Persons Linking in Baptism Records*

Jaroslav Rozman[1][0000−0001−8443−433X] and František Zbořil[2]

[1] Brno University of Technology, Brno, Czech Republic
rozmanj@fit.vutbr.cz
[2] Brno University of Technology, Brno, Czech Republic
zborilf@fit.vutbr.cz

Abstract. This paper describes models that can be automatically created from genealogical records (baptisms, marriages and burials). Those models represent various relationships between people mentioned in the records. The most important relationship is child - father or mother, but there can be found others - grandparents, godparents, best men, midwifes or priests. Usual goal in genealogy is to create models that are based only on child - parents relationships. Such models are called family trees. In this paper we describe whole procedure neccessary for creating such models. We are using rewritten baptisms records of small village that covers year range between cca 1607 to 1899. Those records are loaded from simple database and they are transformed to the structure containing all neccessary information about one person. From this first structure we create another one, that contains all persons mentioned in the record and which is suitable for comparison. After comparison the probability that both persons in two records are the same is computed. If the probability is smaller than threshold, the record is added to the output database, if it is bigger, it is merged with the second record. Because the rewritten records were hand-connected to the family tree in genealogical SW and all persons got its own ID, we are then able to find succes rate of our approach.

Keywords: Genealogy · Baptism records · Records linkage.

1 Introduction

Genealogy, the science about family trees, is widely widespread around the world, e.g. in the USA it is one of the most common hobbies. There are a lot of web servers, where people can upload their family trees and share it with other people (Myheritage, Geni, etc.).

Most genealogy web servers allow people to upload their family trees and then searching engines search for possible matches. But those family trees are without any links to the physical records. So we came with a slightly different approach. First, we rewrite the records with references to the particular record on the page of parish book and the searching is done after it.

* This work was supported by TACR No. TL01000130, by BUT project FIT-S-17-4014 and the IT4IXS: IT4Innovations Excellence in Science project (LQ1602).

The original idea of this approach is to help genealogists to avoid multiple repeated searches in the books. Typically most genealogists do not rewrite records, they just go through it, find records that interests them and continue searching in another book. But usually, when they are far enough in the past, the number of searched persons in one book grows and it is not easy to find them all during one search, so it is neccessary to go through the same parish book multiple times. And at that moment it can appear that it would be easier to rewrite the whole book already in the beginning. Apart of it, many genealogists search independently in the same books. So if the rewrited records would be accessible for everybody during the moment of rewritting then people could cooperate. Then it could make things much easier. Advantage of this approach is that one genealogist can rewrite only those records that interests him/her, other genealogists rewrite records that they are interested in and in such a way the whole book can be rewritten. Rewriting of just a few records that a person is interested in is not so demanding comparing to the case when one person has to rewrite the whole book, which can have few hundreds of pages with thousands of records.

The goal of our project is to create database suitable for community rewriting of parish books. But our project has also higher level part. We want to create another database, where not the records, but the persons and relationships between them will be the main part. We want to perform relationship modelling, i. e. to connect identical persons appearing in the first database. People there can have different roles (child, father, mother, etc.) that will be described in later sections. So after such connection we can have not only the basic family tree but we can also know relationships among children, parents and godparents and other. In the general sense, we have some persons and we know in which records (documents) they are appearing and in what roles.

Even though genealogy is widely spread hobby, the record linkage in this area is not commonly used. The reason probably is the small amount of data. The parish books are being rewritten only in few countries, for example the BALSAC project[3] in Quebec is probably one of the first, another example is HisKi[4] in Finland. There are some examples, where small part of the country were rewritten and data were used for record linkage. The work on connecting records in small town is described in [1]. Here for the determining of score of first or last names authors used weights that depend on the frequency of the particular name in the database. For example name "Joseph" is very common, so its weight is much smaller than "Lucas" which is very rare. This means that if they try to determine if two Josephs are the same person the probability will be much smaller than for two Lucases. The record linking were done in three steps: in first step they tried to find common couples and associate them with all of their children, in second step, they link marriage certificates into pedigrees and in the final step they linked both previous datasets together. Because they did

[3] http://balsac.uqac.ca/english
[4] http://hiski.genealogia.fi/hiski/9asgip?en

not have the ground truth, they validated the results by set of test like number of marriages per individual, number of children per marriage and other.

