Russian Avant-Garde in Exile: Navigating Challenges and Opportunities in Web Scraping of Biographical Dictionaries

Elizaveta Berquin, Université Libre de Bruxelles

In this paper, I would like to share a case study of working with data from biographical dictionaries, which is part of my PhD thesis "Mapping Russian Avant-Garde Networks in Exile: A Digital Humanities Approach." Traditionally, the history of art and the history of European avant-garde have been focused on studying biographies of specific artists, groups, or so-called centers, capitals of culture. Discussions on the history of artistic emigration from the Russian Empire after the 1917 revolution also tend to focus on the lives of prominent figures in metropolitan centers such as Paris, Berlin, Prague, and Belgrade. However, this approach overlooks a vast cultural layer that includes lesser-known artists who, for various reasons, may not have emerged at the forefront of the avant-garde scene. Recent advances in Digital Humanities and its methods, such as data analysis and visualization, allow us to explore the history of the European avant-garde from a wider angle. A data-driven approach makes it possible to reveal unseen patterns, connections, and correlations. Naturally, the first question arises: what kind of data can help us achieve this goal?

Over the past decades, researchers have done extensive work, compiling data on emigrants from various sources to create biographical dictionaries of Russian Emigration. This inspired the idea of creating a dataset based on these dictionaries, containing valuable details such as artists' names, birth and death dates, places of origin, professions, family ties, and other social connections. While still relying on biographical data, this dataset allows us to visualize, map, and quantify the existing knowledge about artistic emigration to Europe on a larger scale.

This paper explores the challenges and solutions encountered in the data extraction process. Initially, these biographical dictionaries exist in the form of multi-volume physical books, and the first challenge was to figure out how to extract data from them. One possible solution is to scan them using Optical Character Recognition (OCR). However, this method requires access to all volumes and a high-quality scanner, which is not always feasible and can be time consuming. Fortunately, some of these dictionaries are available online, but the downside is that the websites are often outdated or lack data export options. In this case, web scraping proved to be a solution.

While web scraping has been widely used and there are many tutorials covering different tools, not all of them are suitable for every situation. Hence, the second challenge: which tool to choose? From Python libraries like Beautiful Soup and web crawlers like Scrapy, to ready-to-use solutions, it is not easy to know which tool would be the most efficient when encountering this task for the first time. I started by writing a Python script to extract the data; however, I quickly realized that it would not be as straightforward as I thought, since the structure of the website turned out to be more

complex due to pagination, nested containers and so on. While looking for practical solutions to these challenges, I discovered a free browser extension called Web Scraper and decided to give it a try. It offers a simple point-and-click interface, clear documentation, and a 5-minute video tutorial that covers most of the process. Web Scraper uses a modular structure that is made of selectors, which guide the scraper on how to navigate the target site and extract data. The paper describes the main steps of using the Web Scraper to extract the data from the “Artists of the Russian Emigration” biographical dictionary. It presents the obtained dataset in CSV and XLSX formats containing records about 1868 emigrated artists, as well as an example of the first test visualization and insight created using Python's libraries such as Pandas, Folium, and GeoPy (Fig. 1).

The third challenge I encountered was related to ethical and legal considerations. Web scraping can sometimes carry a negative connotation because it is often misused for commercial purposes. Ensuring the legality and ethics of my actions was made complicated due to the lack of available information, particularly within the Digital Humanities context. Currently, there is no legislation that directly addresses web scraping, and it is mostly guided by legal principles such as unauthorized data access, misuse of data, contract violations, copyright infringement, etc. Ethical concerns may include website crawling restrictions set by owners, individual and organizational privacy, diminishing value for the organizations, and potential discrimination or bias in data use (Krotov, Johnson, & Silva, 2020). Regarding the Digital Humanities perspective, non-malicious Internet data retrieval can be divided roughly into two categories: formally supported and informally permitted (Black, 2016). Therefore, in many cases, web scraping is considered legally permissible if it is conducted for research purposes and considers the legal and ethical aspects mentioned previously.

To summarize, this short paper will present a case study on how to extract the data using web scraping point-and-click browser extension, considering various technical and ethical nuances of such approach. The aim of showcasing this experience is to demonstrate that, despite certain challenges, web scraping is a powerful method that many digital humanists or scholars from any discipline could benefit from. However, while there is a growing availability of accessible technical solutions for web scraping, ethical and legal concerns remain ambiguous and require further discussion and clarification.

