# The Karjala database – challenges and solutions for digitizing heterogeneous, old genealogical documents for internet use

Jarmo Saarti, Jari Ropponen, Satu Soivanen

# The Karjala database – challenges and solutions for digitizing heterogeneous, old genealogical documents for internet use

Jarmo Saarti, Library director, PhD, University of Eastern Finland Library, Kuopio, Finland, jarmo.saarti@uef.fi
Jari Ropponen, Project Manager, M.A., South-Eastern Finland University of Applied Sciences, Mikkeli, Finland, jari.ropponen@xamk.fi
Satu Soivanen, RDI Advisor, M.Sc. (Econ.), South-Eastern Finland University of Applied Sciences, Mikkeli, Finland, satu.soivanen@xamk.fi

## Abstract

The Karjala database contains digitized demographic data of the parish registers from the regions ceded to the Soviet Union in 1944. The objectives of the digitization project have been to promote access to digitized records for scientific research and genealogy as well as encouraging research on the people of the ceded Karelia region. The main sources for the database have been catechetical lists, lists of children, and registers of vital statistics (registers of births, marriages, migrations and deaths) that are available in Digital Archives of the National Archives of Finland from the period of 1681 – 1949. The data in the database amounts to about 10.3 million entries, but only data older than 100 years is published openly on the Internet. According to decisions by the Finnish data protection authorities, the Personal Data Act is applied to personal registers less than 100 years old. The digitization process is still going on; it has been calculated that there are 1.2 million entries still to be processed. The database is available to users via https://katiha.mamk.fi/. At present, there are about 6.5 million file entries available on the Internet, each presenting data about one individual, e.g. names, the date of birth and death, the cause of death, age, gender, marital status, occupation, residence, migration, the parish.

The Karjala database can be exploited for diverse research purposes; it improves access to the church records that are sometimes very difficult to read. Information in the database can be utilized for historical research, medical genetics, social sciences, and family and onomastics. The database is can be utilized for clarifying family structures, migratory patterns or child mortality. The database also offers excellent opportunities for interdisciplinary research. Our presentation will describe the digitization process management of old, handwritten documents that consist of non-structured data from a historical period that contains varied linguistic material: several languages from a historical period where nations, states and languages were still evolving, different calendars and spelling rules etc. We will also introduce our plans to use text recognition technology so that the handwritten documents such as the Karjala database will be incorporated into the international READ project network http://read.transkribus.eu/network/. We will also discuss the challenges encountered in this type of heterogeneous data and the possibilities for more defined and structured data management that could enable the automated use of the database. We will also include in our presentation a description of the evolution of the different phases of the database, emphasizing the evolution of the database and its linkage with internet technologies e.g. how they have either hindered or enabled the digitization project.

Keywords: digitization, genealogical documents, handwriting, Karelia, Finland

## Introduction

Our paper will describe the digitization process management of old, handwritten documents that contain non-structured data from a historical period that also contains a variety of linguistic material. During this historical period, several languages were used and it was also an era when nations, states and languages were still evolving, different calendars were in use in different regions and there was no consensus about spelling rules etc. We will also introduce our plans to use text recognition technology to digitize handwritten documents. The Karjala database will cooperate with the international READ project http://read.transkribus.eu/network

There are challenges in handling this type of heterogeneous data but with more defined and structured data management, it should be possible to exploit modern automation techniques to digitize the database. We will also describe the evolution of the different phases of the database e.g. how internet technologies have either hindered or enabled the digitization project.

The Karjala database contains digitized demographic data of the Finnish parish registers from the regions ceded to Soviet Union. This digitization project is intended not only to promote access to digitized records for scientific research and genealogy but also to act as a stimulus to research into the inhabitants of the ceded Karelia region.

The idea of storing the Karelian parish registers into a database was initiated by Raimo Viikki, now the past director of the Provincial Archives of Mikkeli. The pilot project started in 1988 as a research project. It examined how whether the Karelian registers suitable could be converted into a feasible database. The results of the pilot project were very successful and subsequently the Karjala Database Foundation was established in 1990. The database is owned by the Karjala Database Foundation; its members are: the city of Mikkeli, the Finnish Karelian League, the University of Eastern Finland, the University of Jyväskylä, the Finnish Ecumenical Council, the National Archives Mikkeli, the Genealogy Society of Finland and the Population Register Centre.

## What is ceded Karelia?

Throughout history, Karelia (Karjala in Finnish) has been a beneficiary but also a victim of its location between the East and West. When the Russian czar, Peter I (the Great), founded the city of St. Petersburg (a.k.a. Petrograd and Leningrad) in 1703, the security of the city became a major concern for Russia. When Finland became an independent nation and its borders with Russia were defined, the shortest distance from the border to the city, then called Leningrad, was a mere 32 kilometers. When World War II broke out, the Soviet Union feared that Germany would attack it also via Finland in the east. Therefore, the Soviet leaders demanded that part of Karelia should be ceded to Russia, a proposal totally unacceptable to Finland, resulting in the Soviet Union launching an attack on Finland on the 30$^{th}$ of November 1939.

