Chapter 8
Record Linkage in the Historical Population Register for Norway

Gunnar Thorvaldsen, Trygve Andersen and Hilde L. Sommerseth

Abstract The Historical Population Register (HPR) of Norway aims to cover the country's population between 1800 and 1964 when the current Central Population Register (CPR) takes over. This may be feasible due to relatively complete church and other vital registers filling the gaps between the decennial censuses—In 1801 and from 1865 these censuses were nominative. Because of legal reasons with respect to privacy, a restricted access database will be constructed for the period ca. 1920 until 1964. We expect, however, that the software we have developed for automating record linkage in the open period until 1920 will also be applicable in the later period. This chapter focuses on the record linkage between the censuses and the church registers for the period 1800 until around 1920. We give special attention to database structure, the identification of individuals and challenges concerning record linkage. The potentially rich Nordic source material will become optimally accessible once the nominal records are linked in order to describe persons, families and places longitudinally with permanent ids for all persons and source entries. This has required the development of new linkage techniques combining both automatic and manual methods, which have already identified more than a million persons in two or more sources. Local databases show that we may expect linkage rates between two-thirds and 90 % for different periods and parts of the country. From an international perspective, there are no comparable open HPRs with the same countrywide coverage built by linking multiple source types. Thus, the national population registry of Norway will become a unique historical source for the last two centuries, to be used in many different multi-disciplinary research projects.

G. Thorvaldsen (✉) · T. Andersen · H.L. Sommerseth
Norwegian Historical Data Centre, University of Tromsø, Tromsø, Norway
e-mail: Gunnar.Thorvaldsen@uit.no

T. Andersen
e-mail: Trygve.Andersen@uit.no

H.L. Sommerseth
e-mail: Hilde.Sommerseth@uit.no

G. Thorvaldsen
Urals Federal University, Yekaterinburg, Russia

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_8

8.1 Background: Population Registers and Early Linking Methods

Currently, the Central Population Register (CPR) covers the population of Norway back to 1964. Earlier, local, card-based population registers covered an increasing number of municipalities from 1906, but these are not considered for computerization due to their volume and scattered archiving (Thorvaldsen 2008). Instead, the national Historic Population Register (HPR) is being built by linking the censuses and church record from 1800 until 1964. About a third of this source material has been transcribed, while the rest is being digitized. The HPR will become a unique historical source for the last two centuries and may be used in many different multi-disciplinary research projects. The potential inherent in the rich Nordic source material will be realized once the nominative records are linked together in order to describe persons, families and places longitudinally with permanent ids for all persons and source entries. This has required the development of new linkage techniques combining both automatic and manual methods, consisting of a composite of several established techniques combined with new methods that increase the linkage rates. Already more than a million persons have been identified in two or more sources.

Prior to the construction of the Historical Population Register, record linkage has been performed on local, nominative source materials, mainly censuses and ministerial records from localities, of which the transcription started in Norway around 1970. Even a bit earlier, family reconstitution was performed on selected parishes with manual methods, and later on these methods were adapted to link digitized records interactively after sorting them according to a number of criteria (Nygaard 1992). This process was still labour intensive and usually it was only possible to link a sample of the population in a local parish to subsets of nineteenth century censuses and church books. The first serious attempt with computer assisted semi-automated record linkage in 44 parishes around 1801 was performed by Jan Oldervoll and his students at the University of Bergen around 1980 (Engelsen 1983). A more automated approach has been used for linking the censuses of 1865, 1875 and 1900 for the province of Troms in the early 1990s (Thorvaldsen 1995, 2000). The database for the parish of Rendalen 1735–1900 was constructed on the basis of a manual family reconstitution in the late 1990s. Its use as a golden standard will be further explained in Sect. 8.5. In parallel, a system dedicated to interactive record linkage was constructed and used on two parishes outside Oslo for most of the nineteenth century (Fure 2000). This interactive process has been carried further with a semi-commercial, semi-automated system used to link censuses, church records, probate registers and other nominative sources in order to create layout for the printing of farm histories and genealogies in local community history books, called *Busetnadssoge* (Kjelland and Sørumgård 2012). Together with older software versions, this has been used to build linked, local population registers for about ten municipalities from the eighteenth century until the late twentieth century. For the last decades these build on digitized oral sources in addition to open written sources such as newspapers and family genealogies.

8.2 Sources

The keeping of church records in Norway spread slowly from 1623 with entries about baptism, burial and marriage (in chronological order) which became compulsory in 1680. It was not until 1812 that priests started to use printed forms with separate columns and defined headings, while the ministerial registers were organized by type of event with baptisms, marriages and burials in pre-defined sections. Høgsæt (1990) has found an undercount of 10–20 % in the eighteenth century records, particularly of children's burials. This, together with the fact that the earliest sources contain poor information on the ages, the names of married women and information about residence, motivates the decision to start the national HPR in 1801 with the first nominative census. It is realistic, however, to build local population registers for earlier periods.

