Reconstituting the Population of a Small European Town Using Probabilistic Record Linking: A Case Study

Sue Dintelman, Tim Maness
Pleiades Software Development, Inc.
sue@pleiades-software.com, timpsd@gmail.com

Abstract

Reconstituting populations is valuable for many reasons. Historians and demographers reconstitute populations to do longitudinal population studies of such things as migration and birth rates. Geneticists use populations to study inheritance mechanisms and to help locate specific genes. Genealogists are interested in tracing ancestry back through time and finding context for their ancestors. Reconstituting populations from vital statistics records such as births, deaths and marriages and from census records is attractive, but until recently, projects of any size required a large investment. Today with available software and low cost hardware smaller organizations and even individuals can undertake population reconstitution projects. This paper outlines the specific steps used to merge birth and marriage records from a small European town into a genealogy.

Introduction

Reconstituting populations is done for a variety of reasons. Historians and demographers reconstitute populations to do longitudinal population studies of such things as migration and birth rates. Geneticists use populations to study inheritance mechanisms and to help locate specific genes. Genealogists are interested in tracing ancestry back through time and finding context for their ancestors. Population reconstitutions until recently have been of two types: large projects sponsored by commercial, academic or religious organizations or smaller projects done manually by individuals or a small group of individuals. Some examples of larger projects are the Church of Jesus Christ of Latter Day Saints Family Search system (Family Search) and Ancestry.com's One World Tree (Ancestry.com). Other projects which have a medical or social science impetus are the Utah Population Database (Skolnick, 1979) (UPDB) which reconstituted the population of Utah starting with the Mormon pioneers up through the 1920s; the BALSAC project in Quebec which linked 660,000 baptisms, marriages, and burials from the Saguenay region between 1842 and 1971 into a genealogy (BALSAC); the Norwegian Historical Population Database (Thorvaldsen, 2009) and the Demographic Database at the University of Umeå, Sweden (Wisselgren, 2009) which uses parish records. Most of these projects use some form of automated record linking and most have significant staffs and budgets.

There are many smaller projects which concentrate on single locations or family lines. See David Hawgood's *One-Place Genealogy* (Hawgood, 2001) for a list of hundreds of projects which focus on a single geographical location. Many of these projects are only collections of records and any attempts to reconstruct the population in a genealogy use manual methods.

With the increasing power and decreasing cost of computers and the availability of software to aid in the process it is now possible to accomplish a population reconstitution project for a relatively low cost. The LDS Church Family History department has a number of locality based projects underway, some of which have been published in the Community Trees section at www.labs.familysearch.org. The Kings County, Nova Scotia; Millville, New Brunswick; and Sør-Aurdal, Norway projects published there are examples of populations which have been reconstituted using automated record linking from source records such as census records, births, and marriages as well as family history data (Community Trees). These projects are by no means easy and still require a level of technical ability in addition to subject matter knowledge. However, inexpensive hardware and improved software puts this type of project within reach of smaller organizations and even individuals.

This case study describes the steps used to reconstitute the population of a small European town using civil birth and marriage records. The project to date has 42,378 original records with references to 151,698 individuals. Although not a large number of records, even this size project makes manual linking of the data a daunting task. Some type of computer aid is required and probabilistic record linking provides a way to quickly produce a good result. The run times for this project are insignificant, taking a total of less than an hour using a generic desktop computer and the GenMergeDB software package (GenMergeDB, 2009). The majority of the time spent on the project was data preparation and the evaluation of the results.

Background

The town is located in a mountainous region in Western Europe and until fairly recent times was isolated. As the largest of the towns in the surrounding area it was a center for merchants and craftsmen during the middle ages. In the late 1800's the political and social environment conspired to make it difficult to make a living and there was considerable emigration to the United States, Canada and South America. The town today has an initiative to conserve their extensive collection of historical records and to make these available to individuals working on their genealogy. Because this is an on-going project the project manager has asked that the town not be identified more specifically.

