Possibilities of Creation of Community Genealogical Database with Semantic Information

Jaroslav Rozman
Brno University of Technology
Božetěchova 2,
612 66, Brno
+420 54114-1190
rozmanj@fit.vutbr.cz

František Zbořil jr.
Brno University of Technology
Božetěchova 2,
612 66, Brno
+420 54114-1173
zborilf@fit.vutbr.cz

Radek Kočí
Brno University of Technology
Božetěchova 2,
612 66, Brno
+420 54114-1171
koci@fit.vutbr.cz

ABSTRACT
This paper deals with design of database system for innovative approach for rewriting records from seriel sources, mainly from old church registers. Suitable database together with user friendly GUI for as comfortable rewriting as possible is needed. Since the records will be rewritten by volunteers – mainly amateur genealogists – we need some reputation system that determines ability of reading of old handwriting. Last part will be focused on getting semantic information from those records. This will be the most important part of the work that allows getting information like average number of children per pair, average age at death and so on.

CCS Concepts
• Information systems→Database management system engines →Database design and models

Keywords
genealogy; serial historical sources; church registers; land records; computer vision; database; uncertainty

1. INTRODUCTION
Currently we are in the middle of the phase of digitizing archival materials in czech archives, including seriel sources (church registers, land records, etc.). These materials are widely used in genealogy and has significant potential for historic-statistic and demographic research. Next logical step following the digitization is data acquiring from those materials. Because the amount of all pages in all church registers is enormous (tens of millions of pages), it is not possible to rewrite it by standard ways.

If we suppose the techniques from computer vision (e.g. nowadays popular deep neural networks) are still too immature for "reading" the records, the only possible way is to use the same approach as in the Wikipedia – let the community of users create such database.

ICSIM2018, January 4–6, 2018, Casablanca, Morocco

ACM ISBN 978-1-4503-5438-7/18/01…$15.00
DOI: https://doi.org/10.1145/3178461.3178465

In this paper we are discussing the possibilities of creating suitable database for such task. Although rewriting of the church registers is not new idea, our approach is quite unique in further work with the data.

Most worldwide known project of rewriting church registers is performed by Mormons [1] – they try to rewrite books from probably the whole world. They use their own application that has to be downloaded to the PC and always two volunteers rewrite one church register. Their work is then checked by somebody from Mormons. Their approach has several disadvantages. First is necessity of downloading the special application that dissuade many people. Second is rewriting the same book by two people and checking by the third. Because of the old handwriting some supervision is necessary, but in this way it is wasting of human sources. Anyway there are still a lot of mistakes in those rewritten records. Last and probably biggest disadvantage is necessity of rewriting whole books.

The idea of rewriting church registers is not new. First attempts were in France in the second half of 20th century. Also e.g. in Iceland there are genealogical databases to ensure not marrying closely relative people. A lot of work was done in project Digitising Scotlands [2]. Kirby et al. for example created tool for automatic coding of historical occupations to standard classification of occupation [3] or similar for causes of death [4]. This normalization is then important, if we want to create some statistics from the data in the database or else we would have same occupation/death written in slightly different way and our statistic would be corrupted. Another important area is family tree matching. It is done in commercial web sites like MyHeritage [5], but for our work we need something like fuzzy matching of family trees as described in work [6], in this work there is also described using of graph database for family trees. Necessity of using some kind of probability or fuzzy based algorithms for matching is because records are not at all complete and there are also mistakes already done by priests.

This paper is organized as follows. Second section describes church registers to get familiar with them. Third section describes our database, this is based on the information that we can find in the church registers. Third section is upgrade of the basic database that adds semantic information. This part creates and stores family trees created from records in the church registers. Last section is conclusion and future work.

2. CHURCH REGISTERS
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 are around 1600. But due to the various wars and other disasters (fire) 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 catholic churches registers as example.