The HPR will in principle link all openly accessible information about historical persons and their locations in Norway. It will be built mainly on church records and censuses, but may eventually include information from emigrant lists, newspaper notices, prison records and tombstones.[1] The HPR links information about the same persons appearing in several sources and their settlements. Combination of many sources will improve the HPR both by providing more information about people and places, and by making more reliable links. The period from 1735 to 1964 includes 9.7 million people and more than 37 million source events in census records, church records and other sources. From the period 1800 to 1960 some 10 million out of 30 million records in censuses and church records have been transcribed, and there are ambitious plans to transcribe the remainder of the church records until 1930 during the next few years (Eikvil et al. 2010).

The evolution of the population is closely linked to the development of settlements, communities and regions. It is an ambition of the HPR project to provide links to this information, which is nearly as dynamic as the population itself, involving the origin of smallholdings, the splitting of farms, urban growth, etc. A residence may be linked to several municipalities over time if administrative boundaries changed. The points of departure are the 1838, 1886 and 1950 farm tax registers. Census records can be connected to these, using farm name, place name, street name, title number, farm number, house number and street number.[2] Results from the project *Historical administrative boundaries* shows that 80–90 % of the farms can be linked to a unique title number. The dynamic farm register developed by the economist Kåre Bævre in this project can be linked to the HPR. Urban places

[1] A project to interpret newspapers with OCR-like methods has received separate funding from the Research Council. Genealogists have photographed and transcribed many thousand tombstones and added this information to a database.

[2] The cadastre from 1838 has been scanned and can hopefully be made available in machine-readable and searchable text form after being processed with OCR. Confer also http://www.dokpro.uio.no/cgi-bin/stad/matr50, http://www.rhd.uit.no/matrikkel/matrikkel1838.aspx and http://www.rhd.uit.no/indexeng.html

will require extensive new work due to the transition from the serial property numbers to street names and numbers in the towns. As will be discussed later, we rely on local and regional expertise to solve these puzzles correctly. Historian Arne Solli at the University of Bergen has done impressive work with the person and location information for cities in the BerGIS project for the period 1696–1906 and is seeking to extend this GIS to other urban areas (Solli 2006). Based on the coordinates for all main farms in the countryside and blocks of houses in the towns, it is possible to create illustrative and analytical graphics, such as displaying the migration patterns in a neighbourhood or the spreading of epidemics.

8.3 Data Processing and Database Structure

Several genealogical databases now use MediaWiki software to link and couple records via the Internet. *WeRelate.org* is the world's largest genealogy wiki with over 2.6 million people registered in the database. These are based on importing family trees via Gedcom files from genealogists. Since the database is not based directly on the source material, this creates problems with duplicates and database quality, and WeRelate administrators make substantial efforts to reduce redundancy (Quas 2011). We have experimented to use the WeRelate platform and to construct our own wiki version, but have decided that building a relational database gives a more flexible and efficient solution, while keeping much wiki-like functionality. The size of the HPR project makes this necessary. In the HPR we read and store all the events directly from the transcribed sources. This places great demands on the system as transparent, simple and designed to strengthen database quality over time rather than degenerating due to redundant duplicates which will pollute statistical use of a demographic database.

The HPR relational database will allow signed up participants to contribute with their expertise. Experience from other projects, such as the Local History Wiki illustrates how different contributors share knowledge on the basis of their particular expertise.[3] Topics include their own family and residence, geographic areas, occupations, ethnic groups, migrants, persons mentioned in historical literature. It is possible to describe the work within a particular theme in terms of project pages, inspiring users with different approaches and techniques to work on the data. The Norwegian Local History Institute is a partner which will link its Local History Wiki as an encyclopaedia with more comprehensive information about places as well-known persons found in the HPR.

The source entries have been transcribed verbatim, faithfully reproducing names, year of birth and other variables. The same person may have variant spellings of names, different age data and other incompatibilities in different sources. For the overall presentation of each person's data we choose the core data in which we have

[3] https://lokalhistoriewiki.no

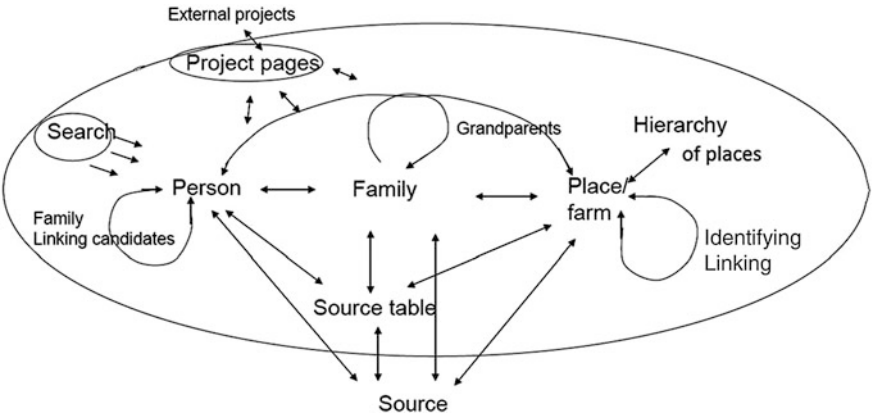

Fig. 8.1 One- and two-way pointers in the open HPR represented by *single* and *double arrows*. Project pages describe community history and local genealogy projects. The hierarchy of places is typically county-parish-farm-sub-farm which need to be linked internally and to the persons

most confidence based on the following priority rules: (1) selection by user during manual record linkage, (2) church records, where the older takes precedence over newer entries, (3) censuses, where recent takes precedence over older and (4) other sources, where recent takes precedence over older. Among church records the oldest will usually be the authoritative baptismal entry, whereas among the censuses the quality improved over time.