Methodology

The remainder of the paper describes the specific steps used to reconstitute this population. Most projects have similar steps although the order of linking steps may vary depending on the type of data available, but the over-all approach is the same.

1. *Data Preparation and Analysis.* The data is first reviewed for errors and values are standardized to improve the chances for records to link. Then cross-tabs are computed

to check the discriminating value of fields to determine which can be used for linking. These frequencies are used to set up the weight tables used for scoring.

2. *Determine Cut-off Score.* The characteristic of probabilistic record linking is the computation of a score to represent the likelihood that two records are for the same individual. For a complete discussion of probabilistic record linking refer to Newcombe(1959), Fellegi, Sunter (1969) and Winkler(1994). A self-score is computed to verify that there is enough information to justify doing the project and to help determine the score used as the cut off between links which are rejected and links that are accepted.

3. *Link Records.* This project was done in three steps:
   Step 1. Link the birth records into sibships.
   Step 2. Link the marriages into sibships and multi-generational families where possible
   Step 3. Link the birth sibships to the marriage sibships to produce the final result.

4. *Validate Results.* The final section of the paper describes various validating steps used to ensure a consistent result.

Data Preparation

The civil birth and marriage certificates were photographed on-site and the images used as the source for a transcription done by local townspeople which produced a set of Excel spreadsheets. Figure 1 summarizes the information about the two data sets used for this project.

	Year Range	Total Records	Individuals	Marriages
Birth Certificates	1811-1900	34,397	104,022	34,677
Marriage Certificates	1809-1900	7,981	47,676	23,876

Figure 1. Project Source Records

The fields of interest for record linking in the birth certificate records are the child's name, birth date and place, the first and last name of the father and the first and maiden name of the mother. The marriage certificates contain the couple's names, their ages and the marriage date and place. The record also contains the names of the bride's and groom's parents. This additional family information on the marriage record provides an exceptional opportunity to more correctly reconstitute families as each marriage record contains information on three couples rather than the usual one.

The data from the birth and marriage records was first imported into a database and field values checked for obvious errors. The corrected files were then processed by a conversion program to produce family data. Converting data into a common lineage linked format allows

comparison of individuals and their family information in a standard way. The result of the conversion is a standard GEDCOM format file with the individuals and marriages contained in the vital records. Each record is sourced using the reference to the original image that was recorded in the transcript.

A birth certificate creates up to three individuals: the child, the father and the mother, and one marriage between the father and mother. The marriage certificate generates up to six individuals and three marriages, the marriage for the couple and one marriage for each set of parents. As shown in Figure 1 above, the birth certificates produce 104,022 individuals and 34,677 marriages and the marriage records produce 47,676 individuals and 23,876 marriages.

Initial Analysis

Prior to beginning any record linking project, cross-tabs of the fields produce information about discriminating power and help identify the fields to use for binning and scoring. The discriminating power of each field in the birth certificates is shown in Figure 2. Also shown in Figure 2 are the total number of distinct values for the field and the number of distinct values that cover 50% of the population.

Field	Distinct Values	Distinct Values (50% of Pop)	Discriminating Power	%Complete
First Name	874/943	13/9	5.41/3.51	99.8/99.9
Last Name	1806	30	6.45	99.88
Middle Name	459/363	4/9	3.28/4.67	11.04/32.1
Birth City	9	1	.0011	33.52
Birth year	90	39	6.46	33.56

Figure 2. Discriminating Power of Birth Certificate Fields

Figure 2 illustrates the difficulty of doing record linking on small populations. Smaller projects tend to be concentrated in a single location which makes place fields poor record linking fields. In this data set, birth year is not a significant linking field because for the most part, few records have an actual birth date. In addition, even names, which traditionally have the most discriminating power, have much less discriminating power in a small population. In this population 30 surnames account for 50% of the records. Compare this with closer to 1,000 names for a population of 58 million British records (Gill, 2001).

