KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

LONGPOP

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

Experiences in Geocoding Historical Population Databases. A contribution towards best practices*

Diogo Paiva

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

Contents

KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

LONGPOP

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

Introduction: Knowing the tools, setting the plan

This report reflects the experiences learned in the processes of georeferencing addresses contained in two historical databases with longitudinal data. It is intended to present a less formal and more personal and honest view on how the processes were thought and applied and its results, for consideration of future developments. One is the Historical Sample of the Netherlands (HSN) and the other one is the so-called COR*-Antwerp database (COR). This paper is written within the context of the LONGPOP project which stands for Methodologies and Data mining techniques for the analysis of Big Data based on Longitudinal Population and Epidemiological Registers (Expected Result 11). This report highlights specific challenges faced and the solutions that were implemented in the course of providing a GIS to these two established historical population databases. It focuses on three moments, namely: the problem description and planning, in the introduction; execution of the geocoding process; and future developments. In addition, a thought exercise is done concerning some relevant aspects that have stand out during the execution of the process. Both projects were thought having in mind a simple checklist to be followed:

- Establishing the goals and defining granularity
    - What is the source data?
    - How consistent is the source data?
    - What is the size of the source data?
    - How much time is available?
    - Who is going to use the output data?
    - What are the (predictable) usages of the output data?
- Listing the available resources
- Designing the methodology

While the general goal of geocoding the historical addresses for both HSN[1] and COR[2] was clearly defined from the beginning, there was still the question of better defining its output.

[1] For a more extensive description of the Historical Sample of the Netherlands database, address structure and geocoding process, see the report on *Geocoding the Historic Sample of the Netherlands Database* and HSN website. Also, for more information on addressing systems used in the past, check

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

Given that the historical addresses were composed, among other elements, of streets and house numbers, a detailed geographic information was available to locate individuals in buildings. However, defining the goal has as much to do with the scope of information as to the available time to achieve it. Finally, because of the two mentioned dimensions (source data scope and time) a third element should be considered for defining the goal: the purpose of the output. As georeferencing the databases is in its core an infrastructure process, the latter frames the output format and content. It is important to understand who are going to be the main users of the data and for what are they going to use it. This has implications on the level of detail/granularity, complexity of format and GIS model. Given the nature of the data the predicted users are researchers in Social History and Historical Demography, especially those concerned with migration studies and spatial analysis. Despite needing some detail, they are mostly concerned with data reflective of movements between countries, provinces, urban and rural settings, towns and, at most, neighbourhoods.

For both the HSN and the COR databases, the source data are historical addresses that were already decomposed at the moment of starting these projects. Their elements consisted in: house numbers (including *wijk* house numbers), street names, *wijk* (districts/quarters) names, locality[3] names and municipality names. However, the historical data is not homogenous. While many addresses identify houses, many others only refer to a *wijk* or even a municipality. This creates an uneven dataset for geocoding purposes were the ambition of having the most detailed dataset can hinder the quality of the final product. In addition, since the scientific domains where the data is expected to be used do not require an exceptional level of detail, the decision was made to establish the granularity of the dataset at a street level and use a multi-level model. Therefore, each address has one or more distinct spatial levels. Since the source data of the HSN and COR are similar but not equal, the goals had to be adjusted. For

Report on addresses over time (19th-20th century). The addresses used are contained in the HSN release of 2010.01.

[2] For a general overview of the COR-Antwerp database, see Koen Matthijs e Sarah Moreels, «The Antwerp COR*-Database: A Unique Flemish Source for Historical-Demographic Research», *The History of the Family* 15, n. 1 (2010): 109–15. For the description of the geocoding process of the COR database, check the report *Geocoding COR*-Antwerpen Database*.

[3] The term locality is used as referring to villages, towns and other geographic features related to where people wold live (fields, agrarian colonies, polders, etc.).

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 HSN it was defined a three-level model (municipality, locality and street) while for COR only a two-level model (municipality and street).

Another important point to consider before establishing the methodology is to list a set of available resources that can help in the geocoding process. This can strongly influence the direction the process takes. Resources can mean a variety of digital objects (or to be digitized) that provide means to partially or completely conclude the geocoding process, like maps (historical and modern), address books, georeferenced datasets of spatial objects, gazetteers, research outputs from other GIS projects, etc.