Other example is from the Val Borbera valley in Italy [2]. Here again authors are using both birth and marriage records. That is because birth records do not contain names of children's grandparents and without this information it is more-or-less impossible to create pedigree. Similar as we do they use blocking where they test is both persons in two records have same sex an if date of birth of one person is in the birth interval of second person. The validation of results is done by searching for the birth record of the child's father, then they find father's marriage record, compare names of fathers parents in birth record and marriage record and finally they check if name of wife in marriage record is the same as name of mother in child's birth record.

To our best knowledge the only papers that work with ground truth are [5] and [3]. In [5] the authors used database created by domain experts with birth, marriage, burials and census data from Isle of Skye between 1861 and 1901. Their results are mainly focused on evaluation of record to record than person to person and as matching tool they used algorithm that was originally meant to use for authors matching, so their results is hard to compare to others. In [3] the authors are using their tool described in [4]. They work with database with more than 100 thousands persons obtained from a domain expert. They used Bayes probability to determine if two records are about the same person. Bayes probability here has similar effect as in [1] - the probability is depends on the frequency of the person's name. Because they have ground truth for the dataset they used, they were able to compute accuracy (between 57% and 65%) depending on the linking method they used.

The neural networks are used in [6], [7] or [8]. In [6] the author also used ground truth, unfortunatelly, there is not explained how the matched data looked and how the matching was done.

The rest of this paper is organised as follows: next section is short introduction to parish books, 3rd section is about preparations of records, 4th section about record linking, 5th is about dataset we used and 6th is testing and finally there is a conclusion.

2 Parish Books

The duty of writing church registers was ordered in 1563 by Trident council. It was ordered again for Czech lands in 1591. So the oldest church registers in Czech Republic has been founded around the year 1600. But due to the various wars and other disasters (fires mostly) it is common for most villages to have church registers since about 1650. There are three kinds of church registers (parish books) -– baptisms, marriages and burials. Since only allowed religion at that time was Catholicism, the church registers started as catholic. Later, when also Evangelicism and Judaism were allowed, there were more kinds of church registers, but together with allowing other religions, the uniform printed form was ordered, so we use only the catholic churches registers as example.

As we stated before there were three kinds of church registers. Because these registers were first used only for ecclesiastical purposes, it does not have information about date of birth (or death), but only about baptisms (or burials). Since 1784, when they were declared as public (not ecclesiastical any more) documents, they started to contain also information about date of birth (death). Church registers (or more exactly, the latter one we can call civil registers) contain private information, so they are freely accessible only when the time from last record is more than 100 years in birth registers and more than 75 years in marriage and death registers. It means we can assume they are freely accessible up to about 1900 in births and 1910-1920 in marriages/burials. Those church registers are usually scanned and they are freely accessible via internet. The language, that was used varies, but we can generally say that the oldest church registers were written in the Czech language, then in Latin and since 1784 in German and then again in the Czech language.

The structure of records since 1784 till about 1900 did not change, so we are stating this structure here (see Fig. 1). The structure before 1784 was similar, but there were less information. Usually the info about child's (spouse's) grandparents and the reason of death was missing. Since in this paper we deal with the birth records only, we provide the structure of birth records as follows:

Parish/birth record

- Date of birth and baptism
- Name of priest
- Name of child
- Male/Female
- Il/legitimate
- Name of father, occupation, place of residence, names and place or residence of his parents
- Father's religion
- Name of mother, names and place or residence of her parents
- Mother's religion
- Names, occupations and place of residence of godfathers
- Usually the name of midwife was added
- There were sometimes remarks about date of death, marriage or other

3 Records Preparation

With cooperation with our colleagues from Philosophical Faculty of Masaryk University we have created a template for rewritting of the baptism records. The template has almost 150 items, because we included full adresses and occupations for all possible people (up to 13 persons) appearing in the records and we also added some items for information that are added extra to the record (date of wedding or death in baptism records, etc). We used names of the child and all its mentioned ancestors for linking, we also assume that we use names of godparents,