After two wars (1939-1940 and 1941-1944) with the Soviet Union, Finland was forced to cede Karelia officially as part of the Paris Peace Treatment in 1947.

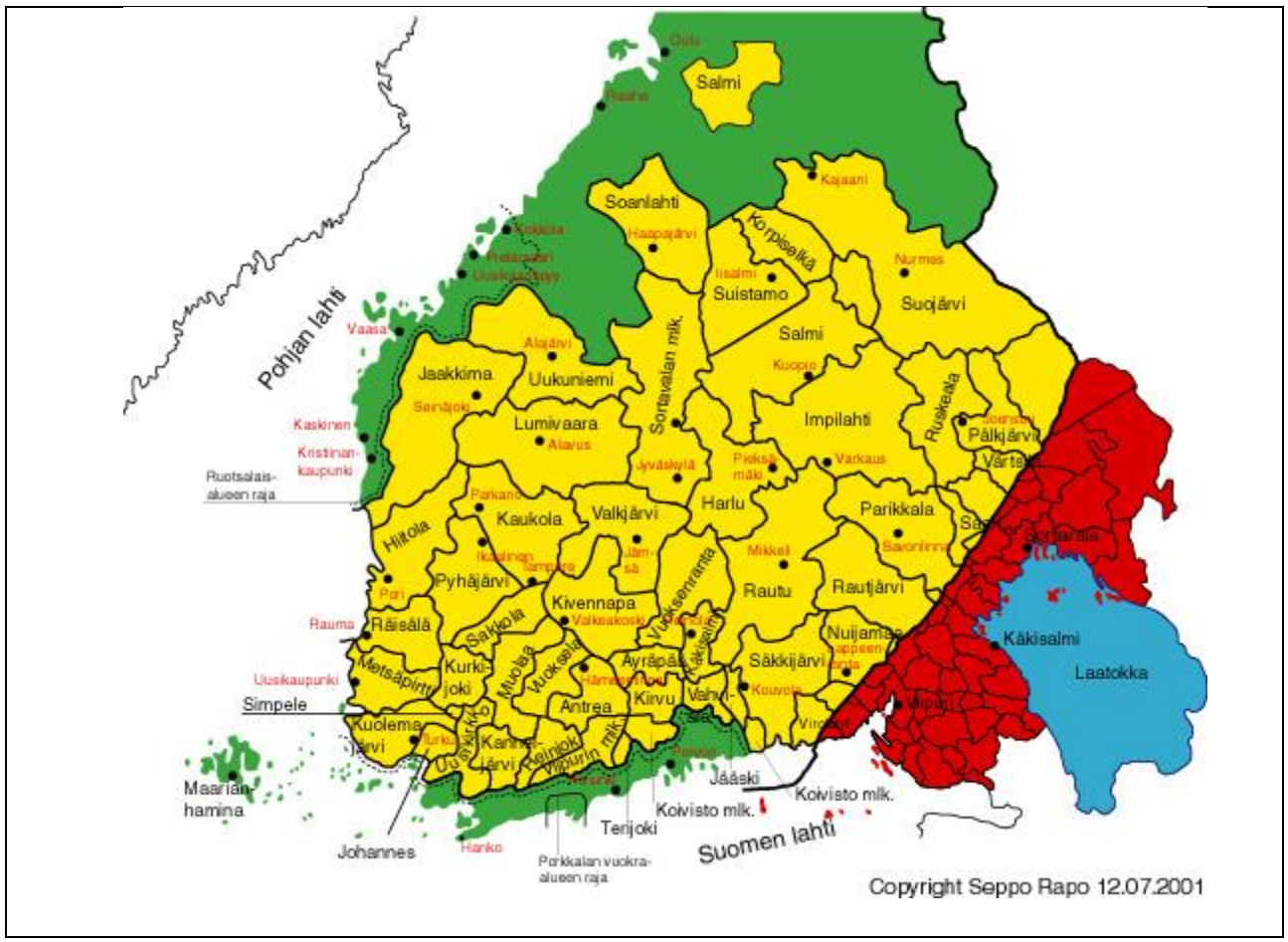



Figure 1. A map of Karelia, https://www.luovutettukarjala.fi/kartat/mhlsijoituss1945.jpg

As the map (fig. 1) shows, the ceded Karelia (red colour) was the South-Eastern region of Finland before the wars. As Finland was still an agrarian country at the beginning of the 20th century, the citizens of ceded Karelia were mostly farmers. There was also an expanding wood processing industry i.e. before the wars Karelia housed about 25 % of Finnish pulp mills, because there were rivers for suitable generating power. A large proportion of the ceded Karelia consisted of water; half of Lake Ladoga (the largest lake in Europe) had belonged to Finland. In 1939, there were about 410 000 people living in that part of Finland (10 % of the total population); they were members of 50 Lutheran and 20 Orthodox parishes. After the wars, the people from the ceded Karelia moved to other parts of Finland (yellow in the map).