The differences of HSN and COR illustrate how accessing different resources shape different approaches, albeit the data is very similar. For the HSN the existence from the beginning of a georeferenced modern official dataset of addresses and buildings (BAG)[4] made the process of geocoding HSN's addresses mainly a record linkage process. Two other datasets were used for record linkage of municipalities and localities. Tasks of georeferencing historical maps and identifying places that no longer exist or changed their names were executed only in a complementary way, just as the use of data from the project Adamlink[5] to provide conversion from old to modern street names (for the city of Amsterdam only).

In the case of COR, a mixed approach of record linkage and georeferencing historical maps was used. The GIStorical[6] Antwerp project provided access to its dataset of historical streets from the city of Antwerp and an historical street inventory. This acted just like the BAG file for the HSN as it provided by record linkage a set of coordinates. However, the scope was limited to the city of Antwerp and it was only useful for about half of the addresses. For the addresses of the rest of the arrondissement it was necessary to georeference historical maps and define the geometry of old and modern streets using ESRI's ArgGIS Pro software. This provided spatial layers (lines) that in turn could be converted to a point geometry compatible with the defined model of historical coordinates.

[4] Basisregistratie Adressen en Gebouwen (BAG) is a nationwide registration of all buildings and public spaces, containing among other information on street names, associated postal code and geographic coordinates, per municipality and town. It was used the version of 2017.
[5] https://adamlink.nl/.
[6] https://www.uantwerpen.be/en/projects/gistorical-antwerp/.

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 planning and the design of a methodology can become much more robust if the above considerations are taken into account. Some of the available time for the project should be spent on this pre-phase, as it will increase the quality of the output. In parallel, too much strictness in executing a well prepared plan can be prejudicial as it is to be expected some unforeseen challenges to arise. A fair amount of flexibility and adaptability can go further if the main principles are not compromised.

Challenges in the Moment of Execution

The longer and more complex the process, the higher the possibility of the execution to diverge from the original plan. Factors like time invested, endurance over time, planning foresight, new relevant elements coming at play later, resistance in executing as planned, among others, can derail a project. The ability to apply a plan is, to some extent, an exercise of self-awareness and monitoring.

Both projects, HSN and COR, required a significant human input on normalization of data. This is especially true for the former, which had a much larger dataset and covered a wider variety of contexts. This made it a longer project which entailed a higher risk on the execution of the plan. Also, because of the predicted time it would take, around two years, it was the first to begin as to prevent delays and be able to deal with setbacks.

At the beginning of the HSN geocoding, normalization was integrally done manually, from a list of original values to be standardized (street names by municipality). The use of municipal codes (Amsterdam Code) that are coherent across time allowed to identify and group entities (in this case streets) even if localized in apparently different municipalities making the process more fluid. However, a challenge appeared when it was found after some months of execution that the algorithm to decompose historical addresses into core elements (street names, house numbers, *wijken*, etc.) misplaced some of the streets values. This algorithm was developed and its output generated prior to the start of this project. The acknowledgement of this fact implicated that an adjustment had to be made to the whole process. To prevent a loss of

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

coherence in the output database the choice was between restarting the process again with the adjustments to the initial plan or a post-normalization correction.

What was then at stake with changing the process? In the planning phase for the normalization of the data, after a general inspection of the values contained in the street name field and the documentation on data entry (particularly important for the coding of blank, unknown or unreadable values) a decision scheme was defined for the normalization itself (how to decide what standard corresponds to each original value) and for the identification of what type of geographic entity was being normalized or signalize if the entity could not be found. The latter included the distinction between public ways (streets, avenues, etc.) and localities (neighbourhoods, towns, fields, colonies, etc.). Therefore, for the cases of misplaced values by the decomposition algorithm normalization was being skipped or wrongly normalized and identification was not finding the entities.

After a moment of evaluation it was decided to continue the process as it was and later the output would be corrected in a kind of second stage normalization. This decision was based in the time already invested and the cost of its loss, the limited impact of the identified problem and the possibility of fixing later the imperfection of the process. In addition, the script designed for the normalization of values was largely based on the use of regular expressions. Selection of values by use of pattern recognition (with regular expressions), inversely to selection by specific values, is faster but works in a sequence. If new values are introduced (by way of correcting the misplacement of values, for example) the already previously written code can select unwanted values and therefore has to be re-written.

