Newly Digitized Database Reveals the Lives and Families of Forced Migrants from Finnish Karelia

JOHN LOEHR, University of Helsinki, Finland
ROBERT LYNCH, University of Turku, Finland
JOHANNA MAPPES, University of Jyväskylä, Finland
TUOMAS SALMI, University of Tampere, Finland
JENNI PETTAY, University of Turku, Finland
VIRPI LUMMAA, University of Turku, Finland

Abstract

Studies on displaced persons often suffer from a lack of data on the long-term effects of forced migration. A register created during 1960s and published as a book series 'Siirtokarjalaisten tie' in 1970 documented the lives of individuals who fled the southern Karelian district of Finland after its first and second occupation by the Soviet Union in 1940 and 1944. To realize the potential value of these data for scientific research, we have recently scanned the register using optical character recognition (OCR) software, and developed proprietary computer code to extract these data. Here we outline the steps involved in the digitization process, and present an overview of the Migration Karelia (MiKARELIA) database now available to researchers. The digitized register contains over 160000 adults and a wide range of data on births, marriages, occupations and movements of these forced migrants, likely to be of interest to researchers across disciplines including demographers, anthropologists, evolutionary biologists, historians, economists and sociologists.

Keywords: births, database, digitization, displaced persons, Finland, forced migration, marriages, mobility, occupations, register information

Introduction

Forced migrations are fixtures of human existence: throughout human history, populations have encountered circumstances in which they have had to move due to environmental, economic, social and political factors. However, studying the consequences of forced migrations are constrained by the availability of individual-based records that detail the life events of migrants prior to and after migration. Between 1940 and 1945 an estimated 40 million Europeans fled their homes in what is widely considered to be the worst refugee crisis in modern history (Lowe, 2012). One interesting opportunity

to study this is provided by the registers that are available on the migrants in Finland who lost their homes during the Second World War.

In Finland, approximately 420,000 citizens were evacuated to southern and central Finland following the loss of Karelia to the Soviet Union. These mass migrations occurred in two distinct waves following Finland's signing of the Peace Treaty of Moscow in 1940 and the Armistice Agreement in 1944, respectively. The initial evacuation occurred in 1939–40 during the Winter War. Although evacuation plans for this region had previously been drawn up by the Finnish government, they were not implemented in time, and the resulting exodus was rather chaotic. Migrants were initially housed in public buildings that were used as shelters and were later transferred to private residences. About half of these forced migrants returned to Karelia in 1941–42 after the territory was recaptured by Finland, although many found that their former homes and workplaces had been destroyed. A second, but more organized evacuation occurred when Karelia was recaptured by the Soviet Union in 1944. Although heavy casualties were avoided, most of these migrants left with few possessions.

The years after the war were marked by the gradual integration of the migrants into Finnish society. This process was aided by a shared culture and language between migrants and their hosts as well as considerable support from the Finnish government to help with resettlement. The migrants were reimbursed for a proportion of their lost possessions, and laws were passed to compel many resident Finnish farmers to sell some of their land to the migrants. By 1951, 100,000 new farms had already been established (Waris et al. 1952), although these new farms were, on average, both smaller and on poorer agricultural land than those that the migrants had left behind in Karelia (Jyrkilä 1975).

Overall the resettlement has been considered a success, especially when one considers the sheer size of the task (e.g. forced migrants comprised approximately 11% of the total Finnish population of 3.7 million), and the fact that the post-war Finnish economy was quite weak. One study even found that forced migration aided certain sections of society. Sarvimäki et al. (2010), showed that male evacuees from rural areas earned greater incomes than rural resident males, possibly because the transition from rural to urban economies was quicker for the displaced males.

Nevertheless, Karelian evacuees often faced serious prejudices, and many resorted to hiding their Karelian accents and identities to avoid encountering negative reactions (Raninen-Siiskonen 1999). In a survey, of 1150 people chosen to be representative of the resident Finnish population in 1950, it was found that 40% preferred to marry local residents instead of Karelian evacuees with the most acceptance found in residents close to the ceded territory and the least acceptance in western Finland (Waris et al. 1952).