For comparison, there have been two major migrations from Finland; one to North America in the years 1881 – 1914 when nearly 300 000 Finns emigrated to the United States and Canada. Another significant wave of emigration from Finland took place in the1960's, this time to Sweden due to the economic opportunities offered there, in that period about 300 000 persons (many of them from eastern Finland) moved to Sweden.

## Background of the documents digitized

In Finland, the Lutheran parishes have kept registers since the 17th century. The instructions for the maintenance of registers came from Swedish church and priests used Swedish for recording the details. This was a normal custom since Swedish was the language of the ruling elite and officers. Nonetheless, in the Karelian region there have been also registers maintained by the Orthodox Church. This was because the Karelian people had lived close to Russia for centuries (fig. 2, The Age of Russia).

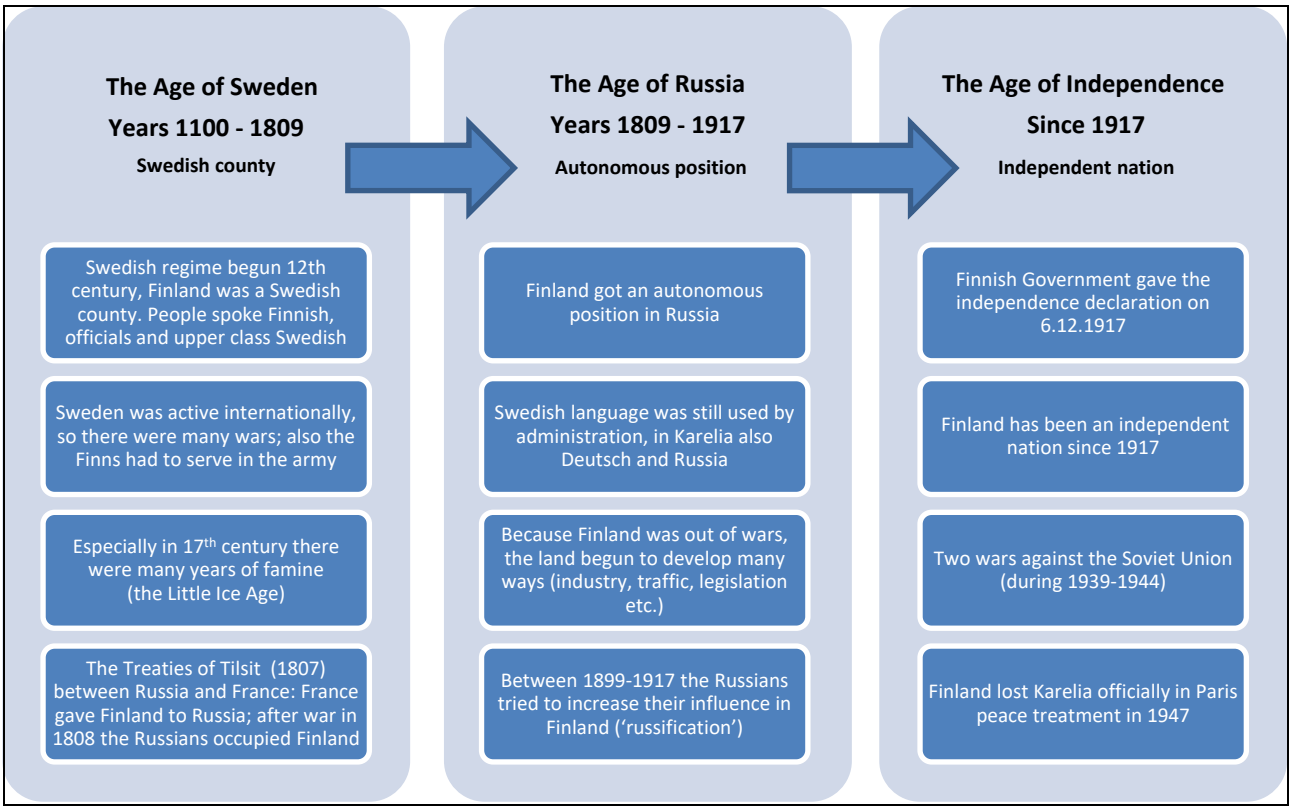



Figure 2. The main historical periods of Finland.

Therefore, both the Lutheran and Orthodox churches both exerted a strong influence on this area. An independent Finland involved two wars against the Soviet Union (the Winter War 1939-1940, the Continuation War 1941-1944). Both wars influenced largely in Karelia and as a result, the Karelian people had to be evacuated from their home districts. This affected also the parish registers since they had to be transported to keep them safe. Most of the registers were eventually housed in the building of the Provincial Archives in Mikkeli, which is a city in South-Eastern of Finland (next page fig. 3, The Age of Independence: The evacuation of the Karelian Church Books).

