Replication, evaluation and quantitative analysis in the DH era: Transparent digital practices and lessons learned from the development of the *GeoNewsMiner*

Lorella Viola[1]

[1] Luxembourg Centre for Contemporary and Digital History (C2DH), University of Luxembourg – Belval Campus, Maison des Sciences Humaines, 11, Porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg
lorella.viola@uni.lu | @ViolaLorella

Introduction

Across fields of scientific inquiry, replication has always played an important role. Traditionally however, this has not been the case for the humanities, for which replication or access to the studies' data are not typically a concern. In fact, scholars in the wider humanities community have grown accustomed to trust published findings and conclusions often without the possibility of accessing any of the sources or without a clear explanation of the used methodology (Faull et al 2016; Jakacki et al 2016, 2015).

With the introduction of quantitative, empirical processes as opposed to qualitative, hermeneutical approaches, the Digital Humanities (DH) have started to change this tradition. Although major results in DH have been based on copyrighted material and large data collections still remain licensed, there is an increasing effort from the research community, libraries, and institutional funding agencies to facilitate research accessibility, transparency and dissemination, for instance through collaborative networks such as Open Science. In this sense, DH are especially well positioned for breaking the old schemes of obscure practices in the humanities. This, however, can only happen by putting a firm stop to the transferring of non-transparent methods into the field while encouraging replicability (O' Sullivan 2019).

This contribution discusses the evaluation steps and concerted efforts towards transparency and replicability taken during the development of the *GeoNewsMiner*[1] (Viola et al 2019 - GNM), an interactive app that maps and visualizes geographical references in historical immigrant newspapers. It describes how the goal of achieving transparency and replicability influenced the methodological decisions made in the process as well as the lessons learned from the experience. The overarching aim is to contribute to the methodological foundations of DH, arguing in favour of clearer explanations of methods and practices both to engage less technical scholars and to advance the field as a whole.

The *GeoNewsMiner* (GNM): Context of the project and research aims

GNM developed within the context of migration studies, linguistics, history, and spatial/digital humanities. The conceptual challenge behind it was to identify the layers of meaning humans

[1] https://utrecht-university.shinyapps.io/GeoNewsMiner/

attached historically to geographical spaces and how these change over time as a result of human movement and migration (White, 2010: 17). However, far from being a mere visualisation tool of geo-references, GNM was in fact conceived to assist researchers interested in geographic and conceptual space, especially from a historical perspective, detached from a specific discipline or study. Thus, rather than being on the potential results, the main focus of the project was on the methodological process which led to the tool's completion. It was precisely this shift in focus that allowed us to go beyond the typical limitations of most tools developed in DH (e.g., the chosen use case's specificities, the tested dataset, the timeframe, the collection's language, specific research questions) in order to build a resource that could be used by others as standard benchmark. Indeed, the intention was to open up a non-study dependent method which therefore could be used with different data or even generate new ones, replicate previous studies and produce similar or different results.

Transparency and replicability

As pointed out by Peels (2019), we should distinguish between *transparency* and *replicability.* Although certainly complementary to each other, these terms entail in fact different things. If on the one hand, a study can only be replicated if sufficient transparency has been observed on the data, the research purposes, the method, the conclusions, etc., on the other some studies can be perfectly transparent and yet not at all replicable. This can be especially true in the humanities as the very nature of some studies can make replication impossible (e.g., a particularly interpretative analysis). Thus, being transparent about both the raw and the processed data, about the methodology and the analytical processes is fundamental for achieving replicability but it may not be enough to make a study replicable. In order to achieve both transparency and replicability, the development of GNM entailed a twofold approach: the development of a layout for the app that could be structurally transparent and the creation of a GitHub repository[2] (Viola et al 2019b) for replicating the study, sharing the data, and guiding scholars towards re-using the methodology.

