Royal Netherlands Academy of Arts and Sciences (KNAW)
KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

Data mining in the (historic) Civil Registration of The Netherlands from 1811 - present

Bloothooft, G.; Mandemakers, K.; Brouwer, L.; Brouwer, M.

published in
Proceedings CNRS-INSHS Workhop "Family name between socio-cultural feature and genetic metaphor. From concepts to method
2010

document version
Publisher's PDF, also known as Version of record

document license
CC BY

Link to publication in KNAW Research Portal

citation for published version (APA)
Bloothooft, G., Mandemakers, K., Brouwer, L., & Brouwer, M. (2010). Data mining in the (historic) Civil Registration of The Netherlands from 1811 - present. In *Proceedings CNRS-INSHS Workhop "Family name between socio-cultural feature and genetic metaphor. From concepts to method* (pp. 1-8). MNHN.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the KNAW public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the KNAW public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
pure@knaw.nl

Download date: 27. Mar. 2025

Data mining in the Dutch (historical) civil registration 1811-present

Gerrit Bloothooft[1,2,3], Kees Mandemakers[3], Leendert Brouwer[2], Matthijs Brouwer[2]

[1]*Utrecht University, Utrecht institute of Linguistics*
[2]*Meertens Institute KNAW, Amsterdam*
[3]*International Institute for Social History KNAW, Amsterdam*
contact: g.bloothooft@uu.nl

I Introduction

Names identify individual persons. As such, names are central in research dealing with individuals, and groups defined by properties of these individuals – such as families. In the latter, also generations come into play, carrying the dimension of time and historical developments in society. The dimension of space equally influences groups: members migrate and interact. For studies of, among others, genetics, health, demography and sociology, the identification of groups and knowledge of their dispersion in time and space is valuable if not essential information.

Identifying individuals needs not to be difficult in contemporary digital systems of civil registration, if not access to this information is severely restricted for privacy reasons. For several countries, telephone directories may provide a significant sample and a useful snapshot, but family relations among people (including generations) remain unknown. For older registrations the privacy limitation does not hold, but (digital) availability and data quality is a serious issue.

In Dutch and other modern civil registrations, people are identified not only by name but also by a persistent ID. By having the parents-IDs in the record of every individual, and a complete and accurate digital registration, all family relations in society are basically known, at least for a couple of generations. In these systems, names are not essential anymore to demonstrate relations between people. However, for older registrations, no IDs were used, and reconstruction of relations between people highly depend on their names and the description of roles in certificates of birth, marriage and decease. Accuracy of these archives is often problematic, completeness rare, and full digitization a long term goal only.

This paper reviews the current status of availability of data from the modern and historical civil registration in The Netherlands . The role of names in projects is discussed and some analytic approaches in the studies of names – under the availability of full population data – are demonstrated, with (in)direct demonstration of social processes.

II Available data and major ongoing projects in The Netherlands

II.1 Modern Civil Registration

In 2000, a new law on the Civil Registration (CR) opened the possibility to acquire data for scientific research. This opportunity was used by Utrecht University and the Meertens Institute to request two selections of data, one focussed on first names, and another on family names. These selections were provided in 2006 and 2007, respectively.

II.1.a First names

Full population data were acquired for all first names of 16 million persons alive in 2006 and of 1,5 million persons deceased since the digitization of the CR in 1994, and of 3,5 million persons deceased before 1994 but who were mentioned as parents in the records of their children. The latter required an extensive reconstruction process since a parent could appear in the records of several children, while especially the deceased parent's information – not essential anymore for the current administration – contained serious numbers of errors. Besides all first names, also the (internal) ID, the first names and IDs of the parents, the date, place and country of birth of all, were provided. This basically constitutes a full population genealogy for several generations – but with only the first name known. The data are largely complete from 1930 onwards, but still provide a 30% sample in 1880. All in all, these 21 million persons entailed 500.000 unique first names which were made public in June 2010 on www.meertens.knaw.nl/nvb. The website provides for each first name the number of namesakes alive in 2006. Information on the first names is differentiated according to first or subsequent name positions and gender. The names are presented by way of a distribution per year from 1880 until present which shows the popularity of names through the ages, a geographic distribution of places of birth of namesakes alive in 2006 (468 municipalities), and an etymological description of the name (limited to 20.000 names only, but many variants still to be linked), see an example screenshot figure 1. All presentations are available in absolute and relative figures.

