This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676060

The Integration of Geographic Context in Historical Databases

Diogo Paiva

Early-stage researcher of the LONGPOP project at the International Institute of Social History of the Koninklijke Nederlandse Akademie van Wetenschappen (KNAW)

This report offers an overview on how database managers deal with historical addresses and geographic context. Specifically, it consider a set of databases present in the EHPS-Net (European Historical Population Samples Network) that collected longitudinal individual data. Given the diversity and heterogeneity of historical sources, database managers have dealt differently regarding the geographic context present in the databases. Consequently, the ability for researchers to employ statistical analysis that cross individual level data and geographic contextual data is fundamentally determined by the decisions made by data managers.

This is the second report on historical addresses related to the LONGPOP project, in a series of three. It differs from the previous, which was more focus on the theoretical framework dedicated on how addresses were assigned in the 19th and 20th century. It seeks to show the present reality on how information is collected, processed and made available for later use for researchers. The third report will mainly consist of a list of best practices in implementing GIS in longitudinal databases, shifting therefore the focus for the future.

The EHPS-Net currently includes 32 databases referring to sources of all over the world. Despite the fact that most are located in Europe there are also some from North America, China and Australia. The information presented in the website was collected through questionnaires given to those responsible for the databases that filled them. These questionnaires inquired on all sort of information regarding the databases, the projects that support it and the people behind their construction and maintenance.

Specifically to geographic content, database managers where asked the geographic scope (territorial coverage and its level), variables and reference/coding systems. For purposes of the present report, it was collected and analysed information regarding both the variables and the coding system used. Some additional information was found in published articles.

Out of the total, seven databases do not collect any geographic information or documentation relating to geographic variables is not available. For the remaining, different levels of capture of spatial dimension exist. Few of them have been geocoded, i.e. use of geographic coordinate system that assigns points or polygons. Therefore, most of the information are textual variables

DISCLAIMER: This publication reflects only the author's view and the Research Executive Agency is not responsible for any use that may be made of the information it contains.

of the names of geographic context or a textual codification. Simultaneously, since sources of different nature were used, the geographic scope vary significantly. Thus, geographic variables may refer to houses, streets, neighbourhoods, villages/towns, parishes and municipalities.

In Table 1 it is summarized the key elements regarding geographic context and its integration in the databases.

Seemingly, two main factors drive the process of integration of geographic context: sources and researchers. Firstly, the information provided by sources determine many aspects of the geographic data that database managers will integrate. Fundamentally, if sources do not provide information there cannot be any data entry. Therefore, the way data is integrated is shaped by the form historical information is supplied. In the same way, heterogeneity in sources will be reflected in the database records. Even in structured historical sources, like civil records, the clerks could include complete addresses, just the locality or even leave it blank. Thus, database managers will have to find the best form to integrate this heterogeneity and provide the best datasets for researchers to use. Nonetheless, it is up to database managers to add more information to it and to georeferenced the data (i.e., to convert textual information into coded data that can be analysed and visualized using specific software), be it municipalities and provinces or house addresses.

Secondly, since the main focus of databases present in the EHPS-Net is to provide micro-level data of individuals, the geographic context plays a secondary role. While data entry process contemplates the inclusion of spatial information, most databases do not go further. Possibly, the effort to geocode is related to the needs of the researchers using the database. For example, executing a spatial analysis on the municipal or locality level may be enough for the large majority of historical demographers that want to capture regional emigration or compare rural and urban contexts.

Nevertheless, some projects did developed their databases to include detailed georeferenced data. The *Utah Population Database* provides detailed information of individuals that may be combined for the purposes of enabling research on the different fields of demography, genetics, epidemiology, and public health.

The *Scanian Economic Demographic Database* goes as far as collecting and georeferencing information on land properties that permits the analysis on wealth, living conditions and individuals' life outcomes.[1]

The project *Digitising Scotland*, currently under development, digitize Scottish vital records and implement a historical address geocoder and a matching algorithm to link individuals in historical sources to a street level[2].

[1] Hedefalk, Finn, Svensson, Patrick & Harrie, Lars, 2017, «Spatiotemporal historical datasets at micro-level for geocoded individuals in five Swedish parishes, 1813–1914», *Scientific Data*, 4.

[2] Communication presented at the European Social Science History Conference 2016 Valencia: Daras, K, Feng, Z, Dibben, C, & Williamson, L, 2016, *Digitising and geocoding historical vital events in Scotland from 1855 to 1973*.

Name	Scope	Levels	Integration	Additional documents and observations
Antwerp COR*-database	Antwerp District, Belgium	Municipalities City quarter Addresses (street + house)	Use of HISGIS and creation of IDS format	De Mulder, Wim & Neyrinck, Ward, 2014, *Documentation construction IDS database with Antwerp COR*-data*, Leuven : Centrum voor Sociologisch Onderzoek. Matthijs, Koen & Moreels, Sarah, 2010, «The Antwerp cor*-database: A unique Flemish source for historical-demographic research», *The History of the Family*, 15(1): 109-115.
Aranjuez Database: Individual and family trajectories	Aranjuez, Spain	No information		
BALSAC: Quebec population database, 1621-1971	Quebec, Canada	Localities Various levels	Numeric codes Georeferenced	
Base TRA Patrimoine	France	Municipalitie Parish	Text?	Bourdieu, J, Kesztenbaum, L., & Postel-Vinay, G., 2014, «The TRA Project, a Historical Matrix», *Population*, 69(2): 191-220
China Multigenerational Panel Database-Liaoning	North and south-central Liaoning, China	Region District Village	Numeric codes Coordinates Georeferenced	Codebooks