The oldest Karelian church book dates from the late 17th century with the newest material from 1949. The sources for the Karjala database have been registers of Vital Statistics (registers of births, marriages, migrations and deaths), Catechetical Lists and Children's books, and; these are now available in Digital Archives of the National Archives of Finland from the period of 1681 – 1949.

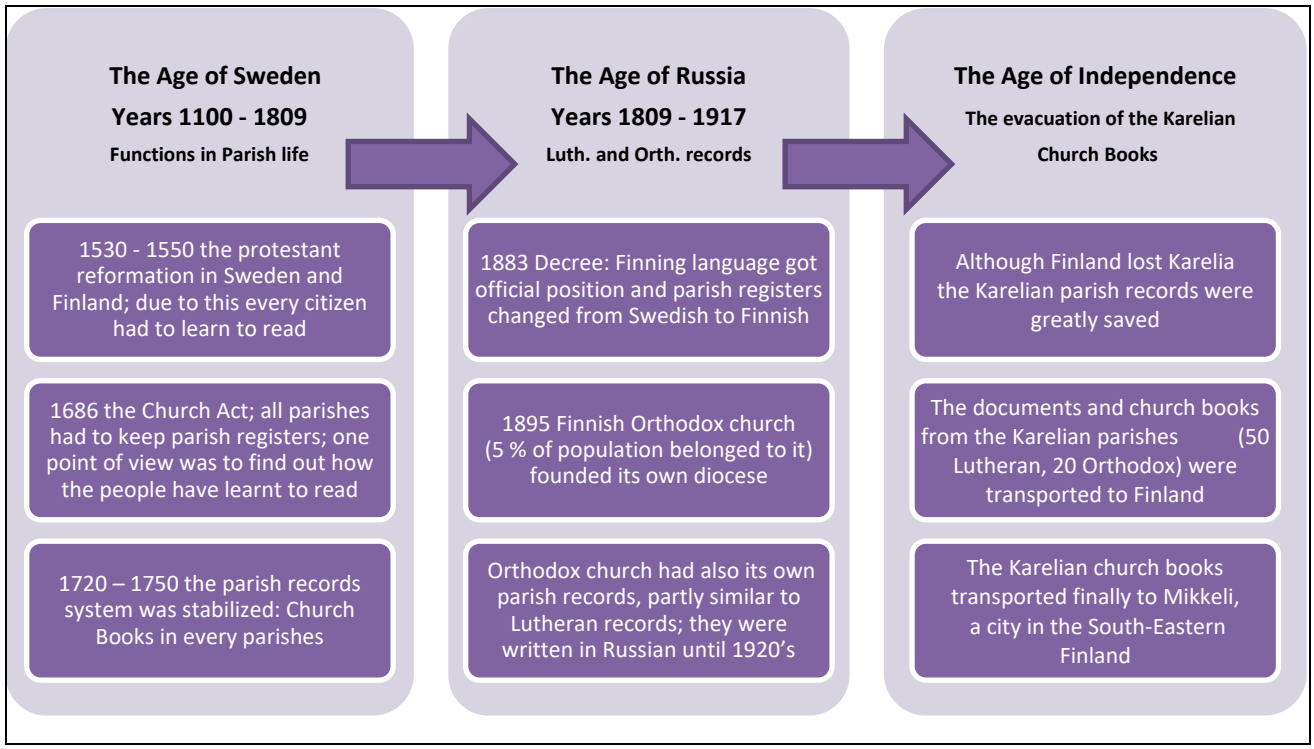



Figure 3. The main periods of the Finnish parish records.

The main bodies of the parish records are:

Birth Records: In the oldest lists of born and baptized, there are only details of the day of baptism, the child´s name, the father´s name and the family´s place of residence at the time of birth.

Marriage Records: The lists of banns and marriages have information: days of banns, the day of the wedding, the names of the bride and bridegroom, their social status and their place of residence.

Death Records: The lists of deaths and burials generally give the date of death, the date of burial, the name of deceased, social status, the place of residence, age and cause of death.

Records of Moving in and out: The lists of migrating persons usually contain the name of the person, the parish from where the person moved and the one to which he/she moved as well as references to a page in the main church book.

Catechetical Lists (so called Confirmation Register): The ecclesiastical code of 1686 decreed that it was mandatory to keep church records in every parish. The primary purpose of the confirmation books was to follow the parishioners´ literacy and Christian knowledge and monitor their participation in the communion. The notes in the confirmation register would often encompass a time span of 10 years.

Children´s lists: There were no guidelines regarding children´s books, so different parishes began to use these at different times. The oldest lists of children date back to the early 18$^{th}$ century. Parish members under 15-years-old were recorded in the children's register and after their confirmation into the congregation; they were transferred to the confirmation register.

