***Demographic Research*** a free, expedited, online journal
of peer-reviewed research and commentary
in the population sciences published by the
Max Planck Institute for Demographic Research
Doberaner Strasse 114 · D-18057 Rostock · GERMANY
www.demographic-research.org

***DEMOGRAPHIC RESEARCH***
VOLUME 3, ARTICLE 11
PUBLISHED 12 DECEMBER 2000
www.demographic-research.org/Volumes/Vol3/11

# Interactive Record Linkage: The Cumulative Construction of Life Courses

**Eli Fure**

# Interactive Record Linkage:
# The Cumulative Construction of Life Courses

**Eli Fure** [1]

## Abstract

In order to carry out demographic analyses at individual and group levels, a manual method of linking individual event records from parish registers was developed in the late 1950s. In order to save time and to work with larger areas than small parishes, systems for automatic record linkage were developed a couple of decades later. A third method, an interactive record linkage, named Demolink, has been developed even more recently. The main new feature of the method is the possibility of linking from more than two historical sources simultaneously. This improves the process of sorting out which events belong to which individual life courses. This paper discusses how Demolink was used for record linkage in a large Norwegian parish for the period 1801-1878.

[1] National Archives of Norway; eli.fure@riksarkivaren.dep.no

## 1. Introduction

The family reconstruction or reconstitution technique developed by Louis Henry and Michel Fleury in the early 1950s made more sophisticated historical demographic analyses possible. The technique, however, developed as it was for small French parishes, encountered difficulties when reconstructing families in bigger areas. The most important problem is the extremely time consuming manual procedure. The technique has also been criticized because stable, resident farmer families were more successfully reconstructed than more mobile families.

To counteract these deficiencies, different strategies have been adopted. In the 1970s and 1980s there were several attempts to automate family reconstitution, or what has become known as record linkage. Not only families but also individual life cycles were reconstructed. Some of these attempts were limited in scope, others more ambitious [Note 1]

The advantages of automatic record linkage are obvious: once the source material is computerized and the system is developed, record linkage is accomplished very quickly. Moreover, the algorithms make the linkage transparent and completely consistent. However, the programs will create more erroneous links than the corresponding manual procedures. Fully automatic systems cannot handle non-systematic errors in the sources. Where it is important that all the links are correct, as in genealogical and some medical studies, the results from fully automatic record linkage systems cannot be used. This is because such programs choose according to probabilities or at random when presented with more than one possible alternative.

The pioneers were optimistic about the future for automatic record linking, but the cost of developing and maintaining such systems has been so high that there are few systems still in use today. The systems are not so general that they can be used for all types of data. New source material and research projects demand comprehensive adaptations. Already in the 1970s some semi-automated procedures were developed, mostly conceived as preliminary or provisional pending the development of fully automatic systems or as small scale alternatives to the more costly automatic systems [Katz and Tiller 1972, M.P. Gutmann 1977.]

## 2. Demolink Compared to Automatic and Manual Systems

This paper looks at the application of an interactive record linkage system, Demolink, to a 19th century Norwegian data set, from a historian’s point of view [Note 2]. The system resembles earlier semi-automatic systems in some respects. Most important of these is that the linkage decisions are taken by the historian and not by predefined computer algorithms. It differs from them by the more efficient way the computer is used, by the flexibility with which different historical sources can be included, and by its user interface which enables the historian to work in a familiar environment, without expertise in computer science. Like them, Demolink offers an alternative to fully automatic systems, not only in respect to the costs but also to the methodology. The interactive technique is not merely a step towards a fully automatic system; it is considered a method in its own right.

In traditional family reconstruction, events are copied from the sources onto cards, sorted in different ways several times, and the information for each family is collected on a family card or sheet. The work is done manually, following rather mechanically fixed rules. Information is compiled chronologically and from one source at a time. At the very end, loose ends and problematic cases are considered and decided.

In automatic record linkage, event records are compared and possibly linked two by two. The degree of manual inspection of the results varies. The Canadian IREP institute uses such high quality sources that a basically automatic record linkage has proved successful [Bouchard 1986], but it also has a procedure for manual control [Bouchard 1996,33]. In the English sources the information is much poorer. The Cambridge Group’s record linkage system chooses according to probabilities or randomly when the information is not sufficiently discriminative [Schofield 1992]. The results are not revised manually. Because many of the cases are impossible to solve no matter what method is used, such a revision may seem unnecessary. Norwegian 19th century sources lie between the two extremes. There is insufficient information for a fully automatic linkage system to be used successfully, but the Cambridge approach would not exploit all the information found in the sources.

Demolink is a general system which can handle nominative, individual information from different sources. Its most important advantage is that it enables the historian to link records from several sources simultaneously. This more dynamic and flexible approach differs both from traditional family reconstruction and from earlier semi-automatic or automatic processes. Demolink enables the historian to approach the data in a more dynamic way. The most secure links are selected first, problematic cases are evaluated throughout the record linkage process, thus producing more secure life histories. The main disadvantage of Demolink is the time required to accomplish the linkage. Some results will be discussed towards the end of the paper to illustrate the success of the method.

## 3. The Data Material and the Demolink System

In my project Demolink is used on data from censuses, church books and land registers from the parish of Asker and Bærum, situated 5-30 km west of Oslo. The sources and dates in the study are shown in Table 1. Its purpose is to study various aspects of demographic behavior in the past.

**Table 1:**
Types and dates of sources from the parish of Asker and Bærum.

| Types | Dates | | | | | |
|---|---|---|---|---|---|---|
| Censuses | 1801 | 1815* | 1825 | 1835 | 1865 | |
| Church records | | 1814 | to | 1878 | | |
| Land registers | 1802 | | 1826 | 1838 | 1866 | 1886 |

* Only half of the parish, Bærum.