As we stated before there were three kinds of church registers – for baptisms, marriages and burials. 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). Only since 1784, when they were declared as public (not ecclesiastical any more) document, they 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 birth and 1910-1920 in marriage/dead. Those church registers are usually scanned and they are freely accessible via internet. The language, that was used can vary, but we can generally say that the oldest church registers were written in Czech language, then in Latin and since 1784 in German and then again in Czech language.

The structure of records since 1784 till about 1900 did not change, so we are stating this structure here (see Fig. 1, Fig. 2 and Fig. 3). 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.

Example of typical information that can be found in all three kinds of registers:

Parish/birth record

- Date of birth and baptism
- Name of priest
- Name of child
- Boy/girl
- 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
-

Marriage record

- Date of marriage
- Place of groom's residence
- Name of the groom, occupation, place of residence, names and place or residence of his parents
- Religion
- Age
- Name of the bride, place of residence, names and place or residence of her parents
- Religion
- Age
- Names, occupations and place of residence of bestmen
- Name of priest
- Usually some notes are added

Death/burial record

- Date of death and burial
- Name of priest
- Place of residence
- Name, occupation, name and occupation of father in case of child
- Religion
- Age (usually guess)
- Reason of death

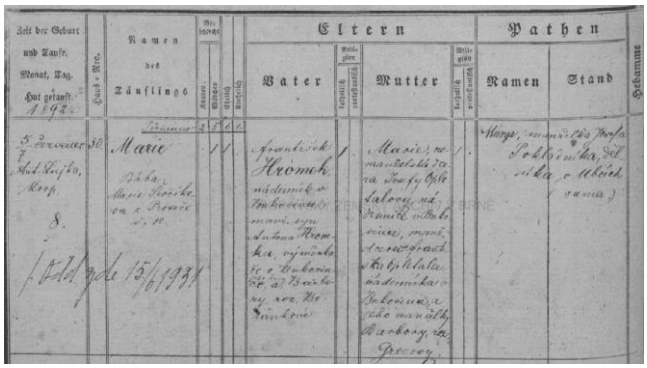

Figure 1 Example of baptism record with heading from the church register.

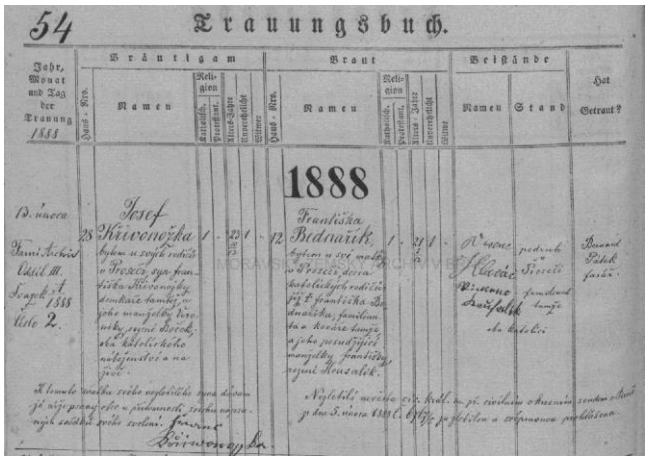

Figure 2 Example of marriage record.

Figure 3 Example of few death records. We can see there is not so much information as in the previous records.

3. DATABASE

To design a database for three kinds of records described in previous section should not be a big problem. But the main idea behind such database is to use it for the genealogical research, so users can show their family trees generated online from records. As far as authors know, such approach is quite unique. Commercial systems (like MyHeritage) allow users to upload their family trees (in GEDCOM format) and when some matches (same persons in two different family trees) are detected, family trees can be merged together. This approach has one big disadvantage. When person that created original family tree finds some mistake and corrects it, it has to send message to all other people, who merged original family tree to correct it. And they have to do it, which is quite rare, so there are a lot of mistakes in the family trees. Because family trees in our case will not be based on persons, but on records, if someone finds record with wrong data field and corrects it, it will be automatically corrected in all family trees that uses that record.