The main feature of the system is the establishment of a separate page for each person and family. A *person* page is of the type "Ole Hansen (21)", which says that there are currently at least 21 occurrences of entries for persons named Ole Hansen registered in the HPR. These sequence numbers are only used in the places, families and persons where it is necessary to provide uniqueness. On this page users can collect all source information about the person, links to family members and even put a picture or write a biography. A *family* page can be entitled "Ole Hansen (21) and Maria Thorsdatter (8)." Their residence information is created from censuses and has pointers to other residents. Settlements and addresses are usually not established from the church records because these are less precise in their description of objects. The pages constructed from the source entries are merged/linked when we have sufficient certainty that it really is the same person, family or residence that is mentioned in various sources.

Figure 8.1 gives an overview of the entities in the HPR. An individual entity is created for each person's occurrence and is linked to the relevant source entry. Family pages are created for each family from a marriage record or one parent with children. Family pages are merged when the parents' records are linked. "Grandparents" is a special case of parent–child links exemplifying how basic family links can be nested to show information about extended families. In addition, there is a free-text page in order to describe peculiarities in the sources or any other

topic related to linking or the population register. The HPR project will provide links to external sites about the project and links to relevant pages of the HPR, such as specific homes or people. There are also entries about each source used in the HPR. A place page for a farm or other types of residence may point to different municipalities in the different censuses because of municipality changes through time, and a family page may point to different place pages due to migration.

It is necessary to establish stable IDs for all persons in the HPR in order to have reliable references to individual occurrences of persons within the HPR for record linkage, and between the HPR and external databases. Today's social security numbers rely on stable and unique identification based on birth date, gender and additional digits. However, date of birth will often not be known precisely for historical persons and can hardly be used as part of the identifier. This is why the National Archives is establishing IDs for all *source entries* (called PKID) and all *properties* (eKIDs) for instance farms listed in the sources. We consider this as a reliable identification of each individual occurrence, referring each entity to a well-defined source entry.

The HPR uses a rule-based definition of each *individual's* ID, an unambiguous PID which is based on one of the PKIDs. Thus, the PID is defined as the PKID with highest priority among records linked to a particular person. This is a system that can handle revisions of the linking, as illustrated in Fig. 8.2. On the left side the PKIDs A and B are initially linked as PID A as A has highest priority. But when a third PKID (C) is linked to PKID B, the link between PKIDs A and B must be discarded. On the right side, PKIDs B and C are linked and priority is given to PKID B as personal identifier (PID) for the person whose records were linked. A separate table will contain the linkage history making it possible to see what links have been attempted unsuccessfully and what PIDs have been replaced due to the revision of links.

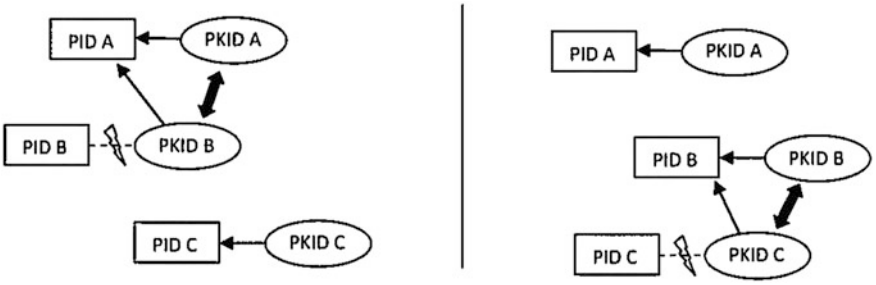

Fig. 8.2 Changing the person's ID (PID) when linking two person records (PKID). In the example on the left the PKID A has priority, and PKID B "inherits" the PID since the records refer to the same person (*thick arrow*). The example on the right shows that record B is instead linked to record C, and the previous link is broken. Record B's PKID becomes the PID because PKID C has lower priority

In the HPR we use the following source priorities to decide about the PID among the PKIDs when there are several to choose from:

1. Census 1910
2. Funerals/lists died before 1910
3. Births after 1910
4. Other censuses where recent censuses have priority over older ones
5. Lists of emigrants 1870–1930
6. Lists of immigrants from after 1910
7. Other entries in the parish registers where older entries (especially baptisms) have priority over later entries
8. Other sources where more recent entries have priority over older ones.

The first five categories are not overlapping, and the first three are presumably of relatively good quality. The census from 1910 will be crucial when linking the open HPR to the closed part of the population register from around 1920.