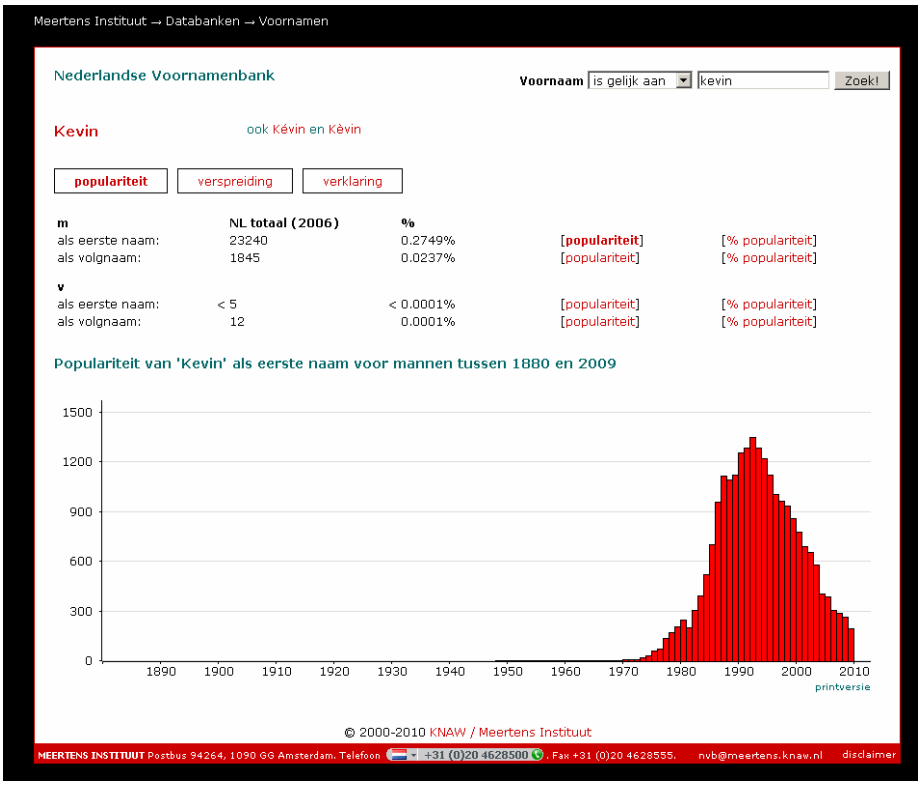

Figure 1. Screenshot of the website of Dutch first names, showing the popularity of *Kevin*.

II.1.b Family names

For the family names, full population data were acquired for the 16 million persons alive in 2007 with information about the following attributes: the family name, date, place and country of birth, and the current place of living. These data were linked to the data from the 1947 census which is available in digital format on the aggregation level of the province (number of persons with the family name); the figures for municipality of living in 1947 are available but not digitized. The 16 million persons proved to be named with 314.000 unique surnames. The

website presenting the surnames was launched in December 2009 on www.meertens.knaw.nl/nfb. It provides the number of persons with the target name in 1947 and in 2007, the geographic distribution in 1947 at the province level and in 2007 at the municipality level (468 municipalities), and documentation references for several thousands of names, see figure 2. Orthographic and onomastic relations between names are presented in a hierarchical network structure with the most frequent name on top.