Fig. 1. Example of baptism/birth record with heading from the church register. On the upper image there is baptism/birth record from May 1899 (17th was the birth, 23rd was the baptism). The child is *Matilda*, father is *Josef Řičánek* and mother is *Matilda*. The father's occupation is *hajný*. For both father and mother there are also names (*Mikuláš Řičánek, Magdalena Sláma*, and *Engelbert Přichystal, Vincencie Polák*), occupations (*dělník, výměnkář*) and villages (*Ubcích, Bukovince*) of their parents. This is what we labeled as 4GP record (4 grandparents known). Additional informations here is date, place and name of groom (on lower left corner) and dates of birth of both parents (16/3 (18)53, 9/3 (18)58). On the lower image there is baptism record from 1.1.1734. The child's name is *Josephus Stephanus* and his father is *Georgius Snaschel* and mother is *Anna*. This record we labeled as F+M (father and mother).

because it can be of a big help in old records where only the first and the last name of father and first name of mother is used. And sometime the father's last name was changed, because at that time the last names were not stable. In such case the names of godparents can help, because people usually used only one or two pairs of godparents for all their children, so we can compare first names of parents and the names of godparents and based on that decide if the children are siblings.

3.1 Structure of One Person

For each baptism record described in previous subsection we create the same structure for every person mentioned in the record. The information about one person is as follows:

- ID - id of the person
- IDrecord - id of the record from parish book, used for not comparing persons from the same record
- IDgedcom - id from genealogical SW where family tree was manually created, used for presicion/recall computing
- role - role of the person in the record (CHILD, FATHER, MOTHER, MOTHERSFATHER, etc.)
- birth date
- birth date range - for parents and grantparents
- baptism date
- death date
- death date range
- weddings date
- weddings date range
- first name - there can be more names
- last name
- multiples
- sex
- religion
- occupation - there can be more occupations for one person
- place of birth (village, street, house No.)
- places of live (village, street, house No.) - for parents, taken from place of birth of their child, there can be more addresses
- identities - array of ID of person's who was determined as identical
- fathers - array of ID of father's of persons from identities
- mothers - array of ID of mother's of persons from identities
- partners - array of ID of husbands/wives (in fact second parent of child)
- children - array of ID of offsprings

Every record for comparing consists of such information about every person that is in the baptism record. During creating of records we fill the ranges based on the date of baptism and some assumptions:

- The birth date range is for ancestors of the child, where we suppose that man can have children between 15 and 65 years (women between 15 and 55).
- Person can have the first wedding in 15 and last at the time of death.
- Person can live up to 100 years.
- If we know, that a person had wedding or child born, we know, that such person had to die after such date (except minus 9 monthes for father).
- If a child is "illegitimate", we know that wedding of the parents has to be after this date, if it is "legitimate" the wedding has to be before this date.

This structure is also used in the second database where the connections are kept. That is why also identities, fathers, mothers, partners and children are part of the record. When the records are created from the baptism record, the IDs of fathers, mothers and children are added and their probabilities/scores are set to 1.0. After the search for the identity is performed, the ID of second person is add to the "identities" together with the probability/score and also other fields are updated (address, etc.).

3.2 Records for Comparing

When comparing records for linking, the obvious way is to compare each record with others. Obviously it is not neccessary to compare children whose birth dates are more than 65/55 years apart because the probability that such old man/woman will have children is very low. There are also compared only persons with same sex. Because we want to find if any person in one record in a parish book is the same as any person in other records, we decided to split every record in parish book to as many records as is the number of persons mentioned in the record. It means that from the upper image in Fig. 1 we got 7 other records (for 1 child, 2 parents, 4 grandparents), from the lower image we got 3 records (1 child, 2 parents).

For every child all its ancestors will be in the new record. Same for father/mother - all his/her ancestors will be in the new record and because name of wife/husband is important information for record linking it will be also added. In case of grandparents there will be only their husband/wife in the new record, in case that their father is also known, he will be also added.

