Building a historical gazetteer: extracting place information from city directories

Vincent Ducatteeuw[1,2]
0000-0003-4493-6268
@VDucatteeuw

[1] Ghent Centre for Digital Humanities (GhentCDH), Ghent University
[2] Antwerp Cultural Heritage Sciences (ARCHES), University of Antwerp

Poster abstract

Gazetteers, also known as geographical indexes, are invaluable for Spatial Humanities *(Bol, 2011)*. Despite the growing availability of FAIR data in the Humanities, spatial information from historical gazetteers often remains unavailable as Linked Open Data. *(Wilkinson et al., 2016; Berman, et al., 2016)*. This has several reasons, including the lack of automated methods for processing historical gazetteers, the need for an adequate ontology for historical geospatial data, and the difficulty of formalizing place in a database management context (*Merschdorf and Blaschke, 2018; Garbacz et al., 2021*). This poster presents a workflow to process and model geospatial information contained within historical gazetteers and make it available as Linked Open GeoData using the upper-level ontology CIDOC CRM *(Ducatteeuw, 2021)*.

This processing method is demonstrated using a digitized collection of city directories called *Wegwijzer der Stad Gent*. The *Wegwijzer* emerged at the end of the 18th century, appearing almost annually from 1770 to 1932 by different publishing houses. The growth of Ghent during this period created a demand for accurate information about trade, industry, and services within the city. The volumes contain a wealth of information, including lists of inhabitants, services, and industries arranged geographically, alphabetically, and/or by activity type. This makes the directories a suitable source for understanding the social, economic, and demographic evolution of the city. Semi-automatic extraction methods using similar directories yielded valuable data on these topics. *(Berenbaum et al., 2016; di Lenardo et al., 2019)*.

To prepare for the semi-automatic extraction of this information, 28 volumes of the city directories were scanned in high resolution by Ghent University Library and made available in IIIF. The *Wegwijzer* was first published as the *Almanach*, after a book printer named Philippe Gimblat acquired the patent for it on 26 October 1769. Subsequently, the *Wegwijzer* was printed by F.J. Bogaert-De Clercq (1802-1827), D.J. Vanderhaeghen (1827-1837), D.J. Vanderhaeghen-Hulin (1838-1854), E. Vanderhaeghen (1855-1905?), A. Vander Haeghen (1906?-1916) and an unknown publisher (1927, 1932). Although many different publishers printed the *Wegwijzer* through time, the editions maintained a uniform structure and scope. The *Wegwijzer* continued to be bilingual in nature. The majority of the editions were written in Dutch, but occasionally, several of the listings would begin with French nomenclature rather than Dutch.

The most salient listings for spatial information were chosen from the *Wegwijzer*, including the already mentioned street index and occupation listings. The street index is an alphabetical listing of Ghentian streets. Every street has a Dutch and French name together with a location description of the street in relation to other areas. Streets are given two digits: the first referring to the police district responsible and the second indicating the jurisdiction of a specific canton of the peace court. The occupational listings are alphabetically and or occupationally organized lists mentioning the name, occupation, and street address of every resident in Ghent. The IIIF canvases of the different listings were processed using a Tesseract-based OCR workflow under development at Ghent Centre for Digital Humanities. To improve the OCR results preprocessing was done on the scans which included splitting, deskewing, and denoising the scans. Tesseract segmented and transcribed every page into text blocks and lines. The transcribed text was converted into structured CSV files using regular expressions. To make the *Wegwijzers* full text searchable by other researchers, the OCR was linked back to the IIIF manifests of the scans as JSON annotations.

The structured data was semantically modeled using the CRM-based data model for an urban gazetteer *(Ducatteeuw, 2021)*. The semantic capabilities of CIDOC CRM and CRMgeo account for all the required properties of space and place. *(Hiebel et al., 2015; Schneider et al., 2020)*. Data in the CSV files were mapped to the relevant CRM classes, converted into RDF, and uploaded to TriplyDB, an RDF triplestore supporting GeoSPARQL. The geospatial data can be queried using the Yasgui SPARQL editor, making it possible to visualize the historical data on a map. In this way, historical street information and demographic information were geographically linked back to each other and the city. The poster will provide some limited statistics illustrating these results. The database can be a starting point for demographic studies or urban morphology studies. Due to a large amount of place data available in the digital gazetteer, it can also function as a system to spatially structure knowledge, i.e., a Knowledge Organization System (KOS) *(Shaw, 2016)*. Currently, two ongoing projects called Ghent Mapped and Collective Active Belgium are investigating whether the digital gazetteer can be used to respectively map cultural heritage data and collective action data.