## Resources for the database project - workers and funds

Three persons began to store the data into the Karjala database in November 1988. The amount of workers varied between three and seven. But fortunately in theyear 1995 the project got more workers, together 21 persons by the contract with the Finnish public employment and business services. So many long-term unemployed people got meaningful work for themselves (5 hours a day and 25 hours a week) and the Karjala database foundation got needed employees for storing the data. One operator worked normally for one year. The project has had about 600 workers in these 29 years, annual numbers are presented in fig. 4.


Figure 4. The annual numbers of the operators in the Karjala database project.

During 2005 – 2010, the number of workers was over 30. At the same time, the number of stored data entries was about 700 000 annually (fig. 5). At present, the data in the database consists of about 10.3 million entries, but only data older than 100 years is published on the internet. This is because the Finnish data protection authorities have decreed that the Personal Data Act protects personal registers less than 100 years old.

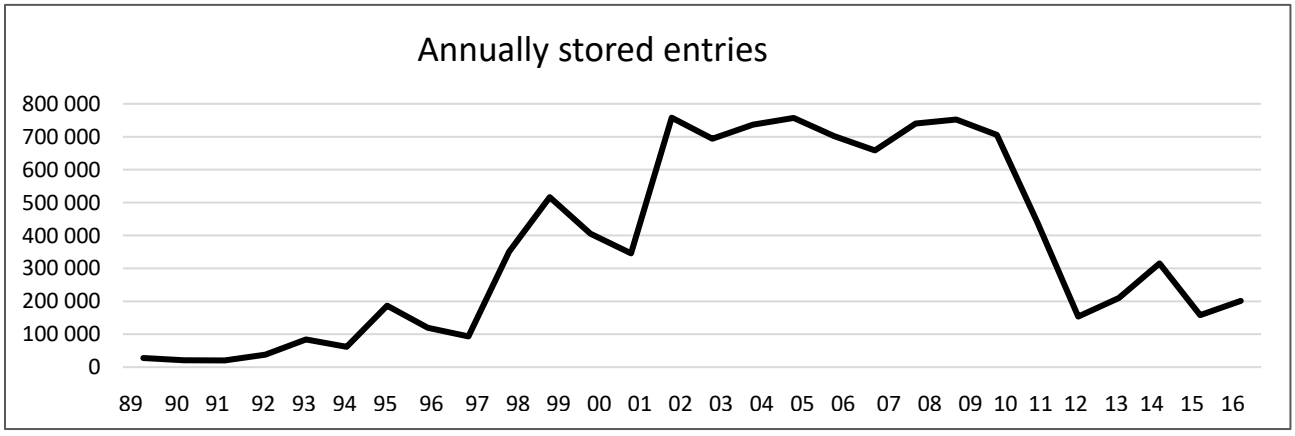



Figure 5. The annual numbers of the stored entries to the Karjala database.

The digitization process is still going on. It has been calculated that there are still 1.2 million entries to be processed. The database is available to users via https://katiha.mamk.fi/. At present, the files available on the internet amount to about 6.5 million entries. At the moment, over 10.3 million entries have been saved in the Karjala Database; the target for the year 2020 is that it should contain about 11.5 million entries.

The Karjala database project has got financial support since 1988, the sum of money will grow up nearly to 9 million euro to the end of 2019. The main sponsor is the government (fig. 6), the support of the ministries of Employment and Education is about 82 % of the funds which the database project has got.

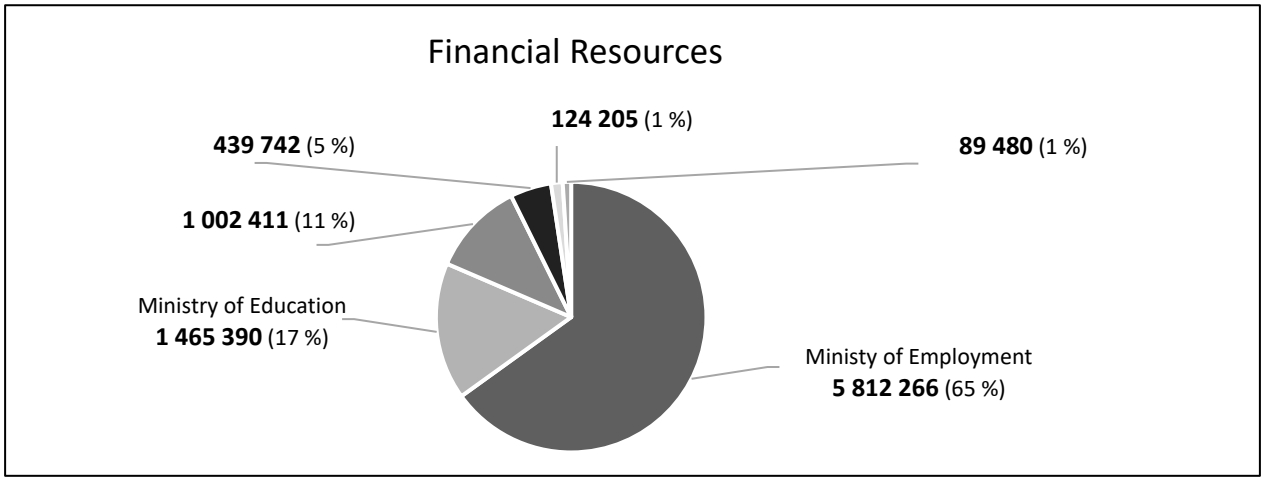



Figure 6. The financial resources of the Karjala database.