From those informations we create one long record whose items will be compared with others. It means that all records have to have the same structure, because of that we have to add items for husband/wife for the records with child, although it is obvious that newborn baby can not have any partner. All structures are shown in Table 1. Substructures contains particular items for comparing and are same for all persons (Ch, F, M, FF, FM,..., MgF):

$$Ch, F, M, ..., MgF = \left[firstname, lastname, religion, occupation, address\right]$$

Child	Ch	F	FF	FM	FgF	M	MF	MM	MgF	0	0	0	0	0	0	0	0
Father	F	FF	0	0	0	FM	FgF	0	0	M	MF	0	0	0	MM	MgF	0	0
Mother	M	MF	0	0	0	MM	MgF	0	0	F	FF	0	0	0	FM	FgF	0	0
F. father	FF	0	0	0	0	0	0	0	0	FM	FgF	0	0	0	0	0	0	0
F. mother	FM	FgF	0	0	0	0	0	0	0	FF	0	0	0	0	0	0	0	0
F. gr.father	FgF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M. father	MF	0	0	0	0	0	0	0	0	MM	MgF	0	0	0	0	0	0	0
M. mother	MM	MgF	0	0	0	0	0	0	0	MF	0	0	0	0	0	0	0	0
M. gr. father	MgF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1. Structures of all persons created from one record for comparing. The zero means this part is empty (for example there is usually not information about father's grandfather from his father side in the baptism record, there is only info about grandfather from his mother's side - FgF). The five columns with all zeros are not neccessary, but they are there for records consistency.

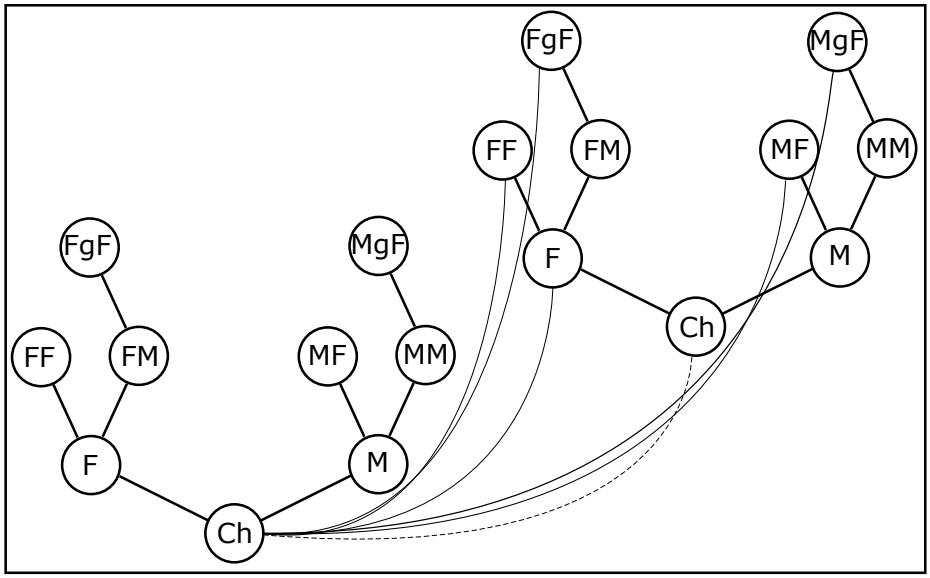

Fig. 2. Scheme of comparing child (boy - is compared only to males) in one record with every possible person (male) in another record. There is also child - child comparison, because there can be two records with the same child (for example when two parish books have time overlap).

4 Records Linking

The records, structured as described in previous sections are compared everyone to everyone. This would means lots of computing, so we limited comparing only for people of the same sex and we also check range of birth date and compare only persons, that are inside this range. Resulting score is between 0 and 1 and the threshold for marking as *identity* is set to 0.85.

Also we applied weights for marking more significant items like names or address where weights are set to 1.0, occupation can change during time, so we set the weights to 0.5 and religion is usually catholic, so we set it to 0.1.

There are three various variable types in the comparison record. There are dates for births, strings for last names and lists. In the lists there are date ranges, occupations and baptism names. Dates are compared as a set of integers, for string comparison we are currently using Levenshtein distance and the resulting real-type number is used for identity score. In date comparison the result is either 0 or 1. Because one person can have more occupations it is neccessary to compare every occupation to every occupation. Similar with given names - a person can be given more names at the baptism, but later only one is usually used, therefore it is neccessary to compare all the given names.