A long term effect of forced migration on mortality has also been observed. Analyzing mortality rates of forced migrants in comparison to the resident population between 1971 and 2010, Haukka et al. (2017) found that Finnish forced migrants had somewhat higher rates of overall mortality and especially elevated risks of dying from

heart disease. The authors suggest that one plausible explanation for these differences is higher stress levels encountered by forced migrants upon evacuation.

Researchers across a wide range of disciplines have long been interested in the consequences of forced migration. Fields ranging from demography and economics to political science and epidemiology have sought to understand the economic, social, psychological, political and health outcomes that result from mass migrations. Because these effects are felt both by individuals and by society as a whole, understanding them often requires examining their impact on multiple levels. Not only are we interested in understanding the long term effects of forced migration on the immigrant and host populations, we are also concerned with how these impacts are felt by individuals and by groups. Previous research on the Karelian migrants has taken advantage of census data and focused on one particular aspect of forced migrations such as the economic integration of refugees into the Finnish labor market (Sarvimäki et al., 2010), the cultural impacts on descendants of the migrants (Alasuutari and Alasuutari, 2009) or the impacts on specific types of mortality (Haukka et. al., 2017). By digitizing the register outlined in this article, we provide a way to use a highly detailed, extensive and previously unanalyzed dataset to gain broad insights into the effects of one of the best documented forced migration events on individuals, their families and the larger population. Thus, these new data complement the census data used in previous work (Sarvimäki et al., 2010; Haukka et. al., 2017), and give avenues for new insight not previously possible. This research lies at the intersection of academic research and politics and, therefore, also has implications for public policy.

'Siirtokarjalaisten tie' register of displaced citizens

The experiences of evacuees were systematically recorded in a register compiled in the book series 'Siirtokarjalaisten tie' (Anon. 1970; title directly translates to: 'Karelian migrants' road'). This project was supported by 'Karjalan Liitto' (Karelian Union), an organization founded in 1940 to facilitate the resettlement of displaced Karelians and to represent them in legal matters dealing with land acquisition and reimbursements. Later this organization became involved in the promotion, preservation and renewal of Karelian culture.

Approximately 300 people were trained to interview Karelians for the register, and interviews took place between 1968 and 1970. Using Finnish state records of displaced citizens, an effort was also made to interview people who were born outside of Karelia but who were living there prior to the occupations. Each entry lists the full name (maiden name if applicable), profession, birth date, birth place, all movements (towns or cities of residence) from birth until the date of the interview, as well as their spouse's names, professions, birth dates, birth places and years of marriage for those who married. Children's names, birth year and birthplaces are also listed (see Figure 1).

These basic demographic data are presented in a standardized format but each entry

███████, ███
maanvilj., synt. 11. 9. -08 Metsäpirtillä. Puol.
███, emäntä, synt. 23. 5. -13 Käkisal-
mella. Lapset: ███ -46, ███ -49.
Molemmat syntyneet Seinäjoella. Asuinp. Karja-
lassa: Metsäpirtti -39. Muut asuinp.: Juva, Kan-
gasniemi, Heinävesi, Seinäjoen mlk. 41—50, Lie-
to 50—. ███ ovat saaneet nykyisen tilansa
maanlunastuslautakunnan kautta ja ovat itse ra-
kentaneet kaikki rakennukset. Tilan pinta-ala on
11 ha. ███ on palvellut korpraalina tal-
visodassa Metsäpirtin RajaK:ssa ja jatkosodassa
PionK:ssa. Hän on toiminut Metsämäen laidun-
yhtymän puheenjohtajana n. 10 vuotta.

Figure 1.
Example of data entry from the register 'Siirtokarjalaisten tie'. Names of individuals have been blacked out for the purpose of publication.

also has a less structured section containing additional personal information on the primary entry and their spouse. For farmers, the size of their property at the time of interview is recorded in hectares, as well as the amount of the property that is dedicated to forestry or agriculture. In addition, many entries list the type of livestock present on the farm. There is also information on military service, medals or merits, data on whether the person was wounded in the war and in cases where they were granted invalid status from the government, the severity of the disability. For women, participation and frequently the role in the paramilitary support organization 'Lotta Svärd' is also mentioned. Civilian merits or awards are listed for service in companies or government employment, as well as any merits awarded through free time pursuits. Approximately one quarter of all entries also include a photograph of the primary person, or their family.