Figures

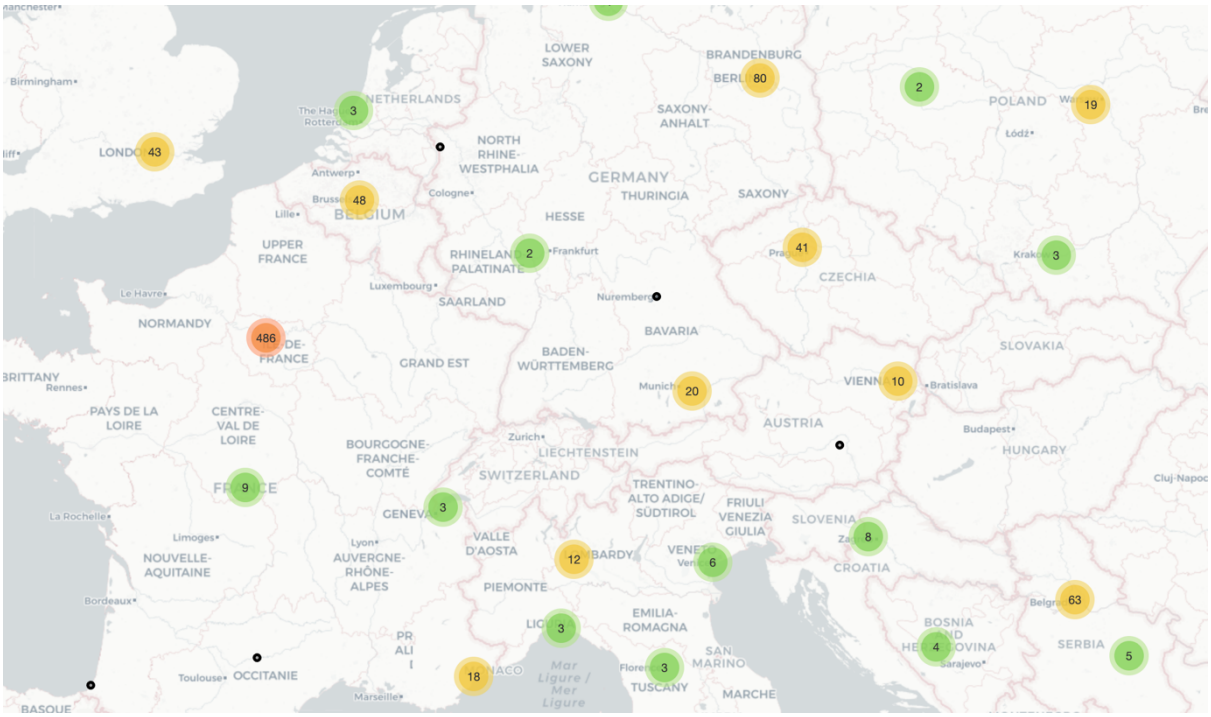

Fig. 1. The interactive map illustrates the geographical distribution of emigrant artists in Europe, showcasing preliminary visualizations of a data sample extracted from the digitized version of the biographical dictionary "Artists of the Russian Emigration: 1917-1941" (Leikind et al., 2019).

Bibliography

Barget, M. (2021). Doing Digital History with Python IV: Web Automation. *Digital History Lab*. Retrieved from https://dhlab.hypotheses.org/1939

Black, M. (2016). The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research. *International Journal Of Humanities And Arts Computing, 10*(1), 95–109. https://doi.org/10.3366/ijhac.2016.0162

Brown University Library. (2024). *Web Scraping Toolkit*. Retrieved from https://libguides.brown.edu/c.php?g=1037232#s-lg-box-23880322

Densmore, J. (2017). Ethics in Web Scraping. *Towards Data Science*. Retrieved from https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01

Iskusstvo i arkhitektura Russkogo zarubezh'ya. (2024). Retrieved from https://artrz.ru

Krotov, V., & Silva, L. (2018). Legality and ethics of Web scraping. In Proceedings of the 24th Americas Conference on Information Systems.

Krotov, V., Johnson, L., & Silva, L. (2020). Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems, 47*, 539–563. https://doi.org/10.17705/1CAIS.04724

Leikind, O. L., Makhrov, K. V., & Severiukhin, D. I. (2019). *Khudozhniki russkogo zarubezh'ia: Pervaia i vtoraia volna emigratsii: biograficheskii slovar'*.

Mitchell, R. (2019). *Web Scraping with Python: Collecting More Data from the Modern Web* (2nd ed.).

Web Scraper Documentation. (2024). Retrieved from https://webscraper.io/documentation

Williamson, E. (2022). Fetching and Parsing Data from the Web with OpenRefine. *Programming Historian*. Retrieved from https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine#why-use-openrefine