The choice of area was determined largely by the availability of sources. While the 1801 and 1865 censuses for Norway are nominal, the 1815 to 1855 censuses are normally only numerical. The Asker and Bærum parish, however, has quite good nominal drafts from the census takers in 1815, 1825 and 1835.

The period chosen was partly limited by sources already computerized, and the courtesy of public institutions involved in the project. The starting point was the 1801 census, already computerized at the University of Bergen in the 1970s. The final point was originally set to the census of 1865, available in machine readable form from RHD at the University of Tromsø. RHD also computerized the church records from 1814-1878 for this project. The censuses from 1815, 1825 and 1835 as well as the land registers from 1802-03 (manuscript), 1826, 1838, 1866 and 1888 (all printed) were computerized with the help of a graduate student.

The 1801 population in the parish was close to 4600 inhabitants, and in 1865 the number of inhabitants had risen to about 8400. The total number of individual event records was about 100,000. The Demolink system facilitates linkage of the records from all sources simultaneously by presenting all the *individual event records*, i.e. each time a person is mentioned in a source, in one *individual event file*. The number of entries in the event file was about 125,000 because people with more first names or more surnames were listed as many times as they had different names. Thus an Anne Kari Hansdatter would be listed both among the Anne Hansdatters and the Kari Hansdatters.

Neither the software, the data preparation, the production of files, nor the design of the system with its user interface, will be presented and discussed in this article. Only the most basic information needed for understanding the use of the system will be given. The computer's part is to calculate, (re-)organize, combine, check, store, retrieve and present the data, while the thinking is left to the historian.

Demolink displays the data principally in two windows, shown in Figure 1. The large, upper window, the *individual event window*, shows a part of the individual event file, which contains records with the first name, patronymic, residence, year of birth (either given or calculated), role in the event (such as groom, father, mother, baptized child, presence in census or land register). The lower left corner of the screen shows the source entry window, a window presenting complete information from the source entries. A useful pop-up window presenting linked life courses is not shown in the figure.

In the individual event window, those records selected by the historian as pertaining to the same person are shown in inverse video. The source entry window gives more detailed information from the sources. Fig. 1 shows three of the events in the life course of Gunnild Sørine Pedersdatter, in this case her baptism, presence in the 1835 census, as well as her marriage.

The individual event file can be presented in different ways according to which sorting criteria are used. The version mostly used was the one where the individual event records were sorted alphabetically after first name, patronymic, residence and year of event. Before sorting, the names were standardized to a code by a separate name standardization system called Foneq [Nygaard 1992]. The standardization is partly etymological, partly phonetic. The names in the sources, not the codes, appear on the screen, e.g. the first names Embret and Ingebrigt receive the same code, and they are sorted to the same place in the file even if the spelling is quite different. The linkage proper was done by selecting individual event records, and then storing them as linked individuals. Each individual received a distinct person number generated by the computer.

## 4. Record Linking Strategy

In Demolink, all individual events are first interactively linked to individual life courses. Afterwards an automatic procedure groups the linked individuals into families. The record linkage was executed, first for all males, then for females. The pocketing or grouping variables were first names and patronymics. Each section of a particular combination of a first name and patronymic was linked before I went to a new combination.

Men were linked first because men had more records than women. More records often mean more identification items and consequently more secure links. Mainly men were mentioned in land registers. The fathers, never the mothers, of the bride and groom were mentioned in the marriage records. When I started to link the females, the ready linked life courses of their fathers and husbands were useful.

To avoid starting with the problems connected with common names, male names with infrequent initials, such as B and D, were first taken up. Thereafter the males were linked alphabetically from A through V, then the females in the same way, except for the names Mari/Marie/Maria/Maren (all standardized to the same name code). These variants were the last to be linked, because they were the most common female name. They also occurred frequently as the second Christian name, so when records from this section of this file were to be linked, many of them were found to be linked already under the first Christian name. This was so because when an individual event record with two first names was linked at one place in the file, the duplicate record, sorted and listed somewhere else in the file, was automatically marked as linked too.

The records belonging to individuals with rare names practically stood out by themselves by virtue of the visual pattern of records in the individual event file. For the most frequent combinations of first name and patronymic, there could be hundreds of individual event records and consequently no overview on the screen. These cases needed a complementary paper print-out of the file. Notes, brackets and arrows prepared the proper linkage on the screen.

When the number of individual event records associated with a particular combination of a first name and patronymic exceeded the screen's capacity, i.e. 30 records, I started with an individual event record type connected to a marriage. It could be a person who was mentioned as father of the bride or the groom, or it could be the groom or bride. Because so much information was connected to marriages, in most cases it was possible to find links to both the family of origin and the family of procreation. By choosing the most informative records first, it made it easier to see where records with less abundant or discriminative information should fit in. Moreover, by taking the people who were mentioned in a marriage first, a substantial number of the records were linked, thus the number of records that initially seemed to fit into different life courses were reduced.

During subsequent examinations of the same combination of equal first name and patronymic, records for people who were married, but where the actual marriage record was missing, were linked. Thereafter unmarried people were linked. Finally, before introducing a new combination of a first name and a patronymic, a last check of missed links in the stored life courses was done. Also, cases with two or more alternative records that might fit into a particular life course, were settled, if possible. The process of choosing between competing links closely resembles the procedure explained for French-Canadian material [Jetté 1989].