In the first pass over data we only add matching IDs to the *identities* field (see information about person in subsection 3.1) of the examined record. In the second pass over data we check all identities in the examined record (and recursively in the records from identity) so we are able to copy information to the examined (first occurence) record. The information means copying info about children, husbands/wifes, occupations and adresses, we also sort all children according to the birth date and then change possible date ranges of births, marriages and deaths for its ancestors.

5 Datasets

Testing was done on the dataset created from birth records for one village between years 1607 (1635 respectively) and 1899. There are 1961 records that were manually connected to the family trees in genealogical software. Then it was exported to the .csv file together with the IDs (here we called it IDgedcom) for every person. Those IDs allow us checking if the matching was correct or not. Here we have to state, that such approach has "disadvantage", because our data are somehow "ideal". That is because we lost small differences that certainly were in the original records - e.g. same person can be once written as "Jan" and once as "Johann" or also the writing of surnames or occupations can be slightly different. But we suppose that we can allow such simplification because we suppose that in our final system the words in the database will be normalized anyway (probably except of the last names, that will be normalized only partially).

From this data we created two kinds of datasets - in the first set we tried to imitate original records so we erased ancestors that were not mentioned in the

original records (because when we created the dataset, we have 4 grandparents in almost every record, which does not correspond with the reality). We got 4 datasets that we marked as 4GP (four grandparents), MF (mother's father), MLN (mother's last name) and F+M (father and mother). In the Table 2 we can see how much information is available.

In the second kind of datasets we chose only those records that contain all 4 grandparents. From original 1961 records we got 1097 records. In these datasets we again deleted ancestors according to the Table 2, but this time we did it for all 1097 records.

	Father	Mother	Father's fath.	Fathers's moth.	Mother's fath.	Mother's moth.
4GP	1+1	1+1	1+1	1+1	1+1	1+1
MF	1+1	1+1	0+0	0+0	1+1	0+0
MLN	1+1	1+1	0+0	0+0	0+0	0+0
F+M	1+1	1+0	0+0	0+0	0+0	0+0

Table 2. Table shows what information is available in various records for 6 ancestors. First is first name, second is last name. 1 means it is known, 0 means it is unknown.

6 Testing and Discussion

First we describe results from the second datasets, because they are better for evaluation of the algorithm, while first datasets more correspond to the real baptism records, but does not give so good idea about working of the algorithm.

From the Table 3 we can see that number of records, comparisons and time is decreasing as the number of persons in records is decreasing. We can see that for MLN and F+M the number of records is the same, because the number of persons in the records does not change - in MLN we know the last name of mother.

Recall is ratio of *true positive*/(*true positive* + *false negative*). This value is approximatelly the same for all four cases, about 96%. This means we are quite successful in finding true matches. What is changing significantly is precision. This value corresponds to what amount of pairs marked as matches are really true matches. This value is about 60%, which means that among matches is about 40% of false positives. This value is quite hing and moreover the precision decreases very significantly for F+M. This means that almost 90% of matches marked as positive matches were in fact false positives. This is caused by lack of information in the records. For example, we know that somebody's name is Maria and only other information we have is time range of her birthdate, which can be 150 years. This means that we connect all Marias whose birth ranges intersects and because Maria was very common first name, we get lot of false positive matches. If we examine what caused false matches in the F+M dataset, we found that from 11 735 false matches, 11 691 were caused by wrongly matching somebody to CHILD. FATHER is wrongly matched only in 998 cases, but MOTHER in 10 781 cases. In MLN dataset, where we know mother's last name, the ratio of false FATHER:MOTHER matches is only 998:643, so the

number of false mothers matches is even smaller then for fathers. From the Table 3 we can see that best results are for MF, so the obvious solution would be to delete father's parents and mother's mother from 4GP records. Unfortunatelly, from Table 4 we can see, that the time where we have all four grandpatents is quite small (in our case 1858 - 1899, for mother's father or mother's last name 1791 - 1857). Luckily, approximately 60% of all people born between 1600 and 1900 were born after 1800.

The resulting presicion mainly in the F+M dataset is not too high, but we have to keep in mind that those results are based only on baptism records and even human genealogists are not able to connect persons among generations when they do not know their last names. This can be solved only by adding other informations like from burial records or better from marriage records where we can usually get brides last names.