To do so, we want to create something like database over database. The first database will be based on the rewritten records from church registers and second database will be based on persons and will take those records and construct family trees from them. Users will be allowed to edit only data in the first database, family trees in the second database will be created automatically.

3.1 Database Structure

There are eight archives that keep church registers in Czech Republic, two in Moravia and six in Bohemia. We have cooperation with those two in Moravia – we have IDs of church registers, contained villages and we also know which kinds of records the register contains (births, marriages or deaths) and time range for them. There can be various kinds of records for various villages for different time range in one book, so this can be kept in mind in designing the database.

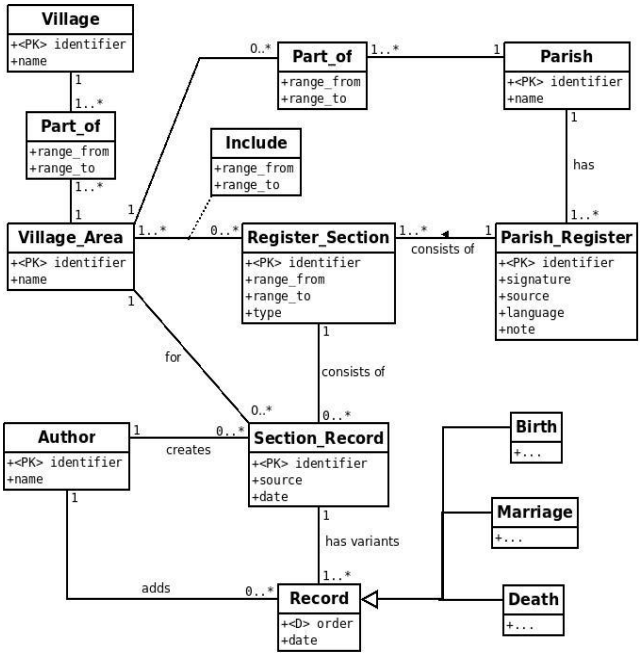

Figure 4 ER diagram of database with particular records from church books (Birth, Marriage, Death).

The description of the significance of the individual entities and their relationships is in Figure 4. Each village (the entity Village) can be broken down into parts (the entity Village_Area), while parts of the village may be assigned over time to different villages. Unless the village is further divided, it has one part with the same name in our model. Each parish (the entity Parish) can then include parts of the villages (Village_Area), which may not always fall under one village. Within one parish, there may be several parish registers (the entity Parish_Register). Each register is broken down into sections (the entity Register_Section). One section records information about either birth, death, or marriages (the attribute type) and links to one or more village areas (Village_Area) that are assigned to the parish under consideration. In addition, the different sections may contain records from different periods for different parts of the village. The individual records in the section are represented by the entity Section_Record. Important information to be stored is the author of the record, i.e., who created the record, and the authors of the individual changes, i.e. who and how they edited the record. For this reason, each record may have several variants (the entity Record) that contain information edited or modified by the author. The entiry Record is generic one for three possible types of records, namely birth (the entity Birth), marriages (the entity Marriage) and deaths (the entity Death) records. These entities contains attributes that corresponds to records described in Section [Church Registers].

This database design allows us to add broad variety of church registers to our database, from cases, where one book contains only baptisms from one village to cases, where in one book there are all three sections (baptisms, marriages, burials) from more villages with different time range for each village.

Structure of the second layer of our database is quite straightforward. The basic element will not be the record, but the person. All persons will have the connections typical for family trees – parents and children. But additional to that, there will be

connections to all other persons mentioned in the record. In the final state it will be possible for single persons find all mentions about it in all historical documents. Also, the date of birth/marriage/death will not be added to the person directly, but as the reference to the record in the first layer. All those connections will be created automatically as described in the next section and users will have limited possibilities to change it. This will ensure, that if somebody finds an error in rewritten record and corrects it, it will be corrected for all family trees that contain this person.

4. MODELS

Main contribution of our system is that it allows to create, display and analyze wide variety of population models. First, we describe main global model that a user can create and then we introduce population sub-models and our approach to their analysis.