In contrast to the traditional manual method, where the events are added chronologically as they occur, Demolink encourages retrospective record linkage. The retrospective strategy offered the opportunity to establish the most obvious links first, thus the number of uncertain links were reduced. If the starting point had been a baptism, the possibility of moving out or dying would have had to be kept in mind before linking the baptism to a later census or a marriage.

The record linking process was also in a certain way iterative, in that stored life courses could always be subject to changes if I found information that altered previously linked life courses. Stored life courses were infrequently split into two separate parts later. It happened more often that the linkage revealed that two or more stored life course fragments belonged to the same individual. The tendency to get fragments rather than erroneous events in the life course is due to inaccuracies or errors in the sources [Note 3]. The record linkage was thus a cumulative process, where insights gained in linking one person could be exploited to link other people, partly by removing apparently competing records, but also by discovering errors in the sources.

## 5. Dealing with Errors in the Sources

Errors in the sources, and particularly errors in names are obviously detrimental to correct record linkage. There is no reason to believe that the data set from Asker and Bærum was unusually replete with errors, but Demolink gave good possibilities for finding and correcting errors, either by looking more closely at a record in its source context, or by checking wholly established life courses of people related to the person in question.

The possibility of viewing the same source entry record from as many angles as there were people mentioned in the record, increased the chances for disclosing errors. The example of Johannes Johannesen may illustrate this. The point of departure for this linkage was his presence in the 1865 census. Johannes, according to his age in 1865, should have been born in 1841, but there was apparently no suitable candidate baptism record in the years 1840-1842. Now the possibility in Demolink to show the previous or next source entry was useful. In the household following that of Johannes Johannesen in the 1865 census, I found a woman, Mari Monsdatter, reported to be the mother of Johannes Johannesen. The section of the file listing the records pertaining to the Mari Monsdatters, was consulted, to see whether there was a Mari Monsdatter as mother at baptism of a child called Johannes. There was no such record, but many records belonging to a Mari Monsdatter married to a Johannes Nilsen. This Johannes Nilsen provided further clues. By looking up the section of the file listing the Johannes Nilsen individual event records, I found many reciprocal events to those belonging to Mari Monsdatter. He was naturally listed as father at the baptisms, where she appeared as mother. In addition there was one where he was listed as father, but the mother’s name was Mari Hansdatter. This baptism was in 1841, but the child's name was not Johannes, but Johanne, the female form of the name. This meant that I finally found the baptism record for Johannes in the female section of the individual event file, among the records with the name Johanne Johannesdatter.

There were two errors in this example. One was already there in the church book, namely the wrong patronymic of the mother. The other was an error introduced by the computerization of the source. The last ‘s’ in Johannes was almost invisible, so the name was interpreted as Johanne. No automatic record linkage system would have been able to resolve these errors, and the link would therefore have been missed.

Since first name was the first sorting variable in the individual event file, the linking strategy was very vulnerable to errors in the first name. Many such errors were indirectly discovered during the work with this file. Errors were difficult to discover for the person to be linked, but could be revealed for a related person.

By the progressive disclosure of data errors in the data set, more and more records were accepted as reliable parts of a life course, despite inaccuracies in names and ages. After the whole individual event file had been worked through, the first links were reviewed. Links that had not been made in the beginning were now established. The reviewing of links was stopped when the return in the form of more complete life courses was negligible. The linking process is thus also a learning process for the historian. This iterative process would not be possible in traditional family reconstitution.

## 6. Linkage Criteria

In his classical manual, E. A. Wrigley clearly advocates the use of all context information in the family reconstitution process, residence as well as occupation are no exceptions. [Wrigley 1966, 130]. The same point is made in a Norwegian handbook in historical demography [Dyrvik 1983,110]. Using the same variables first in the linkage process and then in the substantive analysis is, however, problematic [Wrigley and Schofield 1973,90-91]. Even if the research goals are purely demographic, the overall results can be biased if the reconstructed population differs from the real population. IREP in Canada recommend, therefore, to use only information that is stable for the whole life course in the linkage process [Bouchard 1992,69]. Still, the Cambridge Group has used residence as well as occupation in the automatic reconstitution [Schofield 1992] The same has been practiced in Sweden [Bengtsson & Lundh 1993].

The linkage criteria for Asker and Bærum were the names (both of person to be linked as well as relatives), the year of birth, and the residence, in that order. If the first name was very common, this reduced the importance of this criterion, but only seriously if also the patronymic was among the most frequent. But even in these cases, the age would discern different people. Where age was not given, or several records showed approximately the same age, then identical residence would do in most cases. In this study residence is the area that belongs to a specific land register number. This area typically covers one or two main farms and some rented cottages. Residence is thus a quite limited area. The chance that there should be two different people, with the same name and age that lived at the same place at two consecutive points in time, the first having left and the second having moved in, is small indeed.

To exclude residence from the linking criteria would skew the results unduly. For people with less common names, the names and age were good enough linkage criteria. They would be linked whether they were movers or stayers. Without residence, however, the linkage result for people with common names would often be either a false link or an incomplete life course.

Clearly any study based on family reconstitution benefits from the reconstruction of as many reliable life histories as possible, both for those who stayed at the same place and for those who moved within the parish. Often, families would stay at the same place for some years and then move, stay there for some time, and then move on, or move back to the first place. Such patterns would be much more difficult to disclose without residence as a linkage criterion. Long-distance migrants will be lost in any case.