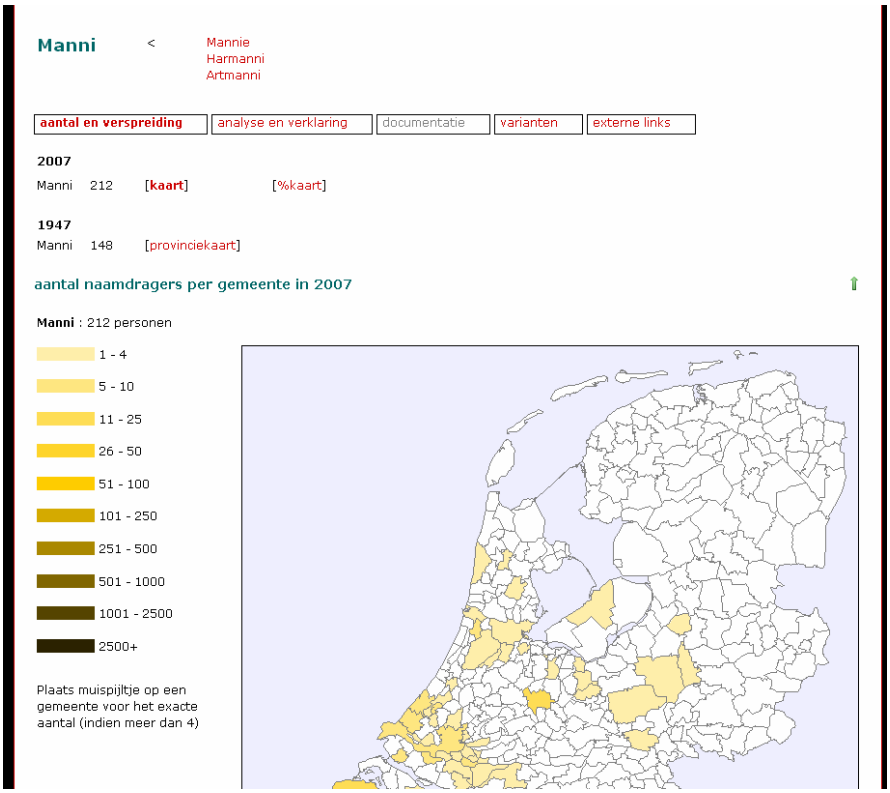

Figure 2. Partial screenshot of the website of Dutch surnames, showing the geographic distribution of the name *Manni* in the Northern part of the country.

Both websites reached 15 million page views in the first month after launching, and a stable one million page views per month afterwards, showing a very high public interest.

II.2 Historic Civil Registration

Since the early nineties of the 20th century, in The Netherlands hundreds of volunteers are working on digitization of all names and dates from the historical registers of birth, marriage and decease. The civil registration system started in 1811 and was based on Napoleontic law. Registers are public with a delay, today are available: registers of birth until 1909, marriage registers until 1934 and registers of decease until 1959. All digitized data are publicly accessible through www.genlias.nl. Currently about half of the job is done. There are now over 16 million registers digitized, containing information on about 70 million (not unique) persons.

Automatic reconstruction of families from these data is now in progress in the LINKS project (*Link*ing *s*ystem for historical family reconstruction). The project started in 2009 and is based at the International Institute of Social History in cooperation with Utrecht University, Meertens Institute and the Leiden Institute of Advanced Computing. For more information about the LINKS project, see http://www.iisg.nl/hsn/news/links-project.php

Ideally, the goal of LINKS is to identify all individuals mentioned in the certificates uniquely, and, just like the modern CR, to tag them with a persistent ID and the IDs of their parents. By using the same techniques in theory it will be possible to link our historical 'population registration' with the current one. However, for reasons of privacy protection and the probably high costs we do not foresee such a match in the near feature. In the LINKS project we systematically explore techniques for automatic record linkage under the condition of incomplete and partly inaccurate data. Geographic, onomastic and phonetic-linguistic knowledge in combination with fuzzy learning techniques will be applied in the reconstruction process.

II.3 Historical Sample of The Netherlands

The Historical Sample of The Netherlands is a project that started in 1991, with the aim to reconstruct life cycles for an unbiased random sample of eventually 78.000 persons (born 1812-1922) on a manual basis. The research persons were sampled from the birth certificates. In addition to standard personal data, also religious affiliation, occupation, household composition, literacy, social network, and migration history are collected from the civil certificates and population registers. By providing this representative dataset the HSN not only supports research with micro-data into social developments in the 19th and 20th centuries, but also a) provides a control group or groups which researchers can compare with their own research population, b) develops the expertise which individual researchers usually cannot acquire in the limited time at their disposal and c) offers the possibility for researchers to use the existing HSN dataset as base for their own research projects (Mandemakers, 2001)

Collecting new data is realized by taking the database as a starting point for further research, both through increasing the number of individuals included (oversampling) and by deepening by means of recording supplementary variables for a specific group of research subjects. Scholars thus kill two birds with one stone. Not only can they use the data already recorded, the software and expertise developed by the HSN are available as well. This expertise can also be considered an important byproduct of the data entering of the past ten years. For using its software and already recorded data, the HSN sets the precondition that new data must be added to the data set, so that they will eventually become available to other researchers too. Over eighteen projects have been realized now collecting additional information, especially in the fields of migration, both inbound and outbound (East Indies). By way of oversampling another 20.000 research persons are added to the HNS-database.

More information can be found on www.iisg.nl/hsn. The HSN supported a wide range of investigations with hundreds of publications in the areas of historical demography, history of the family, and historical sociology.