The app's layout

The app's layout was conceived and designed to allow maximum transparency. This was achieved by allowing the user to access the data behind the interface, i.e., according to the selected different levels of aggregation. There are eight levels of aggregation: by newspaper's title, by date or range of dates, by raw count or normalised count, by country, city or region, by the top highest or lowest percentile, by historical map. Users can share the results of their selections through a sharable link and download the map according to selection. The interface

[2] https://github.com/lorellav/GeoNewsMiner

has four exploratory tabs allowing the user to retrieve further information such as the license, a list of references, suggestion for citation, etc.

The GitHub repository

The GitHub repository includes two main parts: a first, general part (README) which fully describes the project, the context of the study, the dataset, the methodology, the challenges, and the external resources, and a second part, where all the source codes divided per step are provided. Researchers have in this way the possibility to download the entire notebook 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. In addition to the original collections that have been used as the tested use case (i.e., *ChroniclItaly* – Viola 2018 and *ChroniclItaly 2.0* - Viola 2019), both the original and processed data can be downloaded. Finally, to increase readability, the complete list – grouped per type – of the cases that required manual edits is also provided in the form of a narrative so that maximum transparency is also provided regarding the researcher's interventions.

Conclusions

This contribution discussed the evaluation steps towards transparency and replicability taken during the development of the *GeoNewsMiner* app and how these influenced the development process itself. A few considerations can be drawn from the experience. First, creating a fully transparent and replicable project entailed a radical shift in the adopted perspective which went from being results-oriented to being method-oriented. This shift created a space in which the wider research community also played an important role, conceptually and methodologically. The potential value of the study's replication lies precisely in this newly created space which allows for carrying out more replication studies in the humanities, a field in which independent repetitions of published research is hardly ever conducted.

Second, GNM initially stemmed from a highly interdisciplinary project with a clear research question in mind: what can subjective connotations of geographical markers in immigrant newspapers tell us about the socio-cognitive dimension of migration history? Importantly, the realisation of GNM proved that despite such a very specific research question, it is possible to create a generic tool or methodology. This consideration is an important lesson to share in order to demonstrate how transparency and replicability allow us to transcend the limitations and boundaries of specific research projects, which tend to slow rather than advance the field.

Reference

Faull, K. M., Jakacki, D., O'Sullivan, J., Earhart, J. A., and Kaufman, M. (2016). "Access, ownership, protection: the ethics of digital scholarship". *Digital Humanities*, Kraków (July 2016).

Jakacki, D., Faull, K., Porter, D. and O'Sullivan, J. (2015). "The ethics of data curation: the quandary of access vs. protection: Keystone". *Digital Humanities*, Philadelphia (July 2015).

Jakacki, D., Mandell, L. C., Morgan, P., O'Sullivan, J. and Rawson, K. (2016). *Digital scholarship in action,* presided by Patricia Hswe. *MLA Annual Convention*, Austin (January 2016).

O'Sullivan, J. (2019). "The humanities have a 'reproducibility' problem". *Talking humanities*. Retrieved from https://talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/ on the 14 February 2020.

Peels, R. (2019). "Replicability and replication in the humanities". *Research Integrity and Peer Review* (4), 2.

Retrieved from https://utrecht-university.shinyapps.io/GeoNewsMiner/ on 14 February 2020.

Viola, L. (2018). *ChroniclItaly: A corpus of Italian American newspapers from 1898 to 1920*. *Utrecht University*. Retrieved from https://public.yoda.uu.nl/i-lab/UU01/T4YMOW.html

Viola, L. (2019). *ChroniclItaly 2.0. A corpus of Italian American newspapers annotated for entities, 1898-1920 (Version 2.0)*. Retrieved from https://doi.org/10.24416/UU01-4MECRO

Viola, L., De Bruin, J., van Eijden, K., & Verheul, J. (2019a). *The GeoNewsMiner (GNM): An interactive spatial humanities tool to visualize geographical references in historical newspapers*. Retrieved from https://github.com/lorellav/GeoNewsMiner on 14 February 2020.

Viola, L., De Bruin, J., van Eijden, K., & Verheul, J. (2019b). *The GeoNewsMiner (GNM), 1898 – 1920 (v1.0.0).*

White, R. (2010). *Spatial History Project*. Retrieved 8 November 2019, from https://web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29