able to capture part of the migration, but not everything will be possible to document in available sources. Some measurements of linkage rate must be created that can be used to compare the geographical areas and periods. This may be proportion of links between baptism and burial, the proportion with one or two parents identified, or proportion of persons in a census that is linked to at least one other source.

SUMMARY

The HPR will constitute a unique source for this historical period and will be usable in many different research projects in a variety of topics. From an international perspective, there are no comparable historical population registers with approximately the same size built by linking multiple source types. In the same way that Norwegian biobanks attract international research, the HPR can attract international research to Norway to study the Norwegian data. We can make use of the fact that Norway, along with the other Nordic countries, has a good supply of source material. Extensive use of these materials requires the development of new linkage strategies consisting of a composite of several established techniques combined with new methods that increase the overview. In particular, the use of wikis is a significant innovation in this area. It will require a large concerted effort from a large number of people to achieve these goals, but in return this gives us an infrastructure with great utility for many different research projects. We are cooperating with several international projects in the field of computerizing nominative sources, and look forward to future exchanges of ideas, especially on how to better use combinations of the church registers in population research.

ACKNOWLEDGEMENTS

The authors thank Svetlana Boudko and Till Halbach Røssvoll at the Norwegian Computing Center who have implemented the first version of the HPR.

BIBLIOGRAPHY

Bråthen, Torkel Rønold. “Det norske folk i 1814” [The Norwegian People in 1814] i Slekt og Data, nr 4/2011, s. 44-45.

Bull, Hans Henrik. Data and Methods. Universitetet i Oslo, 2006, s. 25-34. http://www.rhd.uit.no/nhdc/Chapter%203%20Hans%20Henrik%20Bull.pdf

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

Eikvil, Line, Holden, Lars og Bævre, Kåre. Automatiske metoder som hjelp tiltranskribering av historiske kilder. [Automatic methods for transcribing historical sources.] Notat SAMBA/44/10, Norsk Regnesentral. Oktober 2010.

Gjelseth, M. Relasjonsdatabaser som verktøy i en historisk-demografisk studie. [Relational databases as a tool for a historic-demographic study.] Hovedoppgave, Historisk institutt, Universitetet i Oslo, 2000.

Heimen, spesialnummer om personvern, [special issue on the protection of privacy] 2005 med blant annet With, Cristian. “Slektsforskning og personvern” [Genealogy and the protection of privacy] i Slekt og Data, 1/2005. Tilgjengelig på Datatilsynet sider, http://datatilsynet.no/templates/Page____1356.aspx

Hovig, Eivind. “Norske gener?” [Norwegian genes?] i Genialt, nr 1/2010, s. 14-15. http://www.bion.no/filarkiv/2010/07/Genialt_1_2010_norske_gener.pdf

Johansen, Hans Chr. “Identifying people in the Danish Past” I Sølvi Sogner, Hilde Sandvik, Kari Telste, Gunnar Thorvaldsen (red): Pathways of the past: essays in honour of Sølvi Sogner on her 70th anniversary 15. March 2002. Novus, Oslo 2002, s. 103-110.

Niemi, Einar, Myhre, Jan Eivind og Kjeldstadli, Knut. “I nasjonalstatens tid 1814-1940” [In the time of the national state 1814-1940]] i Norsk innvandringshistorie [Norwegian immigration history], bind II. Oslo, Pax 2003.

Arne Solli. Urban space and household forms. Artikkel presentert på the Eighth International Conference on Urban History, Urban Europe in Comparative Perspective, Stockholm 30th August — 2nd September 2006.

Suren, Pål, Grjibovski, Andrej og Stoltenberg, Camilla. Inngifte i Norge, Omfang og medisinske konsekvenser, Folkehelseinstituttet, Rapport 2007:2.

Thorvaldsen, Gunnar. “Using NAPP Census Data to Construct the Historical Population Register for Norway” i Historical Methods. Volum 44 (1) 2011, s. 37-47.

Thorvaldsen, Gunnar. “Fra folketelling og kirkebøker til norsk befolkningsregister” [From censuses and church records to Norwegian population register] i Heimen, bind 45, 2008, s. 341-359.