The post-planning adjustment can have a more disruptive character, as it changes the plan into something different, or an incremental character, as it builds upon the previous assumptions. Taking into consideration the influence time has over the successful execution of a project, monitoring it is of great importance. Estimation of time in the planning phase is frequently proven wrong by reality. Thus, keeping track on how well time performance is doing can motivate the development of better processes. In the case of HSN, estimation after completing the normalization of Drenthe (the first province to be dealt) was placing the conclusion of the normalization process after a year of work. Even considering that there is a learning curve and

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

experience increases productivity, manual normalization of streets was endangering the conclusion of the whole geocoding process on time. Time was then invested to accelerate the process by including sub-routines. This partially automatized the normalization, reserving the more complicated cases for human input. Time spent on this stage dropped from the initial estimation of ~280 working days to 87 effective used days.

Thinking about Future Developments and Enrichment

Ending a project without considering any future plans is a missed opportunity. Obviously, most projects come to be from opportunities in funding and human resources that are not repeatable and this discourages contemplations on further development. Nevertheless, one should consider some aspects that justify project leaders to prepare concluded projects to be retaken and improved.

At the conclusion of the HSN and COR processes, the success rate differed but for both there was potential for improvement. This can be achieved by gathering new resources, own or from third parties. Because of time limitations, archival research was very limited and thus can still provide relevant information, for example, with using maps and published address books or gazetteers. Other source of information completely untapped in these projects are administrative sources regarding opening and renaming of streets and house (re)numbering moments. The future publication of online resources from other projects should also be considered as potentially providing useful information. Although it varies from case to case, implementation of new datasets can be achieved with a limited amount of time spent.

A clear example is the use of additional resources like Adamlink, to improve the output dataset. After the record linkage of HSN's addresses with modern addresses was performed, a significant amount could not be linked as some street names are no longer used. With the data from Adamlink a conversion of the historical street names in modern names could be done and in this way new links were obtained.

While COR database is now concluded, the HSN data is still being developed with the addition of new complete live courses and subsequently of new residential careers and historical

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

addresses. This illustrates how the geocoding process with a specific data scope can be used to improve afterwards. Despite the script being specific to the HSN release of 2010.01, the general scheme of workflow can be reused. Also, the output that was created with the current geocoding of HSN can be a fundamental resource for the next release as a dictionary of normalization of historical addresses, which will greatly improve time spent on data normalization. Finally, experience from this project can be used to plan a more efficient data entry and algorithm for decomposing addresses.

The HSN has already started to survey the feasibility of improving its geocodification of historical addresses. Tom Willemsen, a student from the Radboud University, delivered a report on the expansion of HSN's historical addresses system[7]. In his report, Willemsen explores the possibility of introducing the *wijk* as element for geocoding HSN's addresses that would improve precision for the addresses that at this moment are located by the centre of the municipality (half of the addresses in HSN release 2010.01). In order to do this, he uses the province of Gelderland as a case study and estimates time, labour and financial costs of research work that includes visits to municipal archives and libraries, use of maps and address books and development for method of using the collected information. It also considers the problems that he encountered and expects other researchers might deal with in case of implementing the *wijk* system for the whole of HSN.

Insights upon Experience

The experience of designing and implementing two processes of geocoding of addresses contained in historical population databases is not without challenges and insights. In the former sections of this report, it was presented some reflexions and outcomes of these two projects, in their three main stages: planning, execution and future developments. Nevertheless, there are general principles and comments that are not confined by time and are present throughout the process. Upon reflexion on the work undertaken in the past years

[7] Willemsen, T., *Expansion and Improvement of the HSN: addresses .Translating systems of location – from addressed based on wijk-code to a street-based system* (report for Projectcollege: Big Data), Radboud University

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

regarding the geocoding of HSN and COR databases, the following aspects and considerations are thought as important for the projects' success: data coherence, dealing with time, self-evaluation, source critique, relation between data and researchers, visibility and learning with others. This selection is not intend to be an exhaustive list of necessary ingredients for a successful historical GIS. It is rather a product of personal reflexion on the specific work performed in the last years in the LONGPOP project.