China Multigenerational Panel Database-Shuangcheng	Shuangcheng County, Heilongjiang, China	Village	Coded	Codebooks
Female Demographic Biographies: Wald parish, 1880-1938	Wald am Schoberpass, Austria	No information		
Founders & Survivors: Tasmanian life courses in historical context	Australian colonies and states, United Kingdom and Ireland		Georeferenced	Bradley, J., Kippen, R., Maxwell-Stewart, H., McCalman, J., & Silcot, S., 2010, «Research Note: The Founders and Survivors Project», *The History of the Family*, 15(4): 467-477
Geneva Demographic Database	Geneva, Switzerland	Localities (Birth place) *Arrondissements* Streets	Numeric? codes	Alter, G., & Oris, M., 2005, «Childhood Conditions, Migration, and Mortality: Migrants and Natives in Nineteenth-century Cities», *Social Biology*, 52(3-4): 178-191
Historical Database of the Liège Region	Liège, Belgium	Municipalities Locations	Georeferenced (Lambert coordinates)	
Historical population database of Transylvania, 1850-1914	Transylvania, Romania	Localities	Adapted LAU-2 codes (Eurostat)	
Historical Sample of the Netherlands	The Netherlands	Provinces Municipalities Localities Addresses (street + house)	Text (Addresses) Numeric codes (Municipalities) Georeferenced (localities)	
Historical Sample Portuguese Social Mobility, 1850-1960	Portugal	Parishes	Numeric codes	

Hungarian Historical Demographic Database	Szentegyházasfalva, Eastern Transylvania, Romania Kápolnásfalva, Eastern Transylvania, Romania	No information		
Integral History Project Groningen	Groningen, The Netherlands	Municipalities	Standardized text	
Italian Historical Population Database	Casalguidi, Tuscany, Italy Madregolo, Emilia, Italy	No information		
Karelian Database	The Old Eastern Finland, Karelia, Finland	Village	Text?	Räisä, J., & Loponen, M. (2014). The modernization, migration and archiving of a research register
Koori Health Research Database	Australia			A GIS mapping of population movement is under construction using the Police censuses
Melbourne Lying-In Hospital Cohort	Australia		Historically specific SES coding derived from Charles Booth's poverty maps of London, 1890s	
Mosaic project	Europe	Locations with longitude and latitude		Szołtysek, Mikołaj & Gruber, Siegfried, 2016, «Mosaic: recovering surviving census records and reconstructing the familial history of Europe», *The History of the Family*, 21(1): 38-60

National Sample of the 1901 Census of Canada	Canada	Localities	Text	User Guide
Norwegian Historical Population Register, 1800-1964	Norway	Municipalities	Numeric codes	
Odense database: Persons and buildings in Odense, 1741-1921	Odense, Denmark	No coding		
POPLINK	Skellefteå region, Umeå region, Sweden	Parish	Numeric codes	Wisselgren, M., Edvinsson, S., Berggren, M., & Larsson, M., 2014, «Testing Methods of Record Linkage on Swedish Censuses», *Historical Methods*, 47: 138-151
POPUM	Sweden	Parish	Numeric codes	Wisselgren, M., Edvinsson, S., Berggren, M., & Larsson, M., 2014, «Testing Methods of Record Linkage on Swedish Censuses», *Historical Methods*, 47: 138-151
Portuguese Genealogical Repository	Portugal	Parish	Administrative codes	
Registre de la population du Québec ancien	Quebec, Canada	Province (of origin in France) Parish	Text	
Scanian Economic Demographic Database	Scania, Sweden	Parish Village Farms	Georeferenced (Farms)	
Texas Counties Database	Texas, US	No coding		

The Demography of Victorian Scotland: Linked data for 4 Scottish communities, 1861-1901	Scotland		No coding	
The Roteman Database	Stockholm, Sweden		Georeferenced	
Utah Population Database	Utah, US	State County City Census tract Census block Address	Georeferenced	
Digitising Scotland	Scotland	Street		Semi-automated georeference in process, at street level

Table 1 - Summary of GIS integration in EHPS-Net Database

The Historical Sample of the Netherlands, which will be georeferenced in the course of the LONGPOP project, will provide researchers with the ability to track individuals residential "careers" through their lives. Spatial analysis was already conducted to the level of municipalities[3], however HSN will provide a set of coordinates for individual residential history linked to the modern postal codes system. This is a very detailed coding system that combined with a door number refers to any house in the Netherlands.

Finally, the development of the Intermediate Data Structure (IDS), within EHPS-Net, which enables the creation of datasets for analysis that include several distinct databases, thus enabling comparison between structurally different historical databases, also contemplate the use of geographic context[4]. The flexibility of this tool encouraged the development of an extended IDS version for geographic context that augments the scope and applicability of IDS to spatial analysis and comparability between databases.[5]

[3] Kok, J., Beekink, E. & Bijsterbosch, D., 2017, «Environmental Influences on Young Adult Male Height. A Comparison of Town and Countryside in the Netherlands, 1815-1900», *Historical Life Course Studies*, Online first.

[4] Alter, G. & Mandemakers, K., 2014, «The Intermediate Data Structure (IDS) for Longitudinal Historical Microdata, version 4», *Historical Life Course Studies*, 1: 1-26.

[5] Hedefalk, F., Harrie, L. & Svensson, P., 201, «Extending the Intermediate Data Structure (IDS) for longitudinal historical databases to include geographic data», *Historical Life Course Studies*, 1: 27-46.