The results from the automatic linking described in Sect. 8.3 will be copied as lists of groups of PKIDs (the Ids described above) to be revised interactively. It is necessary that we include references to the sources and identify the individuals consistently, and the PKIDs fulfil both requirements. Thus, as soon as a new source is read into the HPR site, the straightforward links (on the two to three highest levels) are established automatically or semi-automatically. The tables will contain information about the quality of the links on the 0–10 scale (see Sect. 8.4.1) and linkage criteria. The web-based, interactive module of the HPR will receive suggestions for record linkage from many users, and such crowd-sourcing requires a standard set of rules with any conflicting links flagged and constant controls to verify consistency (see Sect. 8.3).

8.4 Source Dependent Record Linkage

The HPR will mainly rely on the input of data from censuses and church records, ensuring that we include all person records in these sources only once. There are some duplicate records also in the originals, primarily due to the combined use of de facto/de jure principles in the censuses and some entries of people who died away from their usual residence. These are far fewer and easier to remove by deduplication than all the redundant records which exist in databases built by combining collective genealogies, which are bound to contain more common ancestors as we move backwards in time.

Recognized relationships between spouses and other family members increase the likelihood for successful linkage, but parish registers and censuses provide information about relationship between persons in different ways. The census taker

registered relationships on the level of family and household and these have been made explicit with pointers constructed through the cooperation of the North Atlantic Population Project (NAPP).[4] Parish registers provide information about relationship according to type of event. Marriage lists naturally inform about married couples and from 1820 onwards included the bride's and bridegroom's fathers. For the majority of baptism records we are able to relate explicitly the mother, father and child. The burial represents a more individualistic event, and provides a more difficult source to link. From 1877 the relations between a parent and the deceased child or between spouses are specified in a separate column, but fortunately the practice to register the father's or spouse's name started earlier in some parishes. Elsewhere, the burial lists until 1877 provide information only about the names, age and address of the deceased, which is often insufficient information for linking to other sources. Age can be missing or unreliable in the first phase of the HPR, and birth date is only found in the baptism lists during most of the nineteenth century. For the marriage and burial records a more thorough registration of birth dates started in 1877. Date of birth was introduced later in the censuses, for persons aged under two in 1891 and for all from 1910 onwards, while those older than two only had their birth year reported in the 1891 and 1900 censuses. If place of residence was registered along with first and last name in the parish register, and corresponds to the address in the census, reliable linkage is more likely, especially in the period around 1801 when information on birth place and other characteristics is often missing.

A person in the HPR should not have multiple fathers or mothers. Even if a person's ancestry cannot be linked because of a conflict between competing links to the father or mother, there is always hope that the parents' person records can be linked later when additional information becomes available. Thus, children and parents' potential person records are marked as candidates for linkage. Attempts to merge person entries manually where there is conflict between the candidate parents will result in a warning.

In spite of the fragmentary nature of the migration lists in the church books, the national scope of the project makes it possible to capture migration between different domestic areas. Migrants can usually be retrieved at both place of origin and destination, and special lists for migrating people who are not identified can be established. Similarly, it will be possible to rediscover the roots of most people of Norwegian descent living abroad in the HPR since the emigration lists from around 1870 with about one million records have been transcribed. Returnee emigrants is a bigger challenge, but many are listed explicitly in the 1910 and 1920 censuses, and from World War I onwards the lists of immigrants became more complete.

[4] https://www.nappdata.org

8.4.1 Record Linkage Principles

The main principles behind the automatic and manual linking in the open HPR are:

1. The HPR will build on many available sources of good quality. There should be two-way links, that is, from the HPR to the sources and from the source entries to the HPR.
2. We want the greatest possible openness and transparency. It should be possible to see who made the links and what criteria were applied.
3. We pursue the highest possible quality of links. All users are given opportunity to comment on the quality of the HPR.
4. All links will be marked with quality and linkage rule flags, and the representativeness of linked samples can be compared statistically with the population in the decennial censuses since 1801 in order to estimate bias in the linked part of the population.

These principles ensure unique source references and the combination of data sources ensures increasing data quality over time. The HPR functions as an index to the source entries instead of replacing them. With regard to citations and transparency, the open HPR will be different from the national historical population registry in Iceland, deCode Genetics, which is closed in such a way that people only have access to their own ancestry and is somewhat limited and intransparent with respect to source references and criteria for record linkage. The historical, longitudinal databases in Sweden (the Demographic Database, the Stockholm Roteman Archive, the Scanian Economic Demographic Database), and the Historical Sample of the Netherlands have a better basis for linking in their pre-linked, original sources, but so far cover only parts of the population of these countries (Mandemakers 2000; Thorvaldsen 1998).

In the HPR a distinction is made between *linking* and *coupling*. Whereas *linking* identifies and determines that the same person is referred to in at least two different source entries (e.g. two censuses or two records in a baptism list), *coupling* combines information about relations between persons in a specific event, e.g. in a baptism entry or in a census household list. Linkage and coupling are interrelated processes, since family information can be used to link persons and information from several sources may be necessary to decide which persons are related.

All links will get a quality flag on a scale from 0 to 10, according to guidelines following to what extent there are unique and equal characteristics about a person in two or more sources. These characteristics are first name, family name, gender, age or birth date and birthplace. Address and occupations can be used to obtain uniqueness, and identifying related persons across two sources strengthen the quality of the links. The following grading system is applied:

10: Completely secure link, fulfilling criterion 8 plus identifying the person as part of a family.
8: The same birth date and same or similar names indicated by a high name comparison score, as well as geographical origin on the farm or street level.