Digitization and data extraction of forced migrant register

The amount of data on the Karelian migrants is immense in its original book format, however, they are not well suited to quantitative analysis of migrant life events, and we thus initiated a project to digitize these data resulting in the Migration Karelia (MiKARELIA) database. All personal data entries (6072 pages) were scanned at 300 dpi using a Canon c5250i copier and saved in pdf format. The pdf documents were then scanned for optical character recognition (OCR) using the software ABBYY Fine Reader 12 (ABBYY production LLC 2013) and saved in html format. The software program Kaira-core (Salmi and Loehr 2017) was written to convert Fine Reader produced html files to simpler xml-format containing the data entries. Kaira-core then reads the source text extracting the requisite data and producing a json-file containing all extracted data which then can be used to populate a structured database. A list of variables that have already been extracted and those that are available to be extracted is found in Table 1.

Table 1.
Variables that have been extracted and their validated and expected sample sizes in the Karelian forced migrant database.

Extracted Variables	Validated sample size
Sex	157,479
Name (first, last)	160,843
Maiden name	74,021
Birth date (year, month, day)	151,567
Occupation	150,165
Owned their home	88,254
Marriage year	150,312
Spouses name (first, last)	142,248
Spouses occupation	132,876
Spouses year of death	10,606
Locations of all migrations	479,984
Years of all migrations	704,080
Children (number)	224,461
Children (sex)	224,461
Children (birth dates)	213,555
Children (birth places)	205,929

Variables still to be extracted	Expected sample size
Size of farm	~20,000
Proportion of property dedicated to forestry vs agriculture	~20,000
Type of livestock on farm	~10,000
Military Service, medals and distinctions	Unknown
Wounded	Unknown
Type and severity of disability	Unknown
Lotta Svärd service	Unknown
Photographs	~40,000

Kaira-core extracts data from the source text by finding the user specified text patterns which contain the relevant data. As an example we know that all data entries begin in the following standardized way:

"maanviljelijä, synt. 28. 11. -09 Hiitolassa" ("farmer, born 28. 11. -09 in Hiitola")

Following the general pattern <profession>, <birth date> <birth place>, Kaira-core can pick profession, birth date and place from the text. In practice, of course there are variations on these general patterns which needed to be considered. For instance, some people have maiden names listed before their profession. The source text was carefully studied, however, and the vast majority of the data could be accurately extracted by following some simple guidelines and programming these rules into Kaira to generate specific search algorithms that were able to recognize and extract many of the most common text patterns. While this approach is naive and will not work with unstructured free form text, it worked well with the source material in question. A challenge in the future is to extract the data which does not follow simple text patterns but requires natural language processing. An example of this kind of information are statements like "The spouse has studied at university and has a hobby of horseback riding." Kaira-core at the moment does not extract this kind of data since it cannot connect details such as "university studies" and "horseback riding" to a specific person in the entry, who in this case is the spouse. For this to work, an adequate machine learning and natural language processing approach such as NLTK (Bird et al. 2009) could be used.

Difficulties in digitization occurred primarily due to errors in OCR and in unstructured parts of the data. Errors that were repetitive were easily corrected in the data reading stage by modifying Kaira-core to recognize and correct the errors produced by OCR. However, infrequent or unique errors required further data cleaning and manual work. While much of the data cleaning has been accomplished, the process is still ongoing. For example, the database still contains 11,554 children (5% of total) for which sex is unknown. This is because the name in the database does not correspond to any of the 1947 male and 2558 female names on the reference list used to define sex. Therefore, unique id numbers for all individuals in the database have not yet been generated. Errors in OCR account for the vast majority of the names without matches in the reference list. However, the amount of missing data due to OCR is a small fraction of the total available and of that which is extracted accurately.