## The difficulties of recording parish records

All of the church registers are handwritten with ink and a pen. To help the reading of these old handwritten records, we have begun to examine if it is possible to exploit the international READ (Recognition and Enrichment of Archival Documents) –project.

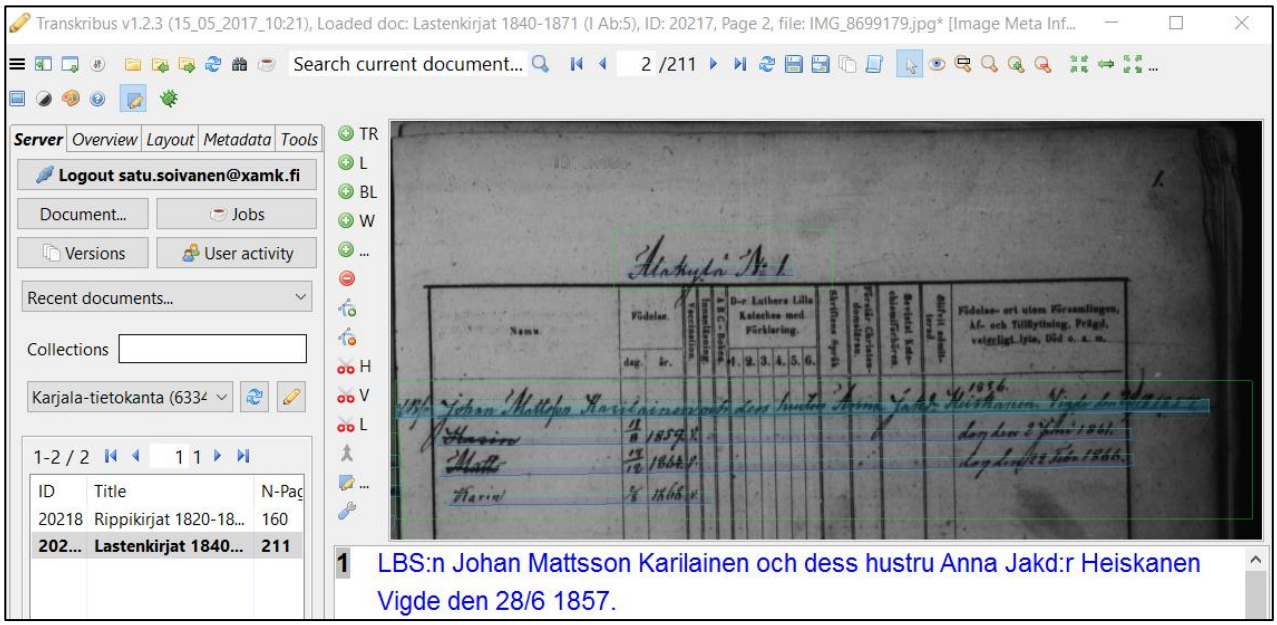



Figure 7. Handwritten text from a church book in the Transkribus-software.

The Transkribus-software of the READ project tries to automatically recognize handwritten text (fig. 7). Unfortunately, the software needs significant ‘teaching’ in order to manage old texts, but we are looking forward to improvements. (See also Blanke, Bryant and Hedges 2012.)

The script varies from time to time and from parish to parish. For this reason, the recording operators have to be talented in interpreting the texts. Since the language is mostly Swedish in Lutheran books and Russian in Orthodox books, the worker must have a good knowledge of these languages as well as Finnish. They also need to know about the history of the language, because many words and even concepts have disappeared.

## Digitization process

The data is stored manually to the Karjala database (fig. 8). The digitization process contains lots of handwork of the operators. At first the data to the database was taken from the original church books. After that the microfilms of the Karelian church books used as the source material to the database. The use of handwritten copies of the parish registers has been one method to make the recording work easier. Nowadays the source material is on internet in the pages of Digital Archives of the National Archives of Finland.

Source data on an original church book | Source data on microfilms

Source data on papers | Source data on the internet

Figure 8. The variety of data sources for the manual data recording.

Today, when digitizing the old and disordered written registers, the staff team consists of experts, many of whom have recorded parish registers for years, almost everyone has graduated with an MA (humanistic sciences) and some speak Russian as their native tongue. In addition, the organization has access to a couple of advisors with special experience who undertake the analyses of the most difficult texts using a variety of sources (fig. 8). The ultimate goal is that the final database should not have many mistakes, e.g. wrong names, erroneous interpretation of an individual's personal details e.g. emigration, crimes, illnesses and the dates of birth, marriage and death.