6: The same or similar name of both spouses and geographic affiliation.
4: Same or similar name, age and related to the same place.
1: A probable link, the records should likely be linked, but further information is required for a certain link.
0: Established as linkage candidate with reasons that specify the uncertainly.

Figure 8.3 illustrates the two competing goals when linking: both to link as many individual records as possible, and to avoid linking entries that do not belong together (Johansen 2002). The ideal is represented by the figure's origo at the bottom left, where all linked entries really belong together and no true links are omitted. This is hardly ever possible to achieve with historical records because of imprecise and missing information. The problem is that when we impose stricter rules in order to reduce the risk of linking records belonging to different individuals (horizontal axis), we easily rule out several links that really should have been included (vertical axis). Conversely, if we introduce more liberal rules in order to ensure that all potential links are realized (moving down on the vertical axis in order to minimize the number of "true negatives"), we may include more links that really should be discarded (introducing more "false positives" and thus moving towards the right end of the horizontal axis).

Johansen assessed that while getting close to the origo is ideal, acceptable linkage results lie between points A and B on the curve in the diagram, with few accepted erroneous links (A) and few excluded appropriate links (B). Linkage results along curve II he deemed unacceptable because this introduced too many false links in the database and too many correct links were omitted. His main point in the HPR context was that Danish and Norwegian source material is quite similar.

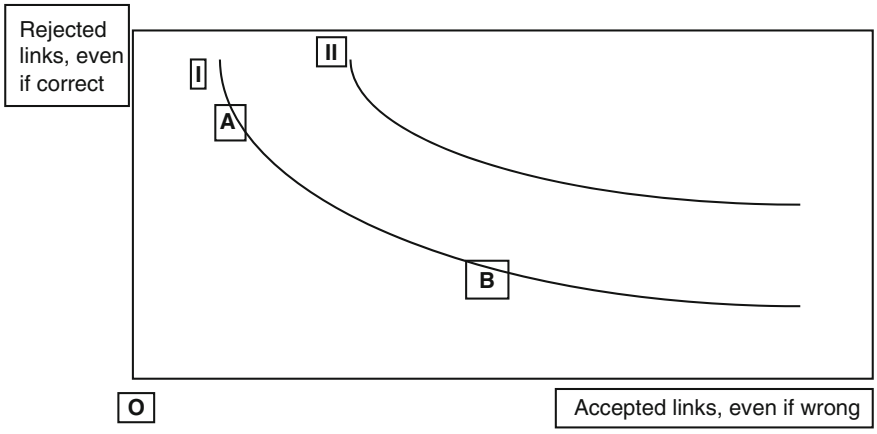

Fig. 8.3 The ratio of accepted links with errors ("false positives") and correct links that were rejected ("correct negatives") in sources from the early 1800s (*curve II*) and late 1800s (*curve I*). "*O*" marks the starting point of the *vertical* and *horizontal* axes—the origo. From Johansen (2002)

Starting with the second half of the nineteenth century, the source material has a precision that allows automatic linking along curve I, while sources from the earlier decades are so imprecise that they must be linked manually with auxiliary information as explained in the next paragraph. Fortunately, over the past decade we have developed techniques and methods that allow us to conclude differently, at least for substantial parts of the material.

First, we can now dynamically bring together variant spellings of the same name, both using standardized lists of names from the nineteenth century censuses developed by expert onomatologists, and algorithms for the comparison of strings (Alhaug 2011). Second, we have developed software which not only links individuals, but takes into account couples or entire families when linking. Both techniques are particularly valuable when dealing with a period when all information, including names and ages was imprecise or sometimes missing from the sources. By utilizing real-time name standardization and information on family relations, we can move a significant part of the population to be automatically identified in several sources from curve II to curve I in Fig. 8.3. Analysis of the 1801 census shows that about 80 % of women and 85 % of men were living with at least one other family member. Even if only a portion of these relations were stable over time, it still shows there is much potential in linking with group criteria. This is born out in our ongoing record linkage work where over a million persons have been linked in two or more sources, including many from the early nineteenth century.

In the next round, the problem of balancing the proportion of correct and incorrect links can be reduced further in the HPR by manual record linkage. The same quality and methods flag requirements apply to all linking, providing a quality indicator and justification for each link. Thus, researchers can make independent assessments and choose what links they will trust in their analyses. In summary, automatic record linkage has the great asset that documentation is inherent in the algorithms. The rules upon which the software is built can be spelt out in a database table, and references to this table can be a variable in the linked data set. This makes the links easier to trust, especially, when using material linked by others. However, there are decisions based on background knowledge and details in the sources which are not easy to automate: nick names used in certain families, knowledge about ancestors who cohabitated, names of neighbouring farms that were used interchangeably for cottars' places etc. Lists of property sales, probate registers and many other sources have been digitized only to a small degree, but have been used by genealogists to build their ancestral pedigrees. This wealth of information they can activate through record linkage crowd sourcing via the web.