Overview of the MiKARELIA database

The digitized database currently contains 88,254 primary person entries of Karelian forced migrants. In total, the database contains information on 160,843 adult individuals when spouses of the primary entries are included. The data contain 212,907 children of known sex which were born between the years 1907 and 1969. The number of indi-
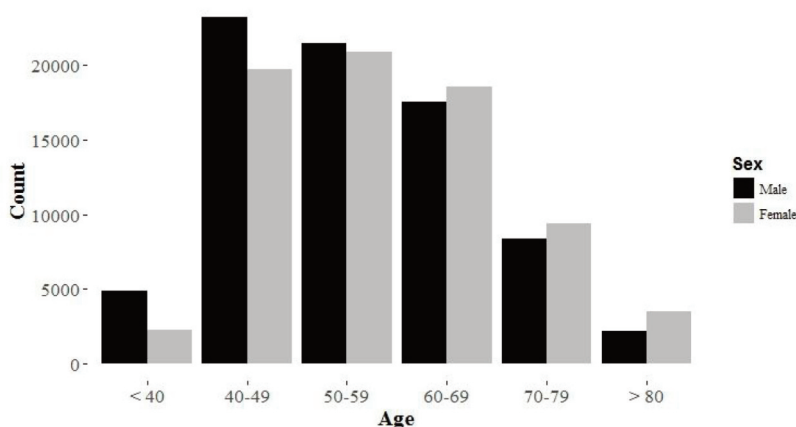
Table 2.

Number of individuals according to sex in forced migrant database based on how they appear in data (primary entry or entered as a spouse of primary) as well as number of unmarried people and number of children.

	Primary	Spouses	Unmarried	Children
Males	47,991	29,364	6,231	109,781
Females	38,364	41,760	9,000	103,126

Figure 2:

Age structure of Karelian forced migrant population in 1970. Total sample size used in figure is 151,802: 74,162 males and 77,640 females. Note that these ages are for individuals at the time they were interviewed in 1970; all were 30 years younger at the beginning of the war (e.g. individuals between the ages of 50 and 59 would have been 20–29 in 1940).

viduals in the data further grouped by sex is shown in Table 2. The age structure of the primary and spousal entries reveals that the majority of interviewees were between 40 and 79 years old in 1970 (see Figure 2) which means that they were between 9 and 48 years old at the time of the first evacuation in 1939.

We grouped the occupations that respondents reported having had in 1970 into seven categories that were based on the 1950 Finnish census, and, as expected, agriculture

Table 3.

Occupations of individuals in forced migrant database in 1970 categorized according to 1950 Finland census.

Occupations	Frequency
Agriculture and forestry	41,196
Factory workers and craftsmen	30,282
Service	10,110
Technical professionals and teachers	9,387
Transportation, mining and industry	7,789
Business and selling	7,472
Directors, office workers and typists	5,953

and forestry related occupations are the most common followed by factory workers and craftsmen (Table 3).

One of the most valuable aspects of the database is the detailed records on the movements of individuals across their lifetime. An overview of the migration frequency over time is presented in Figure 3. This shows the peak migrations events in 1939–1940 and again in 1944 that are linked to the two different Soviet occupations of Karelia at the end of the Winter War (1940) and again immediately following the Continuation War (1944). To further aid in the visualisation of the migration and forced migration, we created an animated heatmap that shows the geographic locations of places moved to for each of the years between 1912 and 1963. The heatmap can be accessed at: https://tuomassalmi.com/siirtokarjalaiset-visualization/.