Thorvaldsen, Gunnar. Migrasjon i Troms studert i folketellingene 1866-1900, [Migration in Troms studied in the censuses 1866-1900] Universitetet i Tromsø 1995. Cf summary in English at URL: http://www.rhd.uit.no/art/summary.html

NOTES

[1] Bråthen 2011.
[2] The first version of the HPR is available at: slekt.nr.no, while the background for the project is described at: URL: http://www.rhd.uit.no/nhdc/hpr.html.
[3] Thorvaldsen 2008 and Thorvaldsen 2011.
[4] URL: http://no.wikipedia.org/wiki/Det_sentrale_folkeregister
[5] Thorvaldsen 2011.
[6] The NAPP project is described at: URL: http://www.nappdata.org/napp/
[7] See URL: http://genetikkportalen.no/for a description of the extent of this work in Norway.
[8] Hovig 2010.
[9] Suren et al 2007.
[10] http://www.ddb.umu.se/english/about-ddb/publications/?languageId=1
[11] URL: http://www.pop.umn.edu/and URL: http://www.nappdata.org/napp/
[12] Niemi et al. 2003: volume II.
[13] Drake 1969: 8.
[14] For information on MediaWiki: URL: http://en.wikipedia.org/wiki/MediaWiki and URL: http://www.mediawiki.org/wiki/MediaWiki
[15] A person record or a person instance is an entry about one person in one source, for example data about a child in a census or about a mother in a baptism.
[16] Professor Ólöf Garðarsdóttir, University of Iceland, personal communication, November 2011. See also: URL: http://www.decode.com and http://en.wikipedia.org/wiki/DeCODE_Genetics
[17] About the Demographic Database, see: URL: http://www.ddb.umu.se/english/?languageId=1. For the Historical Sample of the Netherlands, see: URL: http://www.iisg.nl/hsn/. For the network European Historical Population Samples Network, EHPS-net, see: URL: http://www.esf.org/ehps-net
[18] Thorvaldsen 2008.
[19] URL: http://arkivverket.no/eng/content/view/full/629
[20] URL: http://www.recordlink.org/
[21] URL: http://www.werelate.com and URL: http://en.rodovid.org/wk/Main_Page
[22] URL: http://www.geni.com/
[23] It is a result of the method used to generate the database. WeRelate appears to have a substantial effort to reduce this. See among other things: http://www.werelate.org/wiki/WeRelate:Suggestions/Tweak_to_Duplicates_Report
[24] Hans P. Hosar, Norwegian Institute of Local History, personal communication. Cf URL: http://lokalhistoriewiki.no/index.php/lokalhistoriewiki.no:Hovedside
[25] Johansen 2002.
[26] The variable "famsize" — number of members in own family in the household — can be analyzed with other constructed variables at Minnesota Population Center at the website URL: http://www.nappdata.org
[27] The programs are developed in PL/SQL, an efficient interface to NHDCs Oracle databases. Names are compared with the Jaro-Winkler string comparison algorithm.
[28] Thorvaldsen 1995.
[29] Heimen 1/2005, special issue on privacy and Thorvaldsen 2008.
[30] Cadaster 1950: URL: http://www.dokpro.uio.no/cgi-bin/stad/matr50, cadaster 1886: http://www.rhd.uit.no/matrikkel/excel.html. The cadaster from 1838 has been scanned and will hopefully soon be available in machine-readable and searchable text form after being processed with OCR.
[31] The HAG project is a collaboration between among other things the Norwegian Mapping Authority, the National Archives of Norway, the Norwegian Institute of Local History and the Norwegian institute of Public Health.
[32] URL: http://bergis.uib.no/index.shtml.
[33] Solli 2006.
[34] Thorvaldsen (1995). www.werelate.com has a lot of graphics also using geographic coordinates.
[35] Eikvil et. al 2010 at: URL: http://www.rhd.uit.no/nhdc/HBR_notat_okt-2010.pdf
[36] URL: http://www.lokalhistoriewiki.no is a site run by the Norwegian Institute of Local History.
[37] Copyright on Wikipedia is described at http://en.wikipedia.org/wiki/Wikipedia:Copyrights.
[38] Cf Bull (2006) available on line.
[39] URL: http://www.nappdata.org/napp/linked_samples.shtml