Especially for the earliest period of the HPR we expect that a significant part of the linking and coupling will be done manually with contributions from genealogists and local historians volunteering. This is due to the simple structure of the source material with rather few variables, missing data and lack of consistency. It would be possible to automate much of this manual record linkage, but at a high cost. Typically, over 90 % of the investment in software would go into automating the 10 % special cases where human flexibility and special knowledge play a key

role. A key element in this work is how to motivate many users to provide high quality links and how these contributions are monitored and how the quality is assessed. Most importantly, it must be easy in order to find the potential linkage candidates to search for all the people featured in a specific source, all families in a municipality, and all priests in Norway or other groups of individuals, families and places. Most rural municipalities in Norway have published local community history books which list the ancestries on the farms systematically. In addition, the volunteers can search for persons' records among the chronological events in the church books, all events on a farm over time or other criteria. If in doubt, they are enable to flag person records as linkage candidates and ask for the opinion from others to confirm or reject the link. Next, the automatic and manual links can be informed by checking the HPR against their personal databases or genealogies. National coverage provides a final solution to linkage problems: When only one unique candidate record for linkage remains the link has been ascertained by the elimination principle.

This interactive web-based system lets the user, after logging in, search one or several sources during a specific period for names, age or birthdate, birthplace occupation and place of residence. Once candidate records are displayed, the user can check what records have already been linked automatically. These can be modified, and new links added. Background knowledge about the persons from newspapers can be accessed, and a module with information about tombstones is planned. A beta version of the software is available for testing—so far only in Norwegian.[5]

8.4.2 *Automatic Record Linkage: The Example of Lenvik Parish*

The Norwegian Historical Data Centre has developed software for the automatic linking of married couples and other family members, also utilizing special algorithms for comparing names.[6] In connection with the project to celebrate the constitution of 1814 these programs link the 1801 census and the church records (baptism, burial and marriage) for the northern part of Norway (Bråthen 2011). Figure 8.4 displays the average number of links between source entries for all men who became fathers during the period 1799–1815 in Lenvik parish south of Tromsø. Information about both father and mother was used to link the baptisms. The mean number of births per year was 41.5, however with a wide variation from one year to another, with the extremes of 16 in 1811 and 65 in 1806.

[5] http://hbr2.nr.no/demo/avisproject/avisprosjekt.php

[6] Developed in PL/SQL, the scripting language for Oracle databases. Names can be compared efficiently in real-time with the built-in Jaro-Winkler string comparison algorithm, aiming for similarity levels of over 0.8.

As a result of automatic linking, a rather consistent result emerges when looking at the average number of about four birth list links over time, strengthening our confidence in the linking. In the baptism records the father is consistently linked to the child and the mother, and it is possible to keep track of the couple from one birth to the other. Approximately two-thirds of all parents in the baptism list were linked to the marriage list, which is a stable proportion over time. Both spouses' names are in both lists and the priests were consistent with respect to spelling etc. since they copied information from baptism or confirmation to the marriage protocol.

The census represents only a snapshot of the population in Lenvik parish in 1801, and we typically find persons in different stages of their life. Thus, marital status and place of residence for a couple in the baptism/marriage lists does not necessarily match with the registration in the census. As a consequence, we get good linkage results in the years close to the census year, with a gradual decrease towards the end of the period.

Linking to the burial list clearly was the major challenge. A significant proportion of the burials only provide the first and last name of the deceased, which is usually considered as inadequate linkage criteria. Regarding the entry of child deaths, the lists fortunately provide information about relatives, usually the name of the father. Thus, the burial linkages shown in Fig. 8.4 are in most cases links between a father's entry into a baptism list and the death of his child. Of all buried children under the age of two, approximately 50 % were found in the baptism list. Accordingly, the relatively high linkage rates must after all be understood against the background of the high fertility and child mortality rates in this period (Hubbard et al. 2002).

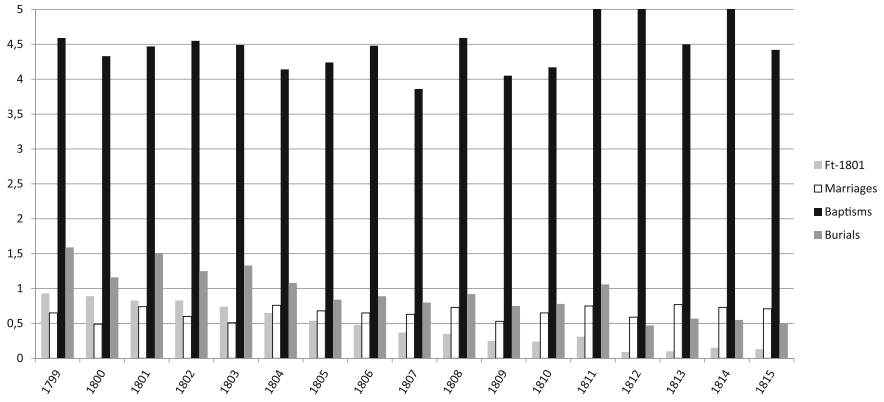

Fig. 8.4 Average number of links of fathers in between entries in the baptism list, the 1801 census (Ft-1801), the marriage list and the burial list. Lenvik parish in the period 1799–1815