III. Data mining, considerations, tools and examples

III.1 Geographic spread

Ideally, families can be reconstructed with fair accuracy from 1811 onwards. Before that year, one has to rely on parish registers and other sources, and reconstruction – though not impossible – becomes difficult. Since more than 70% of the population already had a family name in 1811, for many families the founder of the family (who started a hereditary surname) lived much earlier. He may have had many descendants in 1811 with unknown mutual relations, but usually

with a common surname (including spelling variations). These family branches are then still genetically related. Although the size (today and in 1811) and geographic spread of the family can give an indication whether identified branches have a common founder, this is not guaranteed. A surname may have been come into existence independently in different places. In that case there is no consanguinity among families with the same surname. This especially will be the case for patronymics, occupational names and provenance names.

Current geographic spread of a family name can be shown immediately on the website of the Dutch Family Name Corpus at the municipality level. By providing an online possibility for search by *regular expression*, properties of all kinds of *sets* of surnames can be shown as well, see the example in figure 3. These properties may include all kinds of spelling variation, or require the presence of certain morphemic properties which may be typical for some language or dialect.

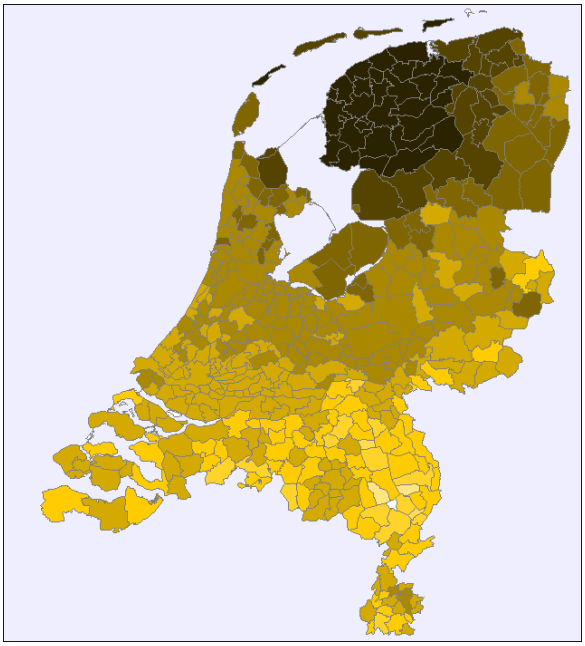

Figure 3. Geographic distribution of all surnames that fulfil the regular expression 'stra$', implying names ending on –stra, in percentage per municipality. This is a typical Frisian name ending, expressing 'coming from'. The map shows the province of Friesland, the circular shape of the decrease of the presence of the name in the North, a relative sharp boundary with the Catholic south of the country - with exceptions in areas of industrial development (in the coal mines of Limburg, around Eindhoven (Philips) and the textile factories in the eastern part).

Reverse analyses, in which we seek surnames with a common geographic spread can be done on the available (modern) data but have not been studied yet.

III. 2 Migration

Once a full reconstruction of the Dutch population from 1811 onwards would become available, migration studies can be performed easily by tracing the places of births of subsequent generations. This could be done for a family but also for the inhabitants of some village or town.

We performed such an analysis on the basis of our first name corpus from the modern civil registration. We did not use the first names themselves, but the information on place and year of birth of a person, and the ID of the parents for the intergenerational links. Starting with all present inhabitants of some municipality with an age between 30 and 50 years, we mapped their places of birth, and the places of birth of their parents, and of their (great)grandparents. Also, it is possible to start with all persons born between 1880 en 1900 in some municipality and map the places of birth of their children, and grandchildren. An interactive online application (expected in the first quarter of 2011) has all these possibilities, see an example in figure 4.