There were, however, times where several individual event records competed to be included in a particular life course, and it was impossible to know which record to choose. Other times no record seemed to fit into an unnatural “hole” in a particular life course. This could be when a person was mentioned in the sources both before and after a census, but was not found in the census. In the case of competing records, two different strategies have been chosen in automatic record linkage to resolve clusters of records for people with equal names [Bouchard 1992,70]. My preference was to follow the philosophy of IREP: To prioritize optimal accuracy over optimal completeness; i.e. it was more important that the links were secure than that a maximum of the individual event records were linked. However, unlike the automatic systems which only compare two links at a time, I could consider several links simultaneously. As such the question of accuracy versus completeness was not generally pressing. The most difficult and the most time consuming individual event records to link were those with common names. Certain combinations of very common first names and patronymics e.g. Hans Olsen, generated up to several hundred individual event records.

## 7. Linkage Results

### 7.1 Completeness of life courses

When records from multiple sources are linked simultaneously, it is impossible to know the definite answer to the question: How many times does each person appear in the sources? One way to evaluate linkage success is to analyse single individual event records that are still not linked, to try to find patterns among them as a possible approach to an explanation of why they are not linked. An overview is given in Table 2.

**Table 2:**
Individual event records, total numbers and total not linked according to type of event.

| | Number of individual event records | | |
|---|---|---|---|
| **Types of individual event records** | Total | Not linked | % Not linked |
| Stillbirths | 462 | 462 | 100 |
| Children baptized before 1866 | 11309 | 1755 | 16 |
| Children baptized after 1866 | 3474 | 2946 | 85 |
| Father of baptized child | 15232 | 718 | 5 |
| Mother of baptized child | 15167 | 637 | 5 |
| Groom | 3365 | 332 | 10 |
| Bride | 3365 | 158 | 5 |
| Groom’s father | 2844 | 828 | 29 |
| Bride’s father | 2838 | 616 | 22 |
| Wedding witness | 1039 | 99 | 10 |
| Person buried | 7998 | 1226 | 15 |
| Spouse of buried person | 946 | 26 | 3 |
| Father of buried person | 2568 | 154 | 6 |
| Mother of buried person | 174 | 45 | 26 |
| Landowner in land register | 3193 | 348 | 11 |
| Present in 1801 census | 4593 | 1868 | 41 |
| Present in 1815 census | 2993 | 391 | 13 |
| Present in 1825 census | 5658 | 310 | 5 |
| Present in 1835 census | 5759 | 269 | 5 |
| Present in 1865 census | 8405 | 904 | 11 |
| Total | 101382 | 14092 | 14 |

All together 14 % of the individual event records were not linked, or to put it in another way: Of the 31230 person numbers generated by the computer, 45% were registered with one event only. This seems to be a high number, but the composition of the raw data (see Table 1) explains many of these non-links. An event type which stand out with a high number of not linked events, are children at baptism. For children born after 1865 the reason why most of them are not linked is obvious. As there is no computerized census after 1865, children born after 1865 will, unless they die, necessarily have only one event before 1878 in their life course. But even before 1865 the percentage not linked is quite high, 16%. This is because there were no censuses between 1835 and 1865. Among the baptisms in the period 1836-45, 33 % are not linked. Generally it can be observed that events from periods where there are many sources are most easily linked, see e.g. the censuses of 1825 and 1835.

The 1801-census, on the other hand, contains many people not linked. Since the computerized ministerial records only start in 1814, and as the census for 1815 only covers a part of the parish, the 1801-records of people who died in the period 1801-1813 are not linked. We can estimate the number by using national demographic information. The age composition of the 1801 population in Asker and Bærum was not at variance with the whole Norwegian population. The average crude death rate for the years 1801-1813 on the national level was slightly above 25 per thousand. Given the same rate in Asker and Bærum, this means that about 1300 of the 1868 not linked people had probably died by 1814. Most of the rest had moved out.

Other events which stand out with a high percentage of no links are the fathers at the wedding, 29% of the fathers of the grooms, 22% of the brides' fathers. The difference is explained by the fact that marriages more often took place in the home parish of the brides than of the grooms. Many of the fathers never lived in the parish. Owners of land also belong to the easily explained 'life courses' with only one record. These must belong to absentee landowners, or to people who moved in after 1865. The last land register was from 1888.

Among the deceased 15% were not linked, altogether 1226 persons. This is probably where the problem of underlinkage is the greatest. The entries in the burial lists sometimes give the name of the husband of a deceased woman, but never vice versa. Age tends to be inaccurately stated at old age. During the linkage process there were several examples of life courses where the age implied that the person in question should be dead within the period of study, but where there was either no apparent candidate record, or there were several equal competitors.

Stillbirths were registered for analytical purposes, but of course there will be only one event record for stillborn children. A preliminary study of infant mortality also revealed that many children who died soon after their birth were only registered in the burial and not the baptismal lists in the church books.

The seemingly high percentage of non-links, 14 %, can thus largely be explained by the composition of the sources used. If we accept that children born after 1865, most of the fathers of brides or grooms and people who are just mentioned in a land register, only have this event in their life course in the Asker and Bærum records, this amounts to almost 5000 people. The stillbirths as well as the estimated number of deaths before 1814, adds 462 and 1300 respectively, to the number.

Among the other more than 7500 not linked events (7%), there are surely some that are underlinked, either because of errors that were not discovered, or because there were too few or too weak identification items in the records. The main reason why they are not linked must, however, be that the people moved out of the parish. The parish is situated close to Oslo, the capital, and this must have been important for migratory movements. Norwegian church books in the 19th century list migrants in and out of the parish, but they notoriously understate the number. For Asker and Bærum the migration registers were not computerized, but manually counted from 1815-1865. According to the church books the total number of outmigrants was 2311.