About 40 % of all burials were not linked, and a large proportion of these were elderly men. They were typically registered without reference to relatives, often with first and last name only. From 1820 age information was usually included, but until the extensive revision of the church register that was undertaken by a royal resolution in 1877, providing a greater variety of data to link by, automatic and semi-automated techniques have clear limits. This is also visible when linking two censuses. Using semi-automated techniques, the 1865 and 1875 censuses were linked for parts of Troms province in Northern Norway. One-third of the inhabitants were not identified in both sources (Thorvaldsen 1995). In other words, we see a considerable scope for complimentary manual linking, employing the detailed knowledge of families and communities that exist in abundance in Norway. With the HPR's system for crowdsourcing via the Internet we activate this knowledge and resolve many of the problems individual investigators struggle with in vain.

8.5 A Ground Truth

To get an idea about the challenges we face and what kind of results can be expected from longitudinal databases at the local level we discuss the rather isolated parish of Rendalen on the border with Sweden. In Rendalen, church registers, censuses and other sources have been manually linked for the period from the local ministerial records that started in 1735 until 1950, when the last census before the one used to build the CPR was taken. The lowest linkage rates are in the eighteenth and early nineteenth century, before the regular, decennial nominative censuses started in 1866. As an isolated census without information about birthplace, the 1801 census increased the linkage rates more marginally. Of all the people that have been observed in the period 1815–1824, half were identified in both baptism, marriage and burial records, while another 20 % were identified in baptism and burial lists, but not in a marriage entry. In view of the high mortality in this period, we expect that a large number of people died unmarried (Bull 2006; Gjelseth 2000). Rendalen was a parish with little migration and the database is the result of thorough manual work on a small geographic area. We cannot expect equal coverage in municipalities with larger migration, but Rendalen can function as a "golden standard" or "ground truth" to evaluate the linking of sources in other parishes. Running the automatic record linkage software on the same source materials, is an interesting comparative exercise that still awaits completion.

Even after manual record linkage using the probate records, containing two generations as has been done for Rendalen parish, it is a challenge to reconstitute the population at a given time. These challenges are not diminished by the fact that mortality was high and that part of the population was geographically mobile during the turbulent times in the early nineteenth century. The high mortality meant that over one hundred marriages in Rendalen where recorded with widows or widowers

from 1801 to 1815. To determine how these settled together with children from previous and new marriages over the following decades is no simple puzzle. The next nominative census was taken in 1866, but some help is rendered by the silver tax lists from 1816 and the farm tax lists from 1838—at least for the heads of households.

8.6 Enhanced Research Opportunities with the HPR

We know from previous local studies and international experience that an HPR can be used in a variety of local, regional and national studies and provide a basis for international comparisons. The HPR promotes international collaboration through the NAPP (Thorvaldsen 2011) and the European Historical Population Samples network (EHPS-net).[7] Our database employs algorithms developed by our partners at the Minnesota Population Center to encode the family structure automatically by creating location variables based on information about family position, sequence number in the household list, gender, age and last name. These pointer variables can be analysed together with other constructed and encoded variables (Sobek et al. 2011).

The HPR provides a new historical and social science understanding of the relevant periods. With longitudinal microdata we can study how family structure and social and geographical mobility changed longitudinally as opposed to the snapshots given in the censuses. In a medicine and health perspective, the HPR will be an important source for studies of the population's gene pool and for instance genetic diseases related to consanguinity (Surén et al. 2007). The National Institute of Health, one of our major partners, has a collection of bio-samples which can be linked to the HPR. The bibliographies maintained by the Demographic Database at Umeå University and the Minnesota Population Center in Minneapolis contain a host of further research topics which can be studied in more detail also in Norway.[8]

Within local history and genealogy it will be easier to place people's own family and community history into a broader context. In practical terms, it will be easier to identify the sources and more efficient to link to other people's work on their ancestries. It will always be possible to find more information, more sources and comparable life histories. But it will be less necessary to retrieve the same individuals and duplicate the same links, and we can instead complement the work of others. The aim is a database where we can follow individuals, families, farms, homes and other locations over time. Not least, the HPR will have a source critical function. Only by linking the sources on the individual level, will it be possible to spot and evaluate the many errors and inconsistencies in the basic source materials. It is not trivial to establish such a population registry, and this chapter describes

[7] https://www.nappdata.org and http://www.ehps-net.eu

[8] http://www.nappdata.org/napp/ and http://www.ddb.umu.se

some of the challenges we face and how they are solved. Record linkage will be done both with automated algorithms and interactively via the Internet. Thus, the HPR represents new technology for collaborating to link personal data.

Acknowledgments We are grateful for contributions from other team member, especially Lars Holden and Torkel Bråthen.

References

Alhaug, G. (2011). *10 001 navn. Norsk fornavnleksikon*. Oslo: Cappelen Damm.

Bråthen, T. R. (2011). Det norske folk i 1814. (The Norwegian people in 1814). *Slekt og Data, 4*, 44–45.