Finally, we combined the information on occupation and migration events over time
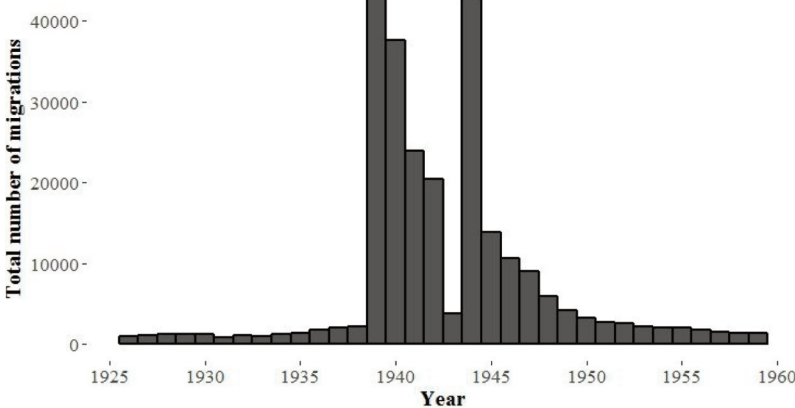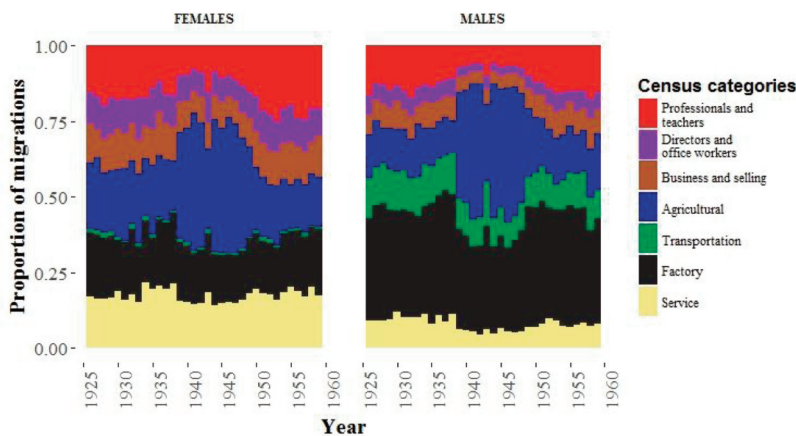

Figure 3:
Frequency of migrations by Karelians across years (1925–1960).

Figure 4:

Proportion of total migrations undertaken by females (left panel) and males (right panel) across occupations by years 1925–1960). Occupations are categorized by the 1950 Finland census.

for both sexes. Figure 4 shows the proportion of individuals from each occupational category (e.g. transportation or service sector) that migrated in each year between the years 1925 and 1960. The Figure reveals both which types of professions were more likely to migrate and the proportion of each professions by sex. Two patterns are worth mentioning here. First the proportion of workers in the agricultural sector (e.g. farmers) spikes during the years 1940 and 1950. In other words, farmers were more likely to migrate during and immediately after the war than in the years before and after. The opposite pattern can be seen for the first 3 categories in Figure 4 where professionals, director's, office workers and those involved in business and selling migrate more before and after the war and less during the war. Second we can see that females are less likely to work in factories or the transportation sector than males across all years (see Figure 4).

The demographic structure of these data are consistent with historical records, census data and data compiled by the Finnish government (e.g. statistics Finland). In order to obtain a rough approximation of the proportion of the original displaced population that are captured in our database we calculated the population in Finland who were between the ages of 10 and 49 in 1939 using Statistics Finland open data (Statistics Finland's PX web databases). We then calculated the corresponding estimate for the Karelian region (11.4% of total population, or 267,900) and compared that to the number of Karelians in our data that were between these same ages in 1939 (141,749). Using this method we arrive at a conservative minimum estimate of 46% of the total number of migrants aged 10–49 who fled Karelia during the war. This estimate is conservative in that it does not include any individuals with missing life history data such as sex, full name and year of birth. These missing individuals can later be integrated into the database as errors in Optical Character Recognition are corrected.

The percentage of Karelians alive in 1970 who were interviewed for the register, however, is likely to be much higher than 46%, because mortality between 1939 and 1970 would have significantly reduced their numbers. The life expectancy for Finns born in 1912 (the mean birth year of individuals in our data) who have survived to the age of 15 is 40.7 additional years of life for males and 47.4 additional years for females (Kannisto et. al. 1999). This puts the expected year of death for males born in 1912 at 1968 and females at 1974. We will need more detailed life history tables on age and sex specific mortality rates over time to generate a better estimate of refugees who are expected to have survived until 1970, but a preliminary analysis of these data suggests that there are records on approximately 75% of the Karelian migrants who were alive at the time of these interviews. Therefore, this is not a statistical sample of the population rather MiKARELIA should be considered a population based database.

Future uses of the database

MiKARELIA is currently being used to investigate the experiences of forced migrants as part of the three year project 'Learning from our past: the effect of forced migration from Karelia on family life' funded by Kone Foundation. The database will allow us to investigate the experiences of the forced migrants on a level of detail not previously possible. This multidisciplinary project will investigate the socio-economic and social integration of migrants into society through marriage and the reproductive outcomes of the forced migrants. These data will be used to answer a wide range of research questions including 1) what are some of the key factors that affect assimilation into the host population, 2) what is the impact of forced migration on social class and status, 3) what effect does high male mortality during the war have on sexual selection, parenting and mating strategies and 4) how does sex, population density and social class affect the degree of integration and attitudes towards out-groups.