## Database for research and information use

The user interface to the database called Katiha https://katiha.mamk.fi/User.action?addLocale=en The Karjala database serves many different kinds of research and improves access to the church records that are sometimes very difficult to read for academics without a background in history. Information in the database can be utilized for historical research, medical genetics, social sciences, and family and genealogical research. The entries of the database presenting data from individuals e.g. name, the dates of birth and death, the cause of death, age, gender, marital status, occupation, residence, migration, the parish; so the database is optimal for investigating family structures, migratory movements or child mortality. The database also offers excellent opportunities for interdisciplinary research. Recently there have also been projects that have exploited data mining methods when analyzing the data. (See Räisä and Loponen 2014.)

The researchers could find data e.g. for questions of variations in fertility and mortality, changes in child mortality, causes of death and inherited diseases and the population age distribution in different parishes. The database is a source for statistics of migration for instance from where and to which parish the people have migrated or where did they find their husbands/wives. Researchers of social history can find answers to their interests: How did the industrialization influence professional titles and women´s work situations? Did the knowledge level notations in the confirmation books predict people's futures? How did the social position inherite through the generations in different groups in the community?

The Karjala database is an excellent aid for genealogists to research their family roots. There are data in the database for onomastics to research family names within an area or to research changes in family names within a certain period. Some examples of researches where the Karjala database has been used as a source material:

1. Population History Research

Happonen, Päivi. "Two Realities? The Parish Registers and the Population Registers in Describing the Sortavala City Population Profile from the Beginning of the 19th Century until 1940". Doctoral thesis. University of Eastern Finland. (2009) http://epublications.uef.fi/pub/urn_isbn_978-952-219-316-2/urn_isbn_978-952-219-316-2.pdf

2. Migration

Loehr, John : PROJECT Learning our past
http://human-life-history.science/learning-our-past-effect-forced-migration-karelia-family-life

3. Social History

Pettay J, Lahdenperä M, Rotkirch A, Lummaa V. "Costly reproductive competition between co-resident females in humans". Behavioral Ecology 27 (6):1601-1608. (2016)
http://human-life-history.science/file/136/download?token=RI27KFV2

Malmi, Eric: DEMO AncestryAI (an application which uses machine learning to infer family trees and enables visualizing and searching the inferred trees. It has been developed to support genealogical research. AncestryAI includes data mining. http://ancestryai.cs.hut.fi/

The Karjala database is already widely used; it receives between 800 and 1200 Internet visits every day. The Statistical Reports from the beginning of January to the end of March showed the following numbers: the use of Katiha2 (the former online version of the program Katiha) 91,582 visits and the use of Katiha.mamk (the revised version of the database interface) 11,421 visits, i.e. 103,003 visits in 90 days i.e. about 1,144 visits per day.

## Technical evolution of the database

The planning for the Karjala database system started in August 1988. The plan for pilot project was as follows:

Step 1: Description of source material (church books: what we have and in which form)
Step 2: Determination of need (technical solutions: what we want and how can we achieve this goal)
Step 3: Analysis of information (variables, relations, values, codes, key fields, exceptions, controls)
Step 4: System planning (model of the recording system)
Step 5: Coding the recording system

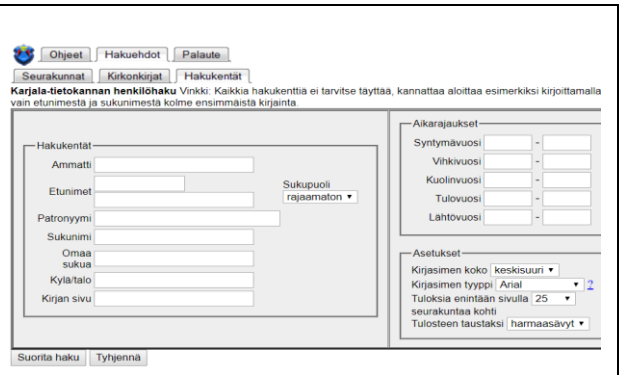
The pilot project had many challenges to resolve some technical problems: how to choose and define the fields which are needed in the recording program and in the database itself and how to express and display the data which has changed with time. The information in different church books varies quite much, there are changes in data (shorter names, synonyms) and lack of data (incomplete birth dates, no definitive names) and language changes (Swedish, Finnish, Russian, German, Latin).

The original database was made with dBase, an old software (management) language. The planned system included six parts with their files: one part for each type of a church book (births 3 files, marriages 3 files, deaths 3 files, migrations 2 files, a book of children 7 files and a book of examinations 10 files). The planned database structure consisted of a total of 28 data files for each of the 70 Karelian parishes. In summary, there were together about 2000 data files in the database.

The original database environment was the dBase -server and mostly housed in the hard drives of computers and CD-ROMs. The dBase was an inexpensive and efficient choice in the 1980's, it suited this kind of research register. It was a spreadsheet type of a database management system but that means there was not any user interface. The backup of the database was arranged by storage on floppy disks. Although convenient at that time, this arrangement was not appropriate for future development. The server required extensive maintenance and simply keeping it running was time consuming. The database had to be divided into two different versions.