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.* (pp. 25–34). Unpublished doctoral dissertation. Oslo, Norway: University of Oslo. http://www.rhd.uit.no/nhdc/Chapter%203%20Hans%20Henrik%20Bull.pdf. Accessed 14 Dec 2014.

Eikvil, L., Holden, L., & Bævre, K. (2010). Automatiske metoder som hjelp til transkribering av historiske kilder. (Automatic methods in transcribing historical sources). Notat SAMBA/44/10, Norsk Regnesentral. http://www.rhd.uit.no/nhdc/HBR_notat_okt-2010.pdf. Accessed 12 Jan 2015.

Engelsen, R. (1983). Mortalitetsdebatten og sosiale skilnader i mortalitet. (The mortality debate and social differentials). *Historisk Tidsskrift 62*(2), 161–202. Abstract in English: http://rhd.uit.no/ht/ht62.html#2169. Accessed 18 Mar 2015.

Fure, E. (2000). Interactive record linkage: The cumulative construction of life courses. *Demographic Research.* doi: 10.4054/DemRes.2000.3.11 http://www.demographic-research.org/volumes/vol3/11/3-11.pdf. Accessed 18 Mar 2015.

Gjelseth, M. (2000). *Relasjonsdatabaser som verktøy i en historisk-demografisk studie*. (Relational databases as a tool for a historic-demographic study). Unpublished master's thesis for master's degree. Oslo, Norway: University of Oslo.

Hubbard, W. H., Pitkänen, K., Schlumbohm, J., Sogner, S., Thorvaldsen, G., & van Poppel, F. (2002). *Historical studies in mortality decline*, II. Hist.-Filos. Klasse Skrifter og avhandlinger Nr.3. Oslo: Novus Forlag in association with the Centre for Advanced Study, at the Norwegian Academy of Science and Letters.

Høgsæt, R. (1990). Begravelsesskikker og trosforestillinger i det gamle bondesamfunnet - en feilkilde når en bruker de eldste kirkebøkene til å studere dødelighet? (Burial customs and ideas of faith in the old agricultural community—a source of error when using the oldest parish registers to study mortality?) *Historisk Tidsskrift*, *69*, 130–145. Abstract in English: http://rhd.uit.no/ht/ht69.html#2591. Accessed 18 Mar 2015.

Johansen, H. C. (2002). Identifying people in the Danish Past. In H. Sandvik, K. Telste & G. Thorvaldsen (Eds.), *Pathways of the past: Essays in honour of Sølvi Sogner on her 70th anniversary 15 March 2002* (pp. 103–110). Oslo: Novus.

Kjelland, A., & Sørumgård O.M. (2012). Databases constructed by the Norwegian extended family reconstitution method as part of a national population register for Norway. Paper presented at ESSHC 2012 9th European Social Science History Conference Glasgow, April 2012. http://tilsett.hivolda.no/ak/FamilyReconstitutionDatabases_HPR.pdf. Accessed 22 June 2015.

Mandemakers, K. (2000). Historical sample of the Netherlands. In P. K. Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (pp. 149–177). Minneapolis: Minnesota Population Center.

Nygaard, L. (1992). Name standardization in record linkage: An improved algorithmic strategy. *History and Computing, 4*, 63–74.

Quas, D. (2011). WeRelate: Suggestions/tweak to duplicates report. http://www.werelate.org/wiki/WeRelate:Suggestions/Tweak_to_Duplicates_Report. Accessed 4 Dec 2014.

Sobek, M. L. C., Flood, S., Hall, P. K., King, M. L., Ruggles, S., & Schroeder, M. (2011). Big data: Large-scale historical infrastructure from the Minnesota population center. *Historical Methods, 44*(2), 61–68.

Solli, A. (2006, August). *Urban space and household forms*. Paper presented at the eighth international conference on urban history, Urban Europe in Comparative Perspective, Stockholm, Sweden.

Surén, P., Grjibovski, A., & Stoltenberg, C. (2007). *Inngifte i Norge, Omfang og medisinske konsekvenser*, (Consanguineous marriage in Norway—prevalence and medical consequences), Folkehelseinstituttet. http://www.fhi.no/dokumenter/9b8f570dcd.pdf Accessed 20 Dec 2014.

Thorvaldsen, G. (1995). *Migrasjon i Troms i annen halvdel av 1800-tallet. En kvantitativ analyse av folketellingene 1865, 1875 og 1900.* (Migration in the province of Troms 1865–1900. A study based on the censuses). Unpublished doctoral dissertation, Registreringssentral for historiske data, University of Tromsø, Tromsø, Norway.

Thorvaldsen, G. (1998). Historical databases in Scandinavia. *The History of the family. An International Quarterly, 3*(3), 371–383.

Thorvaldsen, G. (2000). A constant flow of people? Migration in Northern Norway 1865–1900. *History and Computing, 11*(1–2), 45–59.

Thorvaldsen, G. (2008). Fra folketelling og kirkebøker til norsk befolkningsregister. (From censuses and church protocols to Norwegian population register). *Heimen, 45*, 341–359.

Thorvaldsen, G. (2011). Using NAPP census data to construct the historical population register for Norway. *Historical Methods, 44*(1), 37–47.