Arguably, the most important feature of the geocoding process as it was implemented is ensuring data integrity and coherence. This is the foundation for the credibility of the data that researchers will use. Associated with this is transparency and clarity of the actual process. Ensuring that (geo)data was produced always under the same method, without being influenced by mood swings or change in state of mind of those processing it, is very important for providing quality outcomes. Obviously, many other factors are important for the success of a geocoding process, namely technical and scientific skills of those involved in it. But the reason for highlighting coherence as a key principle is that even if later on something is found incorrect, the rectification is simpler and more easily universal.

Another important factor conditioning the geocoding process is Time. It is one of the finite resources of any project management, alongside with funding, human resources and equipment. Time acts as kind of an omnipresent dictator, in the sense that everything is defined by considering its availability, since the beginning to the end of the project. Goals are set considering how much does it cost in time to achieve them. Monitoring how long tasks and process are taking is essential to keep control over the project and deadlines can be constant worries. Although it does set some hard boundaries, time can also act as a motivator and catalyser of improvements in efficiency and good understanding of this dimension is halfway to a successful project.

Tracking progress provides with useful insights of the project. It is by monitoring and evaluating how the processes are being developed that a project can actually improve. The subroutines introduced in the normalization process of HSN's addresses were a product of understanding the needs but also the potential to improve. Before asking to what and how can a process improve its efficiency, one as to answer: what is the progress so far? And what is

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

failing or lagging behind in the expectations? It is likewise more connected with a mindful attitude, rather than a formal procedure, that promotes a constant awareness of what is being done and its impact for the progress of the project.

A project concerning the geocoding of addresses contained in historical administrative sources, such as the civil registry and population records, is in its nature historiographic. Although both HSN and COR geocoding consisted mostly in record linkage techniques complemented by georeferenced historical maps, source critique is fundamental in understanding more accurately the data that has to be geocoded. In both projects, the source data is actually the product of several collection, maintenance and change phases, starting from the moment the clerks wrote down the addresses in the population records, in the 19th century, until the information was decomposed by an algorithm in 2010. There is a chain of production that should be taken into consideration. A variable number of individuals shaped the data to be geocoded and understanding the different processes of production is highly advantageous at the moment of once more the data suffering another interpretative process with the geocoding.

While the previous paragraph deals with looking at the data's past to understand it properly, a future perspective of the data (i.e. the data output) is also relevant. The scientific community in general concedes a great deal of credit on data output and trusts it has quality to provide the infrastructure of the analysis it wants to develop. It is, therefore, in the hands of those producing the datasets the responsibility of providing the best possible data. Transparency in the methods and decisions is fundamental and goes hand in hand with ethical procedures. There was the concern, when publishing the reports over the geocoding of the HSN and the COR databases, to present clearly the methods employed and under what assumptions and the decisions made. Researchers are therefore informed on how the geocoded datasets were obtained, what are their strengths and also their shortcomings. This empowers the researcher to decide if and how the dataset fits the goals of the analysis.

It seems, that there is a gap that needs to be attended which distances data production and data analysis in the Humanities and particularly in the historical field. The more technical nature of producing datasets that afterwards researchers can use for analysis might contribute

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

for them to distance themselves from this process. They are users of data that are detached from how it comes to be, in the same way a house painter buys paints that were created for him in mind, but that he himself do not consider to be part of the process that made the paints. Likewise, scientific venues and journals are targeted differently by those building datasets (focusing in computer methods and general IT) and those using it (social sciences). This might be contributing to a lower visibility on the value of producing data and increasing limitations of funding from research agencies for data infrastructures.

As a final remark, the need to geocode two distinct databases, albeit similar in content, allowed for a reciprocal learning experience. Much of the process designed for HSN was replicated for the COR database. The latter benefited from the experience gained in HSN and in the processes that are identical it improved the workflow to make it smoother. Planning itself was also greatly improved, especially at the level of understanding where to find more resources and how to better use them. In the other way, the method for converting historical street names, i.e. connecting old names with modern names, firstly applied in the COR geocoding project, was later used to successfully geocode streets in HSN from Amsterdam. It is not so common that parallel projects are developed by the same teams, however this case exemplifies the potential to improve methods if experiences are shared and learned. It was in this spirit that a two-day international workshop was organized by LONGPOP in June 2018.