For data sets spanning a long range of time it is useful to assign an estimated birth year to each individual if possible. This prevents obvious mistakes, for example, where the parents of a child born in 1820 might be incorrectly associated with the parents of a child born in 1900. The birth year on the birth certificates is used to estimate a birth year for the parents. The ages at marriage of the couple are used to estimate their birth year which will be correct within one year.

The ages of the couple are also used to estimate the birth year of their parents. The birth year is estimated to be 20 years before the birth of the child. Since women can bear children over 20 to 25 years and men for a longer period, this estimate can be off by as much as 20 to 30 years. However, having the estimated birth year does prevent making obviously bad links.

Determine Cut-off Score

The frequencies computed as part of the initial analysis become the weight tables that are used as part of the scoring process. Briefly, the weight associated with each value, w_i, for a field is:

$$w_i = \log_2(p/a_i)$$

where p is the population size and a_i is the frequency of the value i. For example, if Smith occurs 60,000 times in a population of 6.5 million.

$$w_{smith} = \log_2 (6.5 \text{ million}/60{,}000) = 6.76$$

So the weights can be handled as integers they are scaled by 10 so the score for Smith would be recorded as 68.

The score for a comparison between two records is computed to be the sum of the weights for each field the records have in common. Missing values are assigned a score of 0, matches are assigned a positive score and mismatches are assigned a negative score. Most data fields are not required to match exactly to result in a positive score. This allows records to match when there are spelling errors, name variations and differences in the value for event dates. Penalties associated with partial matches are assessed against the value of the weight associated with an exact match of the two values being compared.

Prior to starting a linking project it is useful to determine if there is enough information to be successful. GenMergeDB contains a self-score option that compares each record in a set to itself and produces a graph of the resulting scores. In the case of the birth certificates we have only the parent's names and their estimated birth year to use for linking. Figure 3 below shows the results of computing the score for each father with himself. The score is shown on the x-axis with the number of pseudo-pairs having that score on the y-axis. The family score in this case consists only of the individual score of the husband plus the individual score of the wife. This is because in this step birth certificates are being linked and the assumption is that each birth record is for a unique child, so the only family member available to contribute to the family score for the husband is his wife.

The self-score is computed as if each pair is two different records for same person, but in this case the score is computed for each record with itself, so all values will match exactly. The actual score for a pair involving each record will be, at best, the same, but might be smaller if the records have any variations in names, dates or place, so the self-score is actually the best score for each record in the set.

Given the size of this data set, a score of 173 results in even odds that the records are for the same person, that is, the likelihood that the two records are for the same person is 1:1. This is indicated in the graph as line A. For 1000:1 odds in favor of a correct match, a score of 273 is required. This threshold is indicated in the graph as line B. The score that results in 1000:1 odds in favor of a correct match is the score that is used by default for the family score cut-off by GenMergeDB. From this graph we see that the majority of the self-scores are above the default cut-off so we can continue the project with confidence that the record linking will produce results. However, some matching records will not reach the threshold of confidence resulting in under-linking. Setting default cut-off values is done when the weights for the project are computed, but may be adjusted by the user for any run. After the first linking pass, the usual way to check the validity of the cut-off value is to check a sample of links at and just above the cut-off value. Depending on the observed error rate the cut-off score can be adjusted and the linking pass re-run.