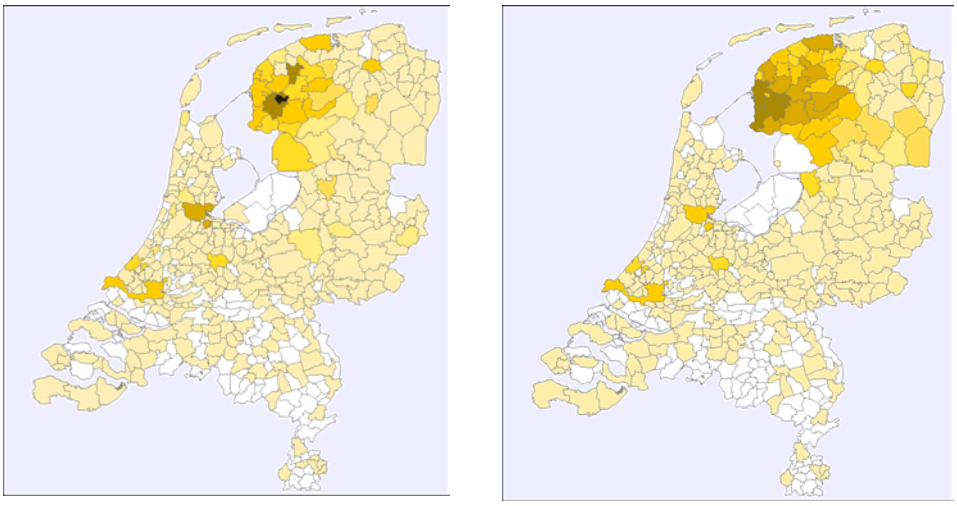

Figure 4. Places of birth of male inhabitants of the town of Sneek (dark spot), between 30-50 years of age, in the left hand panel; and those of their great-grandfathers in the right hand panel.

III.3 Co-variation

An important property of the data in the civil registration (and reconstructed life courses) is that they give fundamental parameters of life of individuals, such as names, and dates and places of birth, marriage and decease, but also the family relations among individuals. The individual data can be aggregated in time and/or space, to study population properties over time or among regions. But on the basis of known family relations, we can perform such studies within families and across generations. Doing this, we stay more closely to the social strata of the population.

We explored this in a study of modern first names. The assumption was that parents do not chose the names of their children at random, but (perhaps unconscious) on the basis of what is fashion or expected in their social environment. This would imply that the names of children in the same family convey a little bit of this fashion. Traditional parents may name their children with old Dutch names like *Willem* and *Dirk*, and this combination of names will appear much more frequent than can be expected on the basis of individual probabilities of the names. By analysing the names of millions of children in families with more than one child, we could cluster the names in such a way that names within a cluster have a higher probability to be found in a single family than across clusters (Bloothooft and de Groot, 2008). For modern naming, about fifteen clusters or name groups gave a fair description of the 1,409 most frequent names (naming 75% of all children). The geographic spread of each name group has significant features across the country, as shown in figure 5 for traditional Dutch names, which mainly follow the Dutch bible belt, while short English names are preferred in the areas of Catholic dominance – which earlier choose traditional Latinized names.

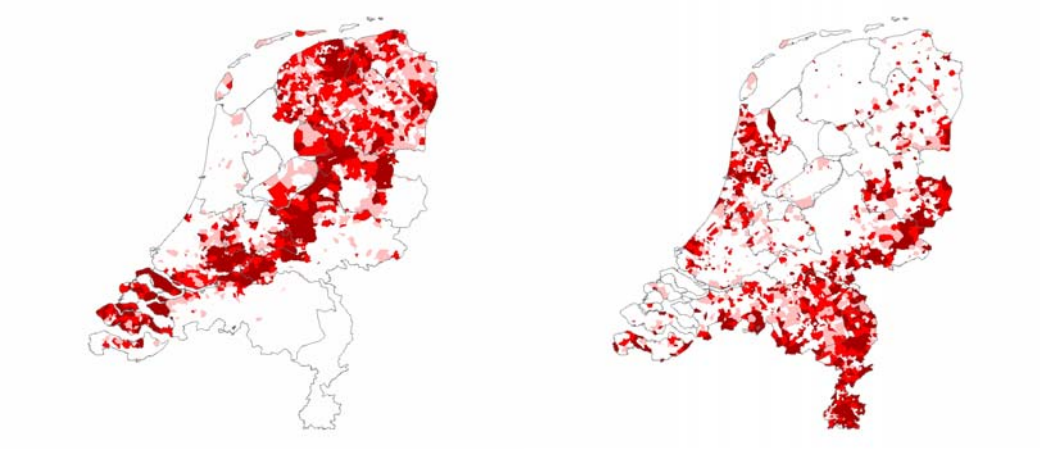

Figure 5. Geographic spread of Dutch traditional first names (left) and short English names (right).

The advantage of such an approach is that the study of names can concentrate on the properties of these 15 name groups, rather than on 1,409 individual names. In a subsequent study we had available all kinds of socio-economic data from about 281,751 households, including the names of the children in the households. This allowed to investigate the relation between socio-economic parameters, such as educational level and income of the parents, and the name groups. We also had lifestyle profiles of the households (summarizing all data), and could map the name groups on major lifestyles dimensions related to them (Bloothooft and Onland, 2011). Results are shown in figure 6, with the horizontal axis related to household income or education of the parents (low-high), and the vertical axis related to affinity to tradition versus trends. Major features are the tendency for well-educated and somewhat traditional parents to choose Dutch, Hebrew or Frisian names, while the medium educated and trendy parents favour foreign or fancy modern names.