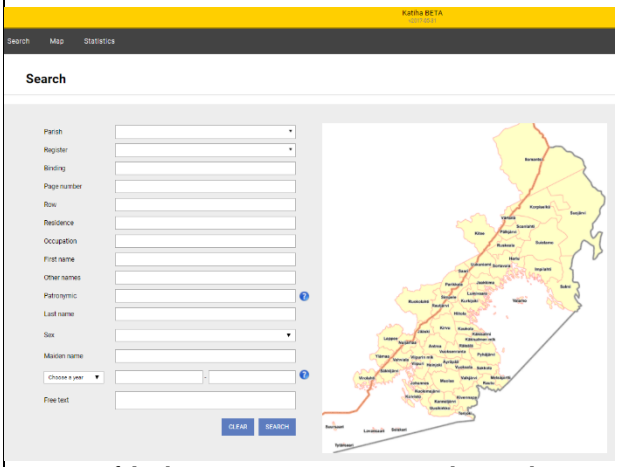
This solution was made when it was decided to allow access to the database on the internet. The Finnish law has defined that personal data can be made public when it is over 100 years old from an individual's birth. However, the data of deceased people is public at 50 years from their death. The Karjala database foundation decided to resolve the privacy issue by dividing the database in two. The public data was accessible on internet and private data housed in the National Archives Mikkeli.


Version of database access in Windows - Katiha




Version of database access on internet - Katiha1


Version of database access on internet – Katiha2




Version of database access on internet – Katiha Mamk

Figure 9. The variety of user interfaces for the Karjala database access.

At present there are four ways to access the data of the Karjala database (fig. 9). This system has too many disadvantages with respect to preserving the data for research access. The objective is to develop the user interface so that there is only one access system and in an open source environment. The modernization will continue in the future in order to ensure an effective utilization so that researchers have easier access to the database.

The original database has already transferred to a new server managed by the South-Eastern Finland University of Applied Sciences. The server has a web service so that the public data – data over 100 years old – is already now widely accessible. The system exploits MariaDB, an open source SQL-environment with the web service running on Java. Hopely these changes will allow the data of parish registers to survive so that it can be accessed by future generations in a format that can not only be read but also modified more easily. After the modernization of the database there are many possibilities to exploit it in other projects and develop an interface for versatile access.

## Conclusions

The need for digitizing old genealogical resources gives rise to new possibilities, especially for digital research. Data mining and data combination methods are completely new ways for expanding knowledge also in humanistic research. It also opens opportunities for other branches of research to utilize this data.

It seems that the older texts still need substantial manual processing and thus are expensive and time-consuming projects. Our case shows that one must also be agile in project management and fundraising. From the point of view of research, it is very important that the original documents are also digitized because one cannot translate all the meanings of the old document especially when trying to make them readable with a machine. On the other hand, for the project leaders this means that their conduct must be critical and ethical: every step of interpretation can change the original meanings of the texts.

The parish registers are only one part of the information originating from ceded Karelia. More information about Karelian citizens is available from the archives of private persons and associations. Some of this information has been digitized and there might be the opportunity to access it. Linking all this data is clearly important but that would need a new project and substantial voluntary work. Some plans have already been made:

1. Integrating geographic information from different areas to the data.
2. Linking the statistical information with the geographic information system.
3. Including different kinds of data into the database, such as photographs, drawings and stories.
4. Connecting the data with the messages in social media.

As the digital realm opens, it offers new possibilities - it is our duty to include human history in this context (see also Parry 2012 and Cimtech 2011).

## Acknowledgements

The authors are grateful to Dr. Ewen MacDonald for linguistic advice and Seppo Rapo for the permission to use the map.

## References

Blanke, Tobias and Bryant, Michael and Hedges, Mark (2012). Open Source Optical Character Recognition for Historical Research. Journal of Documentation (68): 659 - 83.

Cimtech (2011). Crowdsourcing Project Helps Finnish Library to Digitise Historical Documents. Information Management & Technology (IM@T.Online).

http://www.imat.cimtech.co.uk.ezproxy.lib.umb.edu/Pages/IM@T_Online/2011/February-March/IMAT_0211_NewsTrack_11.htm (accessed 22.6.2017).

Parry, Marc (2012). Historians Ask the Public to Help Organize the Past; but is the Crowd Up to it?" The Chronicle of Higher Education 59(2). http://go.galegroup.com.ezproxy.lib.umb.edu/ps/i.do?id=GALE%7CA302048583&v=2.1&u=mlin_b_umass&it=r&p=AONE&sw=w (accessed 22.6.2017).

Räisä, Johanna and Loponen, Mirja (2014). The modernization, migration and archiving of a research register. Proceedings of the DLM Forum - 7th Triennial Conference Making the information governance landscape in Europe. (accessed 22.6.2017).