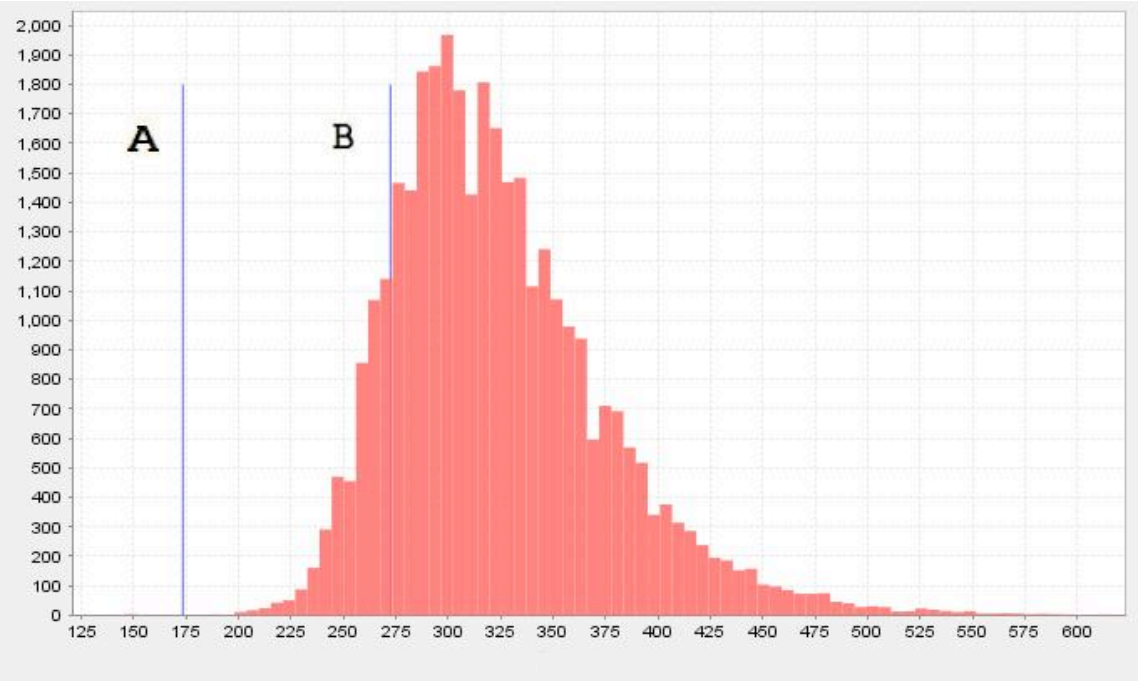

Figure 3. Self-scores for Fathers in Birth Certificates
Line A cut-off value for even odds, Line B 1000:1 odds

Record Linking Step 1. Link Birth Certificates into Families

Figure 4 illustrates the linking that is done in the first step. The goal is to find common couples and associate them with all of their children. In Figure 4 two birth certificates are illustrated that mention the same parents. The result of the linking will be a single record for the father, a single record for the mother and the couple will be associated as parents of the two children. At the end of this step we ideally will have complete sibships for the range of years of the birth certificates.

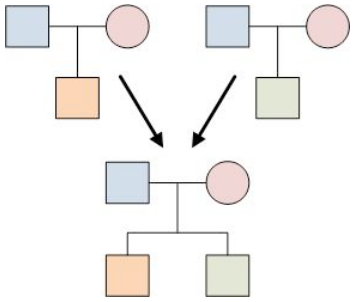

Figure 4. Linking Birth Certificates into Families

For this record linking step and, in fact, for the entire project, we allow no weak links. A weak link is the association of two families based on information for only one person. For example, each child on a birth certificate has the potential to match a father or mother on the birth certificate of their child. Since there is only the information for the one individual involved in making that link, it will not be allowed. Similarly if a man has children by more than one woman, these half-sibships will not be formed at this stage since the only person in common between the two families is the father. Other data sources that have rich information for a single individual (full names, birth and death dates and places) can accumulate scores high enough to allow weak links. This data does not contain enough fields to allow this.

When the data sources being used for a record linking project come from original source documents there should be no duplicates. Of course, this is not true. For census records there is a small percentage of over enumeration and even in this data set there were a few records that were entered twice. The assumption, however, is that each principal event recorded in the record is unique and should not merge with other records of this type. GenMergeDB has a setting on the data source to indicate it is a census, birth, marriage or death record. When this attribute is set for a record, the system prevents this record from merging with other records of that type. Setting the “birth” type for the individual records with the birth dates prevents any merging between individual birth records, even at later stages of the project as the “birth” attribute follows the record through the merging process.

This initial linking run took about 10 minutes to run and produced 13,465 clusters involving 51,976 people. The final merged dataset has 64,877 people and 15,126 marriages, so 38% percent of the original 104,022 individuals created from the birth records merge at this step. 6,644 sibships of size 2 or greater were formed.