### 7.2 Name frequency and linkage results

It was easier to link people with uncommon than common names, but how did name frequency affect the result of the linkage? All the first names in the data material except the people born after 1865 were counted. Double names were counted twice, once for each name. The total number of names were 116,112, distributed among 1052 different male and 1321 female names. The names were grouped into three categories, common, average and rare names. People with double first names formed a fourth category. Each name category contained approximately the same number of people. Among the people with a single common first name, there were about 45 % with only one event in the life course. The frequencies among those with average or rare names were 6-7 percentage points lower. Of those with double first names 32% had only one event in their life course. Thus there is a tendency that people with common names are overrepresented among the non-links. There are, however, many disturbing factors in such an analysis. One is that there was more sharing of the same names at the beginning than at the end of the period. The space and purpose of this article does not allow a more thorough presentation [Fure 1990a, 1990b].

In general, the variation in the number of events linked according to name frequency is not alarming. However, when so many people share the same first name, this is a problem even with an interactive system like Demolink.

### 7.3 Social group and linkage results

Earlier family reconstruction studies have been criticized because stable families have been more successfully reconstructed than the more mobile ones. Such a bias is likely to have consequences for analyses of other types of demographic and social behavior. Stability in this period and context is related to wealth and higher social class: farmers moved less frequently than cottars, craftsmen and workers.

The social status of the heads of households given in the Norwegian sources was coded into three different social groups. Group one consists mostly of farmers who owned their land or tenant farmers; group two consists of persons who were craftsmen, masters of small vessels and small-scale traders; group three consists of workers, servants, fishermen, cottars, sailors, as well as people who received poor relief. A social grouping of a person for the whole lifetime was not made, because of the problem of handling changes in social status. In order to examine the relationship between the number of events in the life course and social group, the social status at a specific time must be chosen, i.e. in one of the censuses. The relationship was examined for all of the censuses, and the pattern was homogeneous. Table 3 shows the information for the 1835-census [Note 4].

**Table 3:**
Number of events in the life course of heads of households in 1835 by social group.

| | Number of events in the life courses of heads of households in 1835 | | | |
|---|---|---|---|---|
| **Social group** | 9 events or less | 10 - 15 events | 16 or more events | Total |
| High | 87 (27%) | 117 (36%) | 120 (37%) | 324 (100%) |
| Middle | 75 (32%) | 80 (34%) | 78 (33%) | 233 (100%) |
| Low | 197 (39%) | 173 (34%) | 133 (27%) | 503 (100%) |
| Total | 359 | 370 | 331 | 1060 |

We see there is a slight tendency that the farmers have more events registered than the heads in the lower social groups. This may not be an effect of a poorer linkage result for the lower social groups, it might just as well be that there were more events or more registered events in the farmers' life courses. Age at marriage was lower for farmers than cottars, thus they might have experienced more baptisms and more burials. They also remarried more often. On the whole the difference between the groups is not dramatic, and it seems safe to conclude that Demolink's interactive linkage of records from all the sources simultaneously, has given good results for all social groups.

## 8. Time Requirements of the Linking Process

The good results are not achieved without investment. The linking of the 100,000 events required about 2400 hours in front of the computer. In addition comes the time to computerize the sources. Some of this work I did myself; some was done by a graduate student, some was done by other agencies. The latter data had to be converted and adapted to the Demolink system. How much time was spent on these parts of the project is unknown.

The average amount of time spent on the interactive linking of one individual with more than one event in the life course was more than 8 minutes. The record linkage required high concentration, and the hours in front of the computer were quite tiring to eyes and brain, five hours a day was the average. Most life courses were linked very smoothly and quickly, very often in 2-3 minutes. Linking people with common names consumed much more time.

It is not easy to find comparable estimates on time spent on traditional family reconstruction; each project seems to have its own sources and its own consumption of time. An investigation for a population of 1000 - 1500 people for 100 years is in Norway considered to give more than a year, i.e. more than 1750 hours, of work. E. A. Wrigley estimates 1500 hours for English parishes of a size of 1000 people for three centuries. J. Dupâquier reckons 2 hours of work for each family of type MF (date of marriage and end of union known) in France [Dyrvik

1983,95, Wrigley 1966,97, Dupâquier 1984,11]. The reason why the Norwegian estimate is higher than the English may be that not only church records, but also census returns and other nominative sources are frequently used in Norwegian studies.

At first glance, then, the worktime saved by using the Demolink system seems marginal. The figures for time spent in manual family reconstitution, however, are, all based on much smaller data sets than the one from Asker and Bærum. The amount of time spent on linkage increases exponentially with the size of the parish. So, there is definitely time to be saved with Demolink. Even more important is that with such a large amount of data, it would be impossible to maintain an overview while using manual methods. Finally the data is ready for efficient analyses with the aid of a computer.

## 9. Concluding Remarks

As an advocate of a fully automatic record linkage process, Roger Schofield of the Cambridge Group has claimed that “if the historian’s judgment has any claim to intellectual respectability, the principles on which it is based must be specifiable in algorithmic form and so be executable automatically by the computer *without further human intervention.*” [Schofield 1992,75]. The experience with the record linkage is, however, that insight into the sources and their historical context is necessary to obtain good results in record linkage. This means that the researcher/linker must know the geography of the area, understand which names, even if not standardized, are so close that they are identical, and finally understand what constitute clear errors. In addition, human qualities like curiosity, imagination, alertness and attention are important. This does not mean that only university educated historians with an interest for this type of work could successfully use this record linkage system. Well educated amateur local historians could be also trained to do the job, but it is definitely not a routine, mechanical task. It is also an advantage that the historian who will examine the sources for substantive analyses also does at least some of the record linking. The work gives one a unique acquaintance with the sources and the data set, and many questions that can later be subject for more systematic research will arise.

