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

Jarmo Saarti, Library director, PhD, University of Eastern Finland Library, Kuopio, Finland, jarmo.saarti@uef.fi

Satu Soivanen, ADP Planner, M.Sc. (Econ.), The Karjala Database Foundation

Mikkeli, Finland, satu.soivanen@karjalatk.fi

Introduction

# The Karjala database contains digitized demographic data of the parish registers from the regions ceded to Soviet Union in 1944. The objectives of the digitization project have been to promote access of digitized records for scientific research and genealogy as well as promote 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 relates to about 10.2 million entries, but only data older than 100 years is published openly via Internet. According to decisions by the Finnish data protection authorities, the Personal Data Act is applied to personal registers less than 100 years old.

Current status

The digitization process is still going on. It has been calculated that there are 1.3 million entries left. The database is available to users via <https://katiha.mamk.fi/>. At the present the files available on the Internet relates to about 6.5 million entries, each presenting data from one person, e.g. name, date of birth and death, cause of death, age, gender, marital status, occupation, residence, migration, parish.

Database for research

The Karjala database serves many different kinds of research and improves access to the church records that are sometimes very difficult to read. Information in the database can be utilized for history research, medical genetics, social sciences, and family and name stocks research. The database is excellently suited for research in family structures, migratory movements or child mortality. The database also offers excellent opportunities for interdisciplinary research.

Digitization process

Our presentation will describe the digitization process management of old, handwritten documents that contain non-structured data from a historical period that also contains varied linguistic material: several languages from a historical period where nations, states and languages were still evolving, different calendar and spelling rules etc. We will also introduce our plans to use text recognition technology to handwritten documents as the Karjala database will join the international READ project network <http://read.transkribus.eu/network/>.

We will also be discussing about the challenges found 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 where the emphasis is in describing the evolution of the database and Internet technologies and how they have either hindered or enabled the digitization project.