Record Linking Step 2. Link Marriage Certificates into Pedigrees

The marriage certificates provide the opportunity to do cross generational links because the marriages mention parents and if we can match this mention to the marriage certificate for the parents this will mention their parents. Figure 5 shows an example of four marriage records that link into a four generation pedigree.

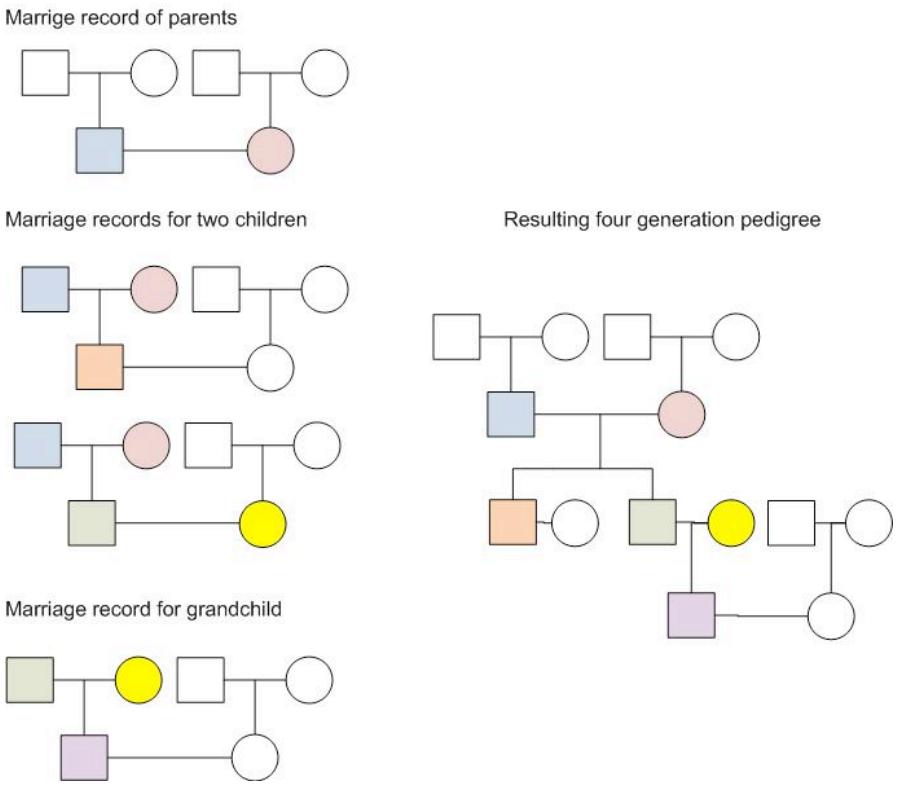

Figure 5. Linking Marriage Records into Pedigrees

The first marriage certificate is for a couple who then appear as parents on the marriage records of their children. The fourth marriage record is for the marriage of a grandchild. Because we have couples names at every stage, it is possible to combine these marriage records into a four generation pedigree.

The cut off scores for this record linking run were not significantly different than for the birth certificates as we would expect since the underlying population is the same. Again, the option to not allow weak links was selected and the marriage records which represent the marriage event (rather than the parents' marriages) were marked to indicate they are original source event records. No other special options were selected for the linking run. This record linking step resulted in 9,090 clusters with 29,028 records. The resulting merged data set contains 27,430 people and 13,831 marriages.

Record Linking Step 3. Link Birth Certificates to Marriage Certificates

The remaining step is to link the two linked data sets together. Before linking the two data sets together, the estimated birth year is recomputed for the parents. In each data set there are now sibships and where the original mother for six children may have had records with an estimated birth year ranging from 1820 to 1835 if her first child was born in 1840 and her last in 1855, the resulting merged record more correctly has 1820 as the estimated birth year. This run produced

23,182 clusters with 48,169 individuals. The final merged dataset contains 67,320 individuals and 21,180 marriages.

