Reviews in Digital Humanities • Vol. 2, No. 5

Review: GeoNewsMiner

Avery Blankenship¹

¹Northeastern University

Published on: May 10, 2021

DOI: 10.21428/3e88f64f.079f1918

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Project

GeoNewsMiner

Project Directors

[Lorella Viola](#), University of Luxembourg

[Jaap Verheul](#), Radboud University in Nijmegen

Project URL

<https://utrecht-university.shinyapps.io/GeoNewsMiner/>

Project Reviewer

[Avery Blankenship](#), Northeastern University

Project Overview**Lorella Viola and Jaap Verheul**

With the ever-growing number of digital sources for research, identifying the layers of meaning humans have attached to geographical spaces over time has become more and more challenging. The GeoNewsMiner (GNM) is an interactive app that addresses this issue by identifying and mapping the spatial dimension of digitized historical sources and by fostering transparency and reproducibility, both to engage less technical scholars and to advance the digital humanities field as a whole.

GNM is currently loaded with a data set that displays geographical references in *ChroniclItaly 2.0* (Viola 2019), an open access corpus of Italian immigrant newspapers published in the United States from 1898 to 1920, which was tagged for entities using deep learning (Viola and Verheul 2020). Rather than being designed specifically for this data set, GNM goes beyond the typical limitations of most digital humanities tools (e.g., the use case's specificity, the data set, the time frame, the collection's language, specific research questions) to allow use with other data sets. Indeed, the intention was to design a tool that could be used with different data, replicate previous studies, and produce similar or different results (Viola 2020).

GNM's development has been influenced by migration studies, linguistics, history, and human geography. In the context of migration, for instance, geographical references to a homeland and a host land can be compared (Vellon 2010; Forlenza and Thomassen 2016) and studied as markers of ethnic identity (Ní Éigeartaigh, Howard, and Getty 2007). Indeed, the struggle of “being *in* a place but not *of* a place” (Ní Éigeartaigh,

Howard, and Getty 2007, p. 2, emphasis in the original) is a recurrent theme in migration experiences, whether voluntary or not.

GNM visualizes geo-coordinates of spatial references retrieved from the Google Places database and in Google Maps via the dedicated Google API. Unique to the GNM app is its layout, which was conceived and designed to allow maximum transparency. At any time, users can access the data behind the interface according to the selected levels of aggregation. For example, with the newspaper data set, users can examine data by newspaper title, date or range of dates, raw count or normalised count, country, city or region, highest or lowest percentile, and historical map. It is also possible to download and share the results through a sharable link and download the map according to selection. The interface has four exploratory tabs allowing the user to retrieve further information such as the license, a list of references, and suggestions for citation.

The full documentation of GNM is available in a dedicated GitHub repository (Viola et al. 2019) from which researchers can download the entire code bank or re-use the parts of the code more relevant to them. The repository also stores both the raw data and the processed data; in the processed data file, in particular, all the manual edits have been marked in red so that all the researcher's interventions and methodological decisions are traceable and therefore visible.

GNM is a project by Lorella Viola and Jaap Verheul; it was funded by the Utrecht University Innovation Fund for Research in IT and received support from the Research Engineering team of Utrecht University. The technical implementation was provided by Jonathan de Bruin and Casper Caandorp. The Shiny app was developed by Kees van Eijden. In addition to being presented at digital humanities conferences and meetings as well as described in a number of publications, GNM is currently used in methodology courses offered to all students in the BA in History at Utrecht University.

References

Forlenza, Rosario, and Bjørn Thomassen. 2016. *Italian Modernities: Competing Narratives of Nationhood*. Italian and Italian American Studies. New York, NY: Palgrave Macmillan.

Ní Éigeartaigh, Aoileann, Kevin Howard, and David Getty. 2007. *Rethinking Diasporas: Hidden Narratives and Imagined Borders*. Newcastle: Cambridge Scholars.

Vellon, Peter G. 2010. "'Between White Men and Negroes': The Perception of Southern Italian Immigrants Through the Lens of Italian Lynchings." In *Anti-Italianism: Essays*

on a *Prejudice*, edited by William J Connell and Fred L Gardaphé, 23–32. New York: Palgrave Macmillan.

Viola, Lorella. 2019. *ChroniclItaly 2.0. A Corpus of Italian American Newspapers Annotated for Entities, 1898-1920* (version 2.0). Utrecht.
<https://doi.org/10.24416/UU01-4MECRO>.

———. 2020. “Replication, Evaluation and Quantitative Analysis in the DH Era: Transparent Digital Practices and Lessons Learned from the Development of the GeoNewsMiner.” May 27. <https://doi.org/10.5281/zenodo.3859535>.

Viola, Lorella, Jonathan De Bruin, Kees van Eijden, and Jaap Verheul. 2019. *The GeoNewsMiner (GNM): An Interactive Spatial Humanities Tool to Visualize Geographical References in Historical Newspapers (v1.0.0)*. Utrecht: Utrecht University. <https://github.com/lorellav/GeoNewsMiner>.

Viola, Lorella, and Jaap Verheul. 2020. “Machine Learning to Geographically Enrich Understudied Sources: A Conceptual Approach - ICAART 2020.” Accessed 1 September 2020. <https://www.insticc.org/node/TechnicalProgram/icaart/2020/presentationDetails/90942>.

Project Review

Avery Blankenship

GeoNewsMiner (GNM) is an interactive spatial humanities tool focusing on geographical references in ethnic newspapers. The front pages of 4,810 issues of Italian language newspapers published between 1898 and 1920 included in *ChroniclItaly* currently serve as the tool’s use case. By focusing on place names in ethnic newspapers, GNM promises to assist researchers who are interested in the complexities of multi-layered, geographical data.

GNM addresses the many layers of information involved in thinking through questions of historic space and place. For instance, one of the features of the GNM visualization tool provides historic maps from the years 1880, 1914, and 1920 to help illustrate the changing geo-political borders throughout their period of study. This attention to changing borders and changing density in locational references in the data prompts users to consider how historic people might have considered geographic space as well

as how this informs how they write about space and place in widely distributed information vehicles such as newspapers.

To tackle some of the challenges inherent in working with the multiple layers of information required of a geographic analysis, GNM has meticulously documented their process and methodology. *ChroniclItaly* has been previously sequence tagged using Tensorflow, which yielded a total of 1,369 unique, timestamped location names. These locations were then geocoded using the Google API to first obtain place identifications in Italian and then reverse geocoded to obtain English-language details for each location. The open-access results are provided through a Shiny app that allows users to filter results according to time period, newspaper title, type of frequency, geographical information, and percentile.

Although historic newspapers notoriously feature messy OCR due to the small font sizes, marks on the pages, and other physical issues that impact machine reading, the GNM team has addressed these potential challenges by limiting their consideration of what counts as a locational reference to exclude regional, continental, historical, and natural locations. They also decided to only include locations referenced at least eight times across the whole collection to anticipate OCR errors.

GNM is a timely exploratory tool which addresses key issues in the spatial humanities, namely the importance of being able to visualize the density of locational references in historic periodicals. By focusing on visualization, this project presents the complex, multi-layered nature of geographical data in a way that is easy to understand and visually provocative. The ability for users to export and share the results of their filtering of the data will no doubt be crucial to researchers in the spatial humanities and presents the work of this project as being truly open access.