The number of identifying items in the sources is also important for record linkage. The good quality of the Norwegian 19th century sources clearly contributed to the success of the interactive linkage. But even when the records contained poor or erroneous identification items, the possibility of viewing many sources simultaneously offered by Demolink made it possible to find the life courses to which the events belonged.

Since the different sources did not completely overlap in time, an expected result was less complete life courses. The case will often be that it is not solely up to the historian to select which sources to be used. Budget, availability and time constraints are legion. Given these constraints, the linkage strategy I chose, standardized first names and patronymics as pocketing variables, males first then females, both basically in alphabetical order, worked reasonably well. An alternative solution could have been to link the people with uncommon names first, regardless of sex. This might have eased the task of linking people with common names. Those of them who were married to people with uncommon names, could more quickly be linked by looking up the life courses of the related people. On the other hand it would not be easy to keep track of what had been done.

Whereas the order of linkage could have been altered, the cumulative and at least partly retrospective linking strategy within a pocket of events belonging to people with the same standardized first names and patronymes, seems more fixed. Rather than starting with a birth and thereupon adding events chronologically, the procedure of starting preferably with a marriage and subsequently adding events with fewer identifying items, seemed to be the most efficient way to produce secure life courses.

The main part of the Demolink system was developed before I could start the actual record linkage, but the process gave ideas as to how the system could be improved. Some of these ideas were implemented during the linking process, like sorting the individual event record file in different ways. Using patronymic or residence as the first sorting key, gave other views of the records, sometimes making it easier to see which records belonged to the same life course. This was particularly useful in the cases where the first name was wrong or lacking. The possibility of linking people by typing their individual event record number, i.e. not necessarily having them on the screen of the computer, was added. Production of lists of not linked records to help complete the final linking of people with common names was also useful.

Other ideas that arose during the record linkage process demand more fundamental changes. Among these is the need for better search routines in order to find candidate records for linkage, when problems like errors or missing information split records that ideally should have appeared close to each other. Some kind of a more automatic record linking procedure, to reduce the time in front of the computer linking the “easy” cases without giving up the claim of correctness, would be welcome. There is, however, always the problem of how to detect and handle non-systematic errors in the sources.

The interactive method, based on the researcher’s knowledge of the sources, historical background as well as cognitive agility, has a qualitative aspect. But it does not necessarily follow that it is less respectable intellectually than the strict applications of formal algorithms. In history there are many ways to evaluate a piece of work. In his seminal article of 1973, Ian Winchester argued that all history basically is “speculation about the past controlled by record linkage operations” [Winchester 1973]. If this is true, it is not obvious that record linkage always should be based on a fully quantitative approach. The method, manual, interactive or fully automated, should be chosen and evaluated according to the goals for the research, the resources available, as well as the quality of the sources.

## 10. Acknowledgements

The work was supported by the National Research Council for the Humanities as well as the University of Oslo. I would like to thank Lars Nygaard, Gunnar Thorvaldsen, William Hubbard, Lars Østby and Stein Sjølie as well as colleagues at the National Archives of Norway for their valuable comments and support.

## Notes

1. The most well-known today are the two Canadian projects, the PRDH at the University of Montréal, the IREP (former SOREP) at the University of Québec at Chicoutimi, connected with other Canadian Universities, and the English Cambridge Group for the History of Population and Social Structure.
2. The Demolink system was developed by the computer scientist Lars Nygaard, at the Department of Informatics at the University of Oslo. There was a close cooperation with historians throughout the development of the system. A more thorough presentation of Demolink's theoretical basis, methodological and technical solutions, and user interface will be published by Lars Nygaard later.
3. An automatic record linkage project in Sweden experienced the same tendency: Even with a hard phonetic name standardization, the result was that almost 20% of the 256 families were incorrectly split up, while only one was incorrectly amalgamated (Bengtsson & Lundh 1993).
4. 56 heads of households had no code for social group. Individual event records from the land registers were excluded from this analysis as, per definition, they only pertain to farmers.

## References

Bengtsson, T. and C. Lundh. (1993). “Name-standardisation and automatic family reconstitution.” *Lund Papers in Economic History. Department of Economic History, Lund University* 29:1-24.

Bouchard, G. (1986). “The processing of ambiguous links in computerized family reconstruction.” *Historical Methods* 19:9-19.

Bouchard, G. (1992). “Current issues and new prospects for computerized record linkage in the province of Québec.” *Historical Methods* 25:67-73.

Bouchard, G. (1996). Annual report 1995-1996. Chicoutimi (Québec):IREP.

Dupâquier, J. (1987). “Pour un rajeunissement des monographies paroissiales.” *Annales de Démographie historique dh. Bulletin d’information* 48:10-28.

Dyrvik, S. (1983). Historisk demografi. Bergen:Universitetsforlaget.

Fleury, M., and L. Henry. (1956). Des registres paroissiaux à l’histoire de la population. Manuel de dépouillement et d’exploitation de l’état civil ancien. Paris: Editions de l’INED.

Fure, E. (1990a). “Oppkalling og familiementalitet.” *Historisk tidsskrift* 69:146-162.

Fure, E. (1990b). “Personnavn og tidsånd.” *Namn og Nemne, Tidsskrift for norsk namnegransking* 7:35-55.

Gutmann, M.P. (1977). “Reconstituting Wandre. An approach to semi-automatic family reconstitution.” *Annales de Démographie Historique*: 315-41.

Jetté, R. (1989). “Preuve de l'identité des quatre couples homonymes Louis Tremblay et Ursule Simard. *Mémoires de la société généalogique canadienne-francaise* 40:18-33.