	Recall [%]	Precision [%]	No. of records	No. of comparisons	Time [s]
4GP	94.0	56.9	7679	6 251 838	97.0
MF	98.2	67.1	4388	3 066 204	33.2
MLN	96.9	55.3	3291	2 159 885	18.0
F+M	97.3	12.5	3291	2 075 102	18.2

Table 3. This results are from the second kind of dataset. The time range here is the same for all four datasets (1735 - 1899) and the number of original records is also same for all four datasets (1097). Number of comparisons differs, because records in different datasets contains different number of persons (see Table 2).

	Year Range	Recall [%]	Precision [%]	No. of records	No. of comparisons	Time [s]
All	1607 - 1899	96.0	42.9	12102	15 747 936	180.7
4GP	1858 -1899	94.4	74.2	4434	2 970 692	52.0
MF	1812 - 1857	97.6	64.9	2349	1 110 679	13.2
MLN	1791 - 1812	96.3	98.9	694	128 485	1.1
F+M	1607 - 1790	94.2	22.3	1425	400 928	3.0

Table 4. Comparison of datasets with data corresponding to the real parish books. "All" means all data (4GP, MF, MLN, F+M) are together in one file. Number of original records here (in All or together in others) is 1961.

In the first kind of datasets we have again 4GP, MF, MLN and F+M and then all those combined together. This datasets better reflects reality, because more into the past, less information were provided (and also more mistakes and missing records). Results from this datasets can be seen on Table 4. Again we can see that values for recall are quite high, about 95% and for precision they goes down from 74.2% for 4GP to 22.3% for F+M, but again with exception for MLN, where they have 98.9%, which is very high value, that can be caused by small size of this dataset.

7 Conclusion

In this paper we described chain of tasks that is neccessary to perform when we want to load baptism records from one database, find identical persons and store resulting family trees into another database. We have created vector-based comparison and tested it on about 2000 baptism records connected into family tree from one small village. Dataset we have used was not from raw baptism data, but it was exported from genealogical software and turned into two datasets, one that correspond to real records and second mainly for testing. Advantage of this approach is that we have IDs from genealogical SW for every person that allow us to find out if the matching was correct or not and then compute recall and precision.

Our future work will be mainly aimed on conditional probability in the comparisons of names. Now we assume that all the names have the same probability, but this is not true. Other improvement can be using of neural network for the decision if two records are the same or not. In the approaches with classical or conditional probability there are lot of weights, because every item of the record has different significance and there is very difficult to tune them all. This problem could be solved by neural networks where the input would be a vector of values as is described in Section 4 and the output then would be matching probability of the persons. We also want to create same datasets from marriage and burial records and we are working on the rewriting of the complete original records that will be supplied by person IDs from genealogical software.

References

1. Dintelman, S., Maness, T.: Reconstituting the Population of a Small European Town Using Probabilistic Record Linking: A Case Study, Family History Technology Workshop, BYU 2009
2. Milani, G., Masciullo, C., et al.: Computer-based genealogy reconstruction in founder populations. J. of Biomedical Informatics, 44: 997–1003, 2011
3. Malmi, E. , Gionis, A., Solin, A.: Computationally Inferred Genealogical Networks Uncover Long-Term Trends in Assortative Mating, Proceedings of WWW Conference, 883-892, Lyon, 2018
4. Malmi, E., Rasa, M., Gionis, A.: AncestryAI: A Tool for Exploring Computationally Inferred Family Trees, Proceedings of International World Wide Web Conference Comittee, 2017
5. Christen, P.: Application of Advanced Record Linkage Techniques for Complex Population Reconstruction, 2016, ArXiv e-prints (Dec 2016), arXiv:cs.DB/1612.04286
6. Wilson, D. R.: Beyond Probabilistic Record Linkage: Using Neural Networks and Complex Features to Improve Genealogical Record Linkage, Proceedings of International Joint Conference on Neural Networks, USA, 2011
7. Pixton B., Giraud-Carrier, Ch.: Using Structured Neural Networks for Records Linkage, In Proceedings of the Sixth Annual Workshopon Technology for Family History and Genealogical Research.
8. Gottapu, R. D., Dagli, C., Bharami, A.: Entity Resolution Using Convolutional Neural Networks, Procedia Computer Science 95, 2016, doi: 10.1016/j.procs.2016.09.306