4.1 Population models

Semantic of individual records will be created as the data which are written in them will be assigned to an object. Each specific type of record like birth, marriage or death as well as those record written in land records etc. In our system we understand records as something that describes an event. Each record is of a concrete type. It has a template in which are specified roles that persons mentioned in the record has (person can be child, mother, priest, bestman, etc.). In the following points we describe some terms that we use.

Genealogical Entity

Entity is an object of interest about which researcher may seek information in an archive books. Usually an entity is a person, family, municipality, possession etc.

Genealogical Records

Records are paragraphs in an archive book describing an event that occur in specific date at specific place. Such event may but need not change state of one or more genealogical entity, usually person that occurs in the record.

Roles

Each entity, usually a person, mentioned in a record has a role that is specific to the event. One person can be a child, a mother, a grunt buyer, a priest in different records. The roles are specified in record templates and model objects may be associated to one or more roles in one or more records.

Templates

Template specifies a relation among roles in a genealogical record. One template is defined for each record of a given type. In the template it is specified which roles and relations are mandatory and which roles may be omitted. Also it is specified which these roles are mandatory in dependence of time period. This is done on the bases of actual rules valid in given period.

Models

Usually we treat models formally as a set of elements with some relations among these elements. Because we have already built up a database of archive records we may assign elements mentioned in the archival books to some objects, make relations among them and by this to create a model of the population. Such a model contains every object that has been assigned to any role in any record.

Instances

Relations arises when a record is mapped to some objects in a model due to the template which is relevant to the record type. Then a relation appears in the model for the objects depending on the template.

4.2 Population model creation

Creation of so called population model is supported in our system. Every user can create a new empty model that contains no elements and relations. Then he or she adds new object that may be associated with a role in a record.

Typically, church registers or another historical book is selected in archival material selection tree. Then a page and a record on the page may be selected. When a record is selected the user can assign any object that is in the current model. To make this assignment more comfortable user may set some filters that reduces the set of objects which are offered for assignment. The filter may specify location, time, age or gender. We are working on advanced filters that can be specified by querying system. The idea is that we may specify more complex conditions, for example relations to some element (f.e. older than, not married etc.). We will discuss this system in subchapter 4.4. When a set of model objects appears on the screen a user may assign particular roles of the record to the elements. Some model element may be also colored and decorated by the user to make them easier to find.

4.3 Population sub-models

In the analogous way like the objects are filtered during the model creation and role-to-object assignment process user may specify some sub models of the population model. On the contrary to the classical genealogical systems that are focused only to the models of family trees our system allows to create wide scale of models. These models are all sub-models of the population model which a user had created. It is quite easy to specify for example that a model should contain only direct relatives to a specific person, but also relatives with given generation constraint, population of municipality in given period, list of professions, population without branches that have not descendants in given time etc. For both, models and sub-models we offer possibility to create basic statistics (total number of people, average birth rate, oldest persons, average marriage rate, average survival rate etc.) and to view them on a map.

4.4 Querying the genealogical model

One of novel and valuable functionality of our system is a possibility to verify some properties of the population model or its sub-models. Formally we consider a population model as a model of a formal predicate logic theory. Genealogical entities are atoms of a predicate logic language and relations are expressed by predicates. Model is then a structure that is transformed to a PROLOG program in the form of set of predicates. We use SWI PROLOG and SICTUS PROLOG systems for our research, but the final release will contain either SWI or another free licensed PROLOG system.

In such system we use PROLOG programs that verifies axioms that we specified for any genealogical structure. On given model the system checks if there are satisfied some basic properties. For example, whether mothers and fathers age is in an acceptable interval (13-70 for mother, 13 - 100 year for father) and that mother was alive when a child was born and its father was alive or not more than 10 months dead, that pair is man and woman and one is not married more than one times (how it was demanded in previous centuries in Central Europe) etc... We try to collect as

much logically independent axioms for genealogical structures as possible and this is still matter of our research.