Katz, M. and J. Tiller. (1972). “Record-linkage for everyman: A semi-automated process.” *Historical Methods Newsletter* 5:144-150.

Nygaard, L. (1992). “Name standardization in record linking: An improved algorithmic strategy.” *History & Computing* 4:63-74.

RHD homepage http://www.rhd.uit.no/.

Schofield, R. (1992). “Automatic family reconstitution - the Cambridge experience.” *Historical Methods*, 25:75-79.

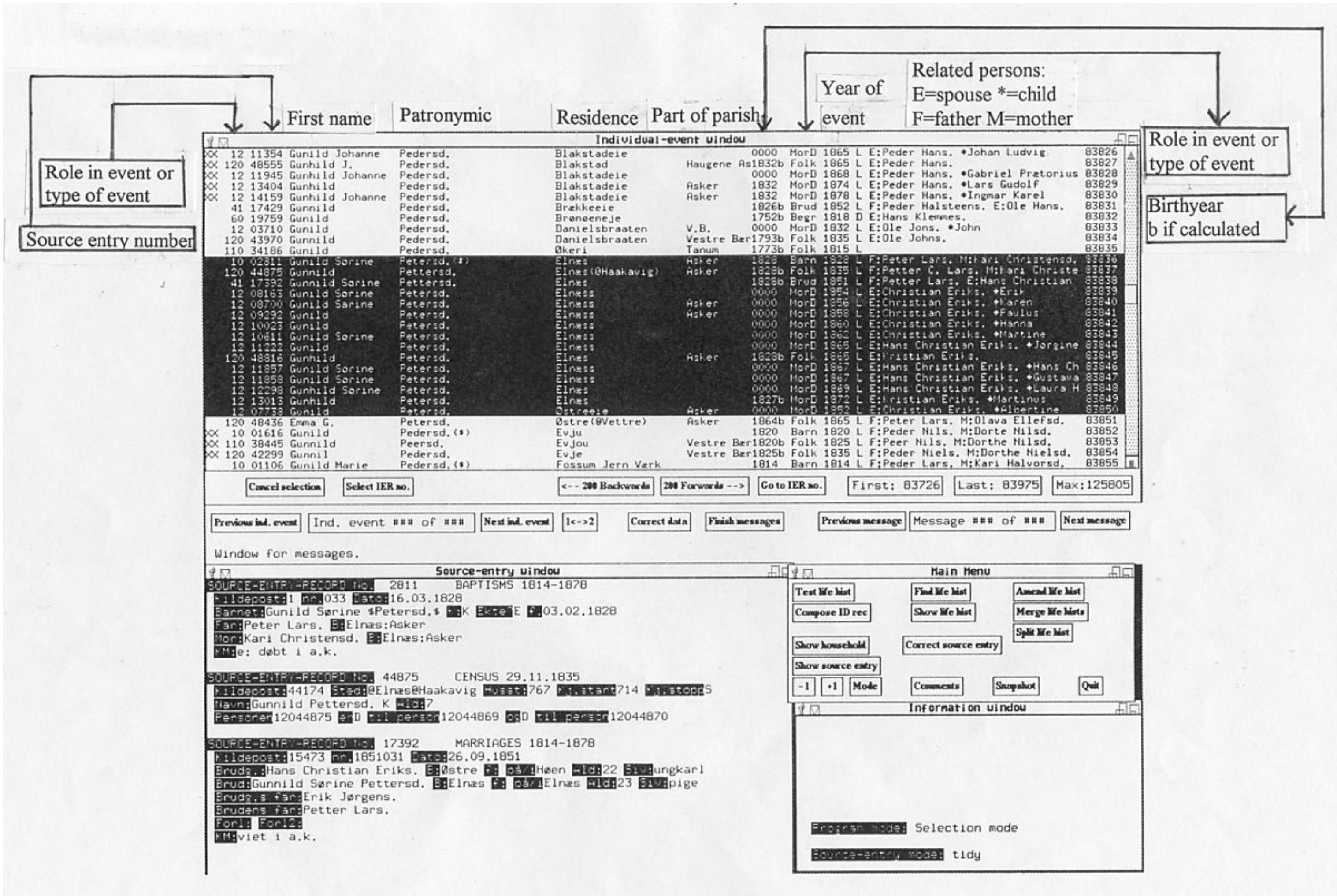
Winchester, I. (1970). “The linkage of historical records by man and computer: Techniques and problems.” *Journal of Interdisciplinary History*, 1: 107-24.

Winchester, I. (1973). On referring to ordinary historical persons. In E. A. Wrigley, editor. Identifying People in the Past. London. Edward Arnold.

Wrigley, E. A. (1966). Family reconstitution. In E.A. Wrigley, editor. An Introduction to English Historical Demography, London: Weidenfeld and Nicolson.

Wrigley, E. A. and R. S. Schofield. (1973). Nominal record linkage by computer and the logic of family reconstitution. In E. A. Wrigley. Identifying People in the Past. London: Edward Arnold:64-101.