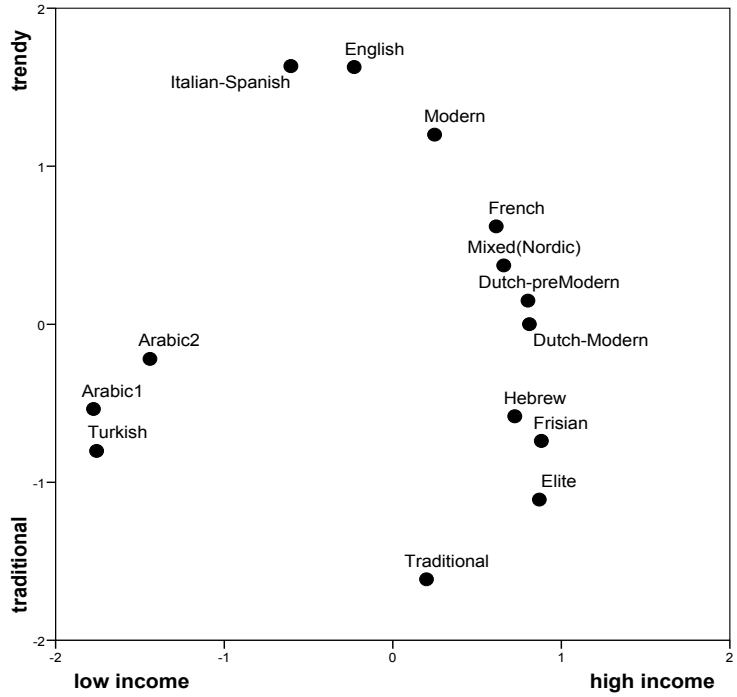

Figure 6. Name groups and life style dimensions.

The relation between the name groups of the parents and those of their children is of interest to study the dynamics in naming across generations: is fashion dynamics social class specific, or is there a much more complex diffusion of fashion among generations and social classes? This topic is currently in focus in our research.

For surnames, relations may exists between names of husband and wife, but establishing groups of surnames on that basis may be difficult because of the relative low frequency per generation of many family names. Data, from sources external to the civil registration, comparable to those shown for first names, like family income, education level, occupation, ethnicity, could be used to describe relationships between surnames and cultural, ethnic and linguistic (CEL) parameters.

IV. Privacy issues

Collecting, working with and publishing person data raises all kinds of privacy issues. For long, this has very much limited research on data from the Civil Registration. Now these data are available to some extent (in The Netherlands), high levels of security and greatest care in publications is a necessity. Errors are unacceptable, given the serious consequences this may have, both for those concerned, but also for future research and developments. Apart from the demand of data security, the rule is that no individual may be identified from published data. Especially with rare names this becomes an issues, especially if data are presented at the level of a municipality. We have considered comparable websites in Norway and France, and decided not to publish any figures less than 5. In the maps we present selected information on the basis of estimation of the risk of identification, while we also carefully looked into situations where a combination of selections (for instance using regular expressions) may convey individual information. We did receive only a couple of complaints about information that indivuals did not want to see published, and for which specific solutions were found. But we got many more complaints about information we deliberately do not publish, whereas the persons concerned want to be visible on the website. In any new applications and publications privacy remains a continuous and serious concern.

V. European co-operation

People migrate across borders. For the understanding of all kinds of onomastic, social and cultural issues, data on a wider scale than the national level would be a prerequisite. There is a need for co-operation at the European level, to discuss the collection and exchange of (historical) personal data in a common frame work under constraints of privacy, the development of tools for spatial and temporal analysis, among others.

References

Bloothooft, G. and L. Groot (2008), 'Name clustering on the basis of parental preferences', Names 56:3, 111-163

Bloothooft, G. and D. Onland (2011), 'Socioeconomic determinants of first names', accepted for publication in Names.

Mandemakers, K. (2000), 'The Netherlands. Historical Sample of the Netherlands', in: P. Kelly Hall, R. McCaa & G. Thorvaldsen (ed.), *Handbook of International Historical Microdata for Population Research* (Minnesota Population Center Minneapolis 2000), 149-177.