Validating Results

One of the most difficult aspects of doing automated record linking on even the relatively small number of records in this project is validating the result. There are two types of errors in record linking: under-linking or missing valid links and over-linking where a link is made between records that in fact do not belong to the same person. The goal is to minimize both types of errors, but there will be errors. It has been shown that even the most conscientious manual linking is also error prone, so this is not a problem unique to computerized record linking (Wisselgren, 2009).

When there is a substantial test set that has been verified by some alternate technique it is common to compute the *precision* (are the links made correct) and *recall* (are the correct links found) of a linking method and to find a cut-off that works for the particular data set and application. In practice, test sets rarely exist and validating the results involves a combination of manual checking and statistical analysis of the final result.

For probabilistic record linking a valuable exercise is manually checking links which score close to the cut-off value. If too many of these links produce inconsistent families, then the run can be repeated with a higher cut-off value. In addition, manually searching for records that did not link can also provide insight into the success of the linking. Both of these manual methods were used at the end of each step of this project.

In addition, after each run the following items were checked:

- Number of marriages per individual
- Number of children per marriage
- Number of sibships with a range from youngest to oldest being greater than 25 years
- Number of sibships with children having the same name
- Number of individuals with children by more than one spouse

Usually, when there is a problem with the cut-off score or another setting used by the record linking, the problems are not subtle and one or more of these counts will differ greatly from the expected value.

We also checked a small sample of pedigrees against those available from on-line sources and found that the results were the same.

Final Linked Population

The objective is to build pedigrees and although there is a known under-linking because of the small population and common names, there are many multi-generational families constructed. Almost 20% of the individuals born between 1880 and 1900 have four or more generations of ancestors.

The result of the record linking can be exported as a GEDCOM file. Figure 6 illustrates how the sourcing is done. Each original individual record has a source reference. The final merged individual carries all these original sources so the users of the final linked set can refer to the original documents. In the example below we have a woman who was referenced in 10 documents: her birth certificate, her two marriages, the birth certificates for five children, and the marriages for two of her children. For the records where this woman is mentioned only as a name (as the mother in either a birth or marriage certificate) the sources are listed with the name as shown in the screen image on top. The information from the birth certificate is listed in the details of the birth event as shown in the screen image in the background. The marriage record details are listed similarly in the source for the marriage event. Searching this GEDCOM file for a specific individual will allow a researcher to find the references to all the original documents that were used to build this person and to easily move up and down in the resulting pedigree to see the ancestors and descendants of this person.