**Figure 1:**
The display of data in Demolink



```
Individual-event window
XX  12 11354 Gunild Johanne  Pedersd.    Blakstadeie                 0000  MorD 1865 L E:Peder Hans. *Johan Ludvig        83826
XX 120 48555 Gunhild J.      Pedersd.    Blakstad         Haugene As1832b Folk 1865 L E:Peder Hans.                       83827
XX  12 11945 Gunhild Johanne Pedersd.    Blakstadeie                 0000  MorD 1868 L E:Peder Hans. *Gabriel Prætorius   83828
XX  12 13404 Gunhild         Pedersd.    Blakstadeie      Asker      1832  MorD 1874 L E:Peder Hans. *Lars Gudolf         83829
XX  12 14159 Gunhild Johanne Pedersd.    Blakstadeie      Asker      1832  MorD 1878 L E:Peder Hans. *Ingmar Karel        83830
    41 17429 Gunnild         Pedersd.    Brækkeeie                   1826b Brud 1852 L F:Peder Halsteens. E:Ole Hans.     83831
    60 19759 Gunild          Pedersd.    Brønøeneje                  1752b Begr 1818 D E:Hans Klemmes.                    83832
    12 03710 Gunild          Pedersd.    Danielsbraaten   V.B.       0000  MorD 1832 L E:Ole Jons. *John                  83833
   120 43970 Gunnild         Pedersd.    Danielsbraaten   Vestre Bær1793b Folk 1835 L E:Ole Johns.                        83834
   110 34186 Gunild          Pedersd.    Økeri            Tanum      1773b Folk 1815 L                                    83835
    10 02811 Gunild Sørine   Petersd.(#) Elnæs            Asker      1828  Barn 1828 L F:Peter Lars. M:Kari Christensd.   83836
   120 44875 Gunnild         Pettersd.   Elnæs(@Haakavig) Asker      1828b Folk 1835 L F:Petter C. Lars. M:Kari Christe   83837
    41 17392 Gunnild Sørine  Pettersd.   Elnæs                       1828b Brud 1851 L F:Petter Lars. E:Hans Christian    83838
    12 08163 Gunild Sørine   Petersd.    Elnæss                      0000  MorD 1854 L E:Christian Eriks. *Erik           83839
    12 08700 Gunild Sarine   Petersd.    Elnæss           Asker      0000  MorD 1856 L E:Christian Eriks. *Karen          83840
    12 09292 Gunild          Petersd.    Elnæss           Asker      0000  MorD 1858 L E:Christian Eriks. *Paulus         83841
    12 10023 Gunild          Petersd.    Elnæss                      0000  MorD 1860 L E:Christian Eriks. *Hanna          83842
    12 10611 Gunild Sørine   Petersd.    Elnæss                      0000  MorD 1862 L E:Christian Eriks. *Martine        83843
    12 11222 Gunild          Petersd.    Elnæss                      0000  MorD 1865 L E:Hans Christian Eriks. *Jorgine   83844
   120 48816 Gunhild         Petersd.    Elnæs            Asker      1828b Folk 1865 L E:Kristian Eriks.                  83845
    12 11857 Gunild Sørine   Petersd.    Elnæss                      0000  MorD 1867 L E:Hans Christian Eriks. *Hans Ch   83846
    12 11858 Gunild Sørine   Petersd.    Elnæss                      0000  MorD 1867 L E:Hans Christian Eriks. *Gustava   83847
    12 12298 Gunhild Sørine  Petersd.    Elnæs                       0000  MorD 1869 L E:Hans Christian Eriks. *Laura H   83848
    12 13013 Gunhild         Petersd.    Elnæs                       1827b MorD 1872 L E:Kristian Eriks. *Martinus        83849
    12 07738 Gunild          Petersd.    Østreeie         Asker      0000  MorD 1852 L E:Christian Eriks. *Albertine      83850
   120 48436 Emma G.         Petersd.    Østre(@Vettre)   Asker      1864b Folk 1865 L F:Peter Lars. M:Olava Ellefsd.     83851
XX  10 01616 Gunild          Pedersd.(#) Evju                        1820  Barn 1820 L F:Peder Nils. M:Dorte Nilsd.       83852
XX 110 38445 Gunnild         Peersd.     Evjou            Vestre Bær1820b Folk 1825 L F:Peer Nils. M:Dorthe Nilsd.        83853
XX 120 42299 Gunnil          Pedersd.    Evje             Vestre Bær1825b Folk 1835 L F:Peder Niels. M:Dorthe Nielsd.     83854
    10 01106 Gunild Marie    Pedersd.(#) Fossum Jern Værk            1814  Barn 1814 L F:Peder Lars. M:Kari Halvorsd.     83855
```

```
First: 83726   Last: 83975   Max:125805
```

```
Ind. event ### of ###
Message ### of ###

Window for messages.
```

```
Source-entry window
SOURCE-ENTRY-RECORD No.  2811     BAPTISMS 1814-1878
 Kildepost:1 nr.033 Dato:16.03.1828
 Barnet:Gunild Sørine $Petersd.$ K:K Ekte:E f:03.02.1828
 Far:Peter Lars. B:Elnæs:Asker
 Mor:Kari Christensd. B:Elnæs:Asker
 KM:e: døbt i a.k.

SOURCE-ENTRY-RECORD No.  44875    CENSUS 29.11.1835
 Kildepost:44174 Sted:@Elnæs@Haakavig Husst:767 Kj.start714 Kj.stopp:S
 Navn:Gunnild Pettersd. K Alder:7
 Personer12044875 er D til person12044869 og D til person12044870

SOURCE-ENTRY-RECORD No.  17392    MARRIAGES 1814-1878
 Kildepost:15473 nr.1851031 Dato:26.09.1851
 Brudg.:Hans Christian Eriks. B:Østre f: på/i Høen Ald:22 Siv:ungkarl
 Brud:Gunnild Sørine Pettersd. B:Elnæs f: på/i Elnæs Ald:23 Siv:pige
 Brudg.s far:Erik Jørgens.
 Brudens far:Petter Lars.
 Forl: Forl2:
 KM:viet i a.k.
```

```
Information window

Program mode: Selection mode

Source-entry mode: tidy
```