As part of this project, we also intend to combine the Karelian forced migrant data with other available datasets, such as the digitized Katiha database (Karjala-tietokantasäätiö 2017). This database contains vast amounts of data from Karelian church registers from 1700–1949, which will allow us to access more detail on the migrants not available in the current database (e.g. religion, vaccination records, previous professions etc.), and allow for multigenerational analysis as well. We have also digitized other registers similar to the 'Siirtokarjalaisten tie' register (Anon. 1970), which will eventually be combined with the current database. For example, the registers 'Suomen rintamamiehet' (Anon. 1975–1991; 'Finland's Front Line Soldiers') and 'Suomen pienviljelijät' (Anon. 1966–1969; 'Finland's Small Farms'), contain additional information on forced migrants, and they have been digitized and data extracted using similar methods to those described above. These other registers also contain valuable information on Finns who lived in areas that did not require evacuation, allowing us to directly compare the lives of Karelian evacuees with Finns who were resident in their homes throughout the war.

The MiKARELIA database will offer excellent opportunities for multidisciplinary research and allow insight into forced migration not previously possible. The extensiveness of the digitized database provides opportunities for academic study on a forced migration event documented in unprecedented detail. To make the most efficient use of this resource we encourage researchers interested in accessing the data to contact the corresponding author.

Acknowledgements

References

ABBYY Production LLC. (2013) ABBYY Fine Reader [computer software]. Version 12.0. Moscow, Russia.

Alasuutari, P., & Alasuutari, M. (2009) Narration and ritual formation of diasporic identity: The case of second generation Karelian evacuees. *Identities: Global Studies in Culture and Power*, 16(3), 321–341.

Anon. (1966–1969) Suomen pienviljelijät. Turku, Finland: Turunmaan Kirjakustantajat.

Anon. (1970) Siirtokarjalaisten tie. Turku, Finland: Nyky-Karjala.

Anon. (1975–1991) Suomen rintamamiehet. Lieto, Finland: Etelä-Suomen Kustannus.

Bird, S., Klein, E., & Loper, E. (2009) *Natural Language Processing with Python*. Sebastopol: O'Reilly media.

Haukka, J., Sarvimäki, M., Suvisaari, J., & Martikainen, P. (2017) The Impact of Forced Migration on Mortality: a Cohort Study of 242,075 Finns in 1939–2010. Epidemiology, 28(4):587–593.

Jyrkilä F. (1975) Siirtoväen sopeutuminen kylmillä tiloilla. Jyväskylä: Jyväskylän yliopiston sosiologian laitos.

Kannisto, V., Nieminen, M., & Turpeinen, O. (1999) Finnish life tables since 1751. *Demographic Research*, 1. DOI: 10.4054/DemRes.1999.1.1

Karjala-tietokantasäätiö (2017) web URL: http://www.karjalatk.fi/ Accessed 15.7.2017

Lowe, K. (2012) Savage Continent: Europe in the Aftermath of World War II. Macmillan.

Raninen-Siiskonen T. (1999) *Vieraana omalla maalla*. Tutkimus karjalaisen siirtoväen muistelukerronnasta. 387 pp., PhD thesis, University of Joensuu.

Salmi, T. & Loehr, J. (2017) Kaira-core [computer software]. Lammi Biological Station. Available at https://github.com/Learning-from-our-past/Kaira-core

Sarvimäki, M., Uusitalo, R. and Jantti, M. (2010) Long-Term Effects of Forced Migration. IZA Discussion Paper No. 4003. Available at SSRN: https://ssrn.com/abstract=1351167

Statistics Finland's PX, web URL: http://www.stat.fi/tup/pcaxis/px_web_ominaisuudet_en.html Accessed 13.7.2017.

Waris, H., Jyrkillä, V., Raitasuo, K., & Siipi, J. (1952) *Siirtoväen sopeutuminen. Tutkimus Suomen karjalaisen siirtoväen sosiaalisesta sopeutumisesta*. Otava, Helsinki 1952.