Title: The long road to text as data in Brazilian archives: the case of the Digital Library of Brazilian Newspapers

Sumary:

In this paper, I intend to analyze the characteristics of graphical user interfaces and archives of Brazilian digitized sources and evaluate possible ways to travel the long road that separates the collection of digitized documents from its transformation into a text capable of being analyzed as data in Brazil.

To do so, I will take as an example the graphical interface and the collection of the Hemeroteca Digital Brasileira (Brazilian Digital Newspaper Library), which is part of the Fundação Biblioteca Nacional (Brazilian National Library), and pyHDB, a heuristic tool developed in Python specifically for researching, collecting and organizing data from the *Hemeroteca Digital Brasileira*. The pyHDB uses web scraping techniques and aims to assist researchers in the methodological construction and recording process, generating reports and providing tabular data and image datasets from the search parameters defined by the user. From the analysis of this case, I seek to present a set of possibilities to analyze the Hemeroteca Digital Brasileira collection and advance in the transformation of raw data into computer-readable text. Ways of analyzing the metadata organized by pyHDB with Pandas and Plotly and the treatment of digitalized images with sets of tools for layout and character recognition, such as LayoutParser, OCR-d, gImageReader and Transkribus will be presented.

Therefore, this paper intends to discuss the limitations, challenges, and possible paths for conducting research in digital humanities that intend to work with text as data from sets of primary digitalized sources in portuguese. It is increasingly common for such documents to be made available online by Brazilian institutions, such as the Digital Library of the Federal Senate, with a collection related to the legislative power; the Historical Collection of the Official Gazette of the Union - DOU, maintained by the National Press, with a digitalized collection between 1808 and 2001 and the aforementioned Brazilian Digital Newspaper Library. The latter has proven to be a landmark in historical research, especially in excerpts from the 19th and 20th centuries, either through its vast digital collection of newspapers from all over the country and from other parts of the world, or through the word search engine. In general, these collections have graphical user interfaces that allow searches by simple parameters (date, authorship, subjects, publications, etc.) - with the exception of Brazilian Digital Newspaper Library, which searches for words in the body of documents - and returns lists of results to users. allowing the download of pages in PDF or various image formats. However, there are no resources for accessing the document’s text itself, or data on OCR accuracy, word frequency, the possibility of bulk downloading, or the generation of reports on searches and results.

Thus, despite the importance and growth of the digitization process and making collections available openly, the character of these collections and graphical interfaces, their search tools and types of results returned to users limits and often poses challenging obstacles for those who wish to apply text-as-data analysis methods in Brazil.

Key-words: Web scraping, digital methods, Historical Digital Newspapers, Python, OCR

References:

BRASIL, Eric. pyHDB - Ferramenta Heurística para a Hemeroteca Digital Brasileira: utilizando técnicas de web scraping para a pesquisa em História. *História da Historiografia: International Journal of Theory and History of Historiography*, Ouro Preto, v. 15, n. 40, p. 186–217, 2022. DOI: 10.15848/hh.v15i40.1904. Disponível em: https://www.historiadahistoriografia.com.br/revista/article/view/1904. Acesso em: 15 jan. 2023.

EHRMANN, Maud; BUNOUT, Estelle; DÜRING, Marten. Historical Newspaper User Interfaces: A Review. In: *LIBRARIES: DIALOGUE FOR CHANGE*. Atenas, 2017. Anais […], Atenas: IFLA WLIC, 2019, p. 1-24. Disponível em: http://library.ifla.org/id/eprint/2578/. Acesso em: 14 set. 2021.

JARLBRINK, Johan; SNICKARS, Pelle. Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *Journal of Documentation*, Bingley, v. 73, n. 6, p. 1228–1243, 12 out. 2017. Disponível em: https://doi.org/10.1108/JD-09-2016-0106. Acesso em: 12 set. 2022.

MARRES, Noortje; WELTEVREDE, Esther. Scraping the Social? *Journal of Cultural Economy*, Londres. v. 6, n. 3, p. 313–335, 2013. Disponível em: https://doi.org/10.1080/17530350.2013.772070. Acesso em: 12 set. 2022.

PFANZELTER, Eva; OBERBICHLER, Sarah; MARJANEN, Jani; et al. Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. *Journal of Data Mining and Digital Humanities, v. HistoInformatics*, jdmdh:6121, 2021. Disponível em: https://jdmdh.episciences.org/7069. Acesso em: 12 set. 2022.