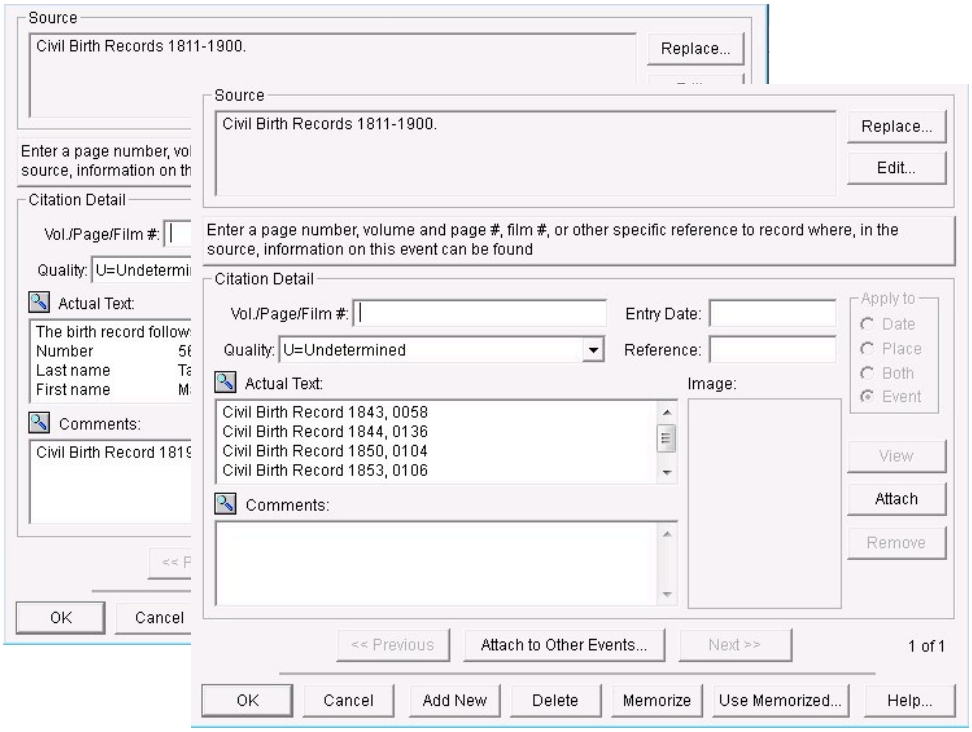

Figure 6. Source Examples

Next Steps

The project is on-going and as additional data sources are digitized they will be added to the project. Some of these data sets include a 1901 Census, various church census records and other church records including baptism and marriage records. Currently the results of this first linking are being reviewed by a professional genealogist familiar with the area.

Conclusion

Although this project is a small one, the same methodology can and has been used for much larger datasets, including building multiple genealogies of 50+ million individuals from lineage linked data. Choosing good binning fields significantly cuts the number of comparisons that must be done for a large project and most steps in the linking process can be done in parallel across multiple machines. Scoring is done on the fields and relatives available, so the algorithms work whether the records being linked are for individuals without any family information, vital records with limited family information, or full lineage linked datasets.

Reconstituting historical populations is valuable for demographers, geneticists and genealogists. Current hardware and software make these projects more accessible to smaller organizations and individuals.

References

BALSAC Project, Université du Québec à Chicoutimi, 555, boulevard de l'Université, Chicoutimi (Québec) G7H 2B1. Website: http://www.mdeie.gouv.qc.ca/index.php?id=4794

Community Trees, FamilySearch Labs, The Church of Jesus Christ of Latter-Day Saints, http://www.labs.familysearch.org/

FamilySearch, The Church of Jesus Christ of Latter-Day Saints, http://www.familysearch.org/eng/default.asp

Fellegi I.P., Sunter A.B. A theory of record linkage. *Journal of the American Statistical Association*, 65 (1969), 1183-1210.

GenMergeDB User Manual (2009), Salt Lake City, UT: Pleiades Software Development, Inc.

Gill, Leicester, Methods for Automatic Record Matching and Linkage and their Use in National Statistics, National Statistics HMSO. John Charlton, Ed., 2001., pp.

Hawgood, David (2001), *One-Place Genealogy* , D. Hawgood, pub.

Newcombe, H.B. , Kennedy, J.M., Axford, S.J., James, A.P. “Automatics linkage of vital records.” *Science* 130 (3381) (1959): 954-59.

One World Tree©**,** Ancestry.com, <http://www.ancestry.com/search/rectype/trees/owt/>

Skolnick M. H. (1977) Prospects for population oncogenetics, in Mulvihill J., Miller R., Fraumeni J.J. (eds): **Genetics of Human Cancer.** New York: Raven Press, 1977

Thorvaldsen, G. (2009) *Record linkage in the late and early 19th centry. The case of the Norwegian Historical Population Database*, University of Tromsø, Paper presented at the Record Link Workshop, University to Guelph, April 6-7, 2009.

Utah Population Database, University of Utah and Huntsman Cancer Institute. <http://www.hci.utah.edu/groups/ppr/>

Winkler, W.E., (1994) Advanced Methods for Record Linage. *Research Report 94/05*. Found at www.census.gov/srd/www/byyear.html.

Wisselgren, M.J., Larsson, M.(2009) *Demographic Data Base*, Umeå University, Sweden, Paper presented at the Record Link Workshop, University of Guelph, April 6-7, 2009.