On the other hand, our system offers a database of PROLOG rules that allows to answer some predefined queries to the database. We have interviewed some professional genealogists that helped us to collect typical queries that they should be able to answer in family trees that they have made. Based on their experiences we are building a database of predefined questions, for example questions of degree of consanguinity or number of generations between two persons, number of ancestors of given professions, nationality etc.

Our recent research also focuses on a questioning system that would allow to make questions by users for example "which percent of children in 17th century survived in village XY adult age" or "is how many families from 18th century has descendants in 20th century but no paternal line (it means its surname had not survived)". We understand that some users of such system need not be familiar with computer as genealogy is hobby of seniors, so they hardly can make rules in PROLOG neither use most of the current querying systems.

As an example, we show how PROLOG may help to analyze genealogical records. First rule takes all surnames from database that had been transformed to a set of facts / predicates. If we have the fact in simplified form (in our real system each database records consists of thirty items)

Record (book, page, day, month, year, name, surname, fathers_name, mothers_name).

Then definition of such a rule is very simple:

```
surnames(SNM):-
        setof(NM,B^P^D^M^Y^NM^FN^MM
                ^mrecord(B,P,D,M,Y,N,NM,FN,MM),SNM).
```

A query surnames(X) then binds variable X to a list of surnames that are written in a record of the database.

Second definition shows a rule that for each surname in a list detects its first and last occurrence in the database.

```
printlastoccurnames([]).

printlastoccurnames([NM|T]):
        bagof((NM,B^P^D^M^Y^NM^FN^MM
                ^mrecord(B,P,D,M,Y,N,NM,FN,MM),SNM),
        nl,print(NM),
        max_member(LMAX,L),min_member(LMIN,L),
        print(','),print(LMIN),print(','),print(LMAX),nl,
        printlastoccurnames(T).
```

Such PROLOG queries allow for example creation of whole family trees and consequently testing its consistence. Our next research will be focused on adding probability to those queries, so we will be able to tell, with which probability are those parents parents of this particular child.

5. CONCLUSION

In this paper we have presented our ongoing research of possibilities of genealogical database creation and further semantic research over such database. The database that we have created is designed as two layered. First layer is based on records from the church registers. Second layer is based on persons and it is created from the first layer. Also this second layer is necessary for creation of semantic models. Description of theory about semantic models was presented in the last part.

6. ACKNOWLEDGMENTS

This work was supported by the BUT project FIT-S-17-4014 and the IT4IXS: IT4Innovations Excellence in Science project (LQ1602).

7. REFERENCES

[1] Mormon Genealogy Website, https://www.familysearch.org/, [visited 5.11.2017]

[2] Digitising Scotland, https://digitisingscotland.cs.st-andrews.ac.uk, [visited 5.11.2017]

[3] Kirby, GNC, Carson, JK, Dunlop, FRJ, Dibben, CJL, Dearle, A, Williamson, L, Garrett, E & Reid, A, 2015, 'Automatic methods for coding historical occupation descriptions to standard classifications'. In: *Population Reconstruction*, Bloothooft, G, Christen, P, Mandemakers, K & Schraagen, M (Ed.), Springer, pp. 43-60, ISBN 978-3-319-19884-2.

[4] Carson, JK, Kirby, GNC, Dearle, A, Williamson, L, Garrett, E, Reid, A & Dibben, CJL, 2013, 'Exploiting historical registers: Automatic methods for coding c19th and c20th cause of death descriptions to standard classifications'. In *New Techniques and Technologies for Statistics*. Eurostat, pp. 598-607, New Techniques and Technologies for Statistics (NTTS 2013), Brussels, Belgium, 5-7 March.

[5] Free Family Tree, Genealogy and Family History - MyHeritage. url: http://www.myheritage.com [visited 5.11.2017].

[6] Lundberg, H., Fuzzy Matching and Merging of Family Trees using a Graph Database, Master Thesis, Lund University, 2015