HEMDIG(pt)

First, I would like to thank Joana and Daniel for inviting me to participate in this event. It is a pleasure to be here and to present this research developed last year at the DH Lab of Universidade NOVA de Lisboa. HemDig(pt) is a framework I developed to handle digitized newspapers, and the idea here is to provide a summary of what was done, the challenges faced, and the results obtained.

Motivation

- **Historical Newspapers:** In recent years, I have observed the growing use of these sources in research in Brazil, especially with the popularization of the Brazilian Digital Newspaper Library, an initiative by the National Library of Brazil. Many journal articles, theses, and dissertations have utilized these sources to investigate various aspects of Brazilian history.

But at the same time, there is a methodological silence about their use: how the research is conducted, what methods are used for processing these digital objects, and moreover, how such issues epistemologically impact the research carried out. I have been researching and publishing on these issues, and HemDig(pt) is an attempt to systematize some of these reflections and collaborate with researchers who deal with digitized newspapers. - **Challenges:** Large volumes, diverse formats, and varying quality. Challenges are also related to the graphical interfaces for accessing digitized newspapers and the research possibilities these collections offer.

- **Current Gap:** Few debates in Portuguese, few models trained for OCR in Portuguese. Do repositories in other languages face the same issues?
- **Objective:** Develop a comprehensive framework for efficient data collection and processing of digitized newspapers in Portuguese.

Research Background: History, Versioning, and Repository

- **History:** The HemDig(pt) framework was developed to address the challenges of handling digitized newspapers. It is based on my research and publications on the topic.
- **Versioning and repository:** The entire research process was versioned using Git and hosted on GitHub. The framework is published as a Jupyter-Book, and the link is available in the footer of the presentation.
- **License:** The framework is available under a Creative Commons license for non-commercial use.

HemDig(pt) Framework Phases

I will present each of the 5 phases, highlighting the objectives and characteristics of each. The goal is to provide an overview of the framework and point out some of the results obtained in each phase.

Phase 1: Planning and Preparation

The first phase consists of methodological and technical planning and preparation. This phase introduces a set of tools and strategies to develop basic knowledge and resources, ensuring subsequent phases are executed more efficiently and accurately.

- **Literature Review:** In this stage, I present Zotero as a tool for managing bibliographic references; and the conduction of a bibliographic survey on topics related to digitized newspapers in the Scopus and Web of Science databases. As a result, a collection called HEMDIG(pt) was created and added to the public reference library Digital History.
- **Basic Technical Preparation:** In this chapter, we aim to present some tutorials from Programming Historian organized into the following topics: Basic Computational Knowledge, Data Management, and Writing and Publishing.

We do not seek to conduct an exhaustive survey of themes or fill gaps in training, but rather to point out directions and define a minimum level of technical knowledge to begin the research.

- **Research Planning:** In addition to the bibliographic survey and minimum technical training, as indicated in chapters 1 and 2, it is important to define, even if generally and preliminarily, the specific planning aimed at documentation, organization, preservation, and licensing of the research.

In other words, it is essential to develop a Data Management Plan (DMP), which should be created in this first phase and updated throughout its development.

Here, we discuss the FAIR principles, version control systems, tools like Zotero, Tropy, and data licensing.

Phase 2: Critique of Collections

The second phase involves the critical evaluation of the collections and graphical interfaces of the selected repositories. This task is crucial for understanding the biases, limitations, and possibilities of each collection, and for developing consistent methodological strategies and theoretical reflections aligned with the technical, political, and theoretical aspects of each repository.

- **Case Studies:** Evaluation of three repositories: Brazilian Digital Newspaper Library, National Digital Library of Portugal, and Municipal Newspaper Library of Lisbon. Detailed analysis in Chapter 4.

I sought to present the general data of each collection and provide visualizations of this data so that future research can be developed. The general characterization of the digitized collections from these institutions is not currently available in their own search interfaces, which directly affects the research and use of these collections. Therefore, the organized and graphical presentation of this data, as done in this chapter, is a contribution to the research and use of these collections.

- **Analysis of Graphical Interfaces:** Evaluation of the graphical interfaces of the selected repositories using the Impresso project's method. Detailed in Chapter 5.

I conducted a critical evaluation of various aspects of the graphical interfaces of each collection, using the methodology proposed by Ehrmann, Bunout, and Düring [Ehrmann et al., 2019].

In their research, the authors developed a comprehensive and detailed questionnaire covering six evaluation criteria subdivided into about 140 items. In the original study, this questionnaire was applied to twenty-four graphical interfaces of historical newspaper collections.

The article presents interesting results about the different stages of each interface, as well as common characteristics and general limitations. However, the research was focused on repositories in Western Europe and the United States, and did not include any collections in Portuguese or Spanish.

Given this limitation, I sought to apply the method to the graphical interfaces of historical newspaper collections in Portuguese.

I translated and adapted the questionnaire, applied it to the graphical interfaces of the selected collections, and found significant similarities in the results obtained.

These results are presented in the next chapter.

Collection Characteristics

Interface Characteristics

Phase 3: Data Collection

In this phase, we outline strategies for the collection and organization of data and metadata from the studied collections.

- **Reports and Documentation of Searches:** Importance of consistent documentation and methodological registration.

To this end, I created small Python scripts to generate search reports, which function as methodological support tools. These scripts record data about the repositories, the searches conducted, the criteria used, and the results obtained. These reports are generated automatically and can be used to document and record the search process, facilitating the replication and review of the research.

- **Data Scrapers:** Tools and strategies for metadata and data collection.

When there are no open data policies or APIs to access data, it is necessary to develop data scraping strategies.

To collect data from the HDB, it is recommended to use pyHDB, a tool written in Python that allows you to search and download data from the HDB. The tool collects metadata from searches and downloads images of pages with occurrences of the searched terms (when permitted by copyright).

I also created a scraper for the Digital Library of Portugal and the Digital Newspaper Library of Lisbon: Collect data and metadata using Selenium and Beautiful Soup.

- **Organization of Datasets:** Strategies for organizing collected data.

Phase 4: Data Processing

This phase organizes tutorials, documentation, and tools for executing OCR (Optical Character Recognition) and OLR (Optical Layout Recognition) on digitized Portuguese-language newspapers.

This phase organizes tutorials, documentation, and tools for executing OCR (Optical Character Recognition) and OLR (Optical Layout Recognition) on digitized Portuguese-language newspapers.

We conducted a series of tests with different tools, ranging from CLI (Command Line Interface) to GUI (Graphical User Interface), and with varying levels of complexity. We also listed a series of lessons from Programming Historian that are useful for those seeking to perform OCR on newspapers.

This chapter is divided as follows:

- CLI Tools: Command Line Interface tools, namely Tesseract, OCR-D, and Kraken.
- GUI Tools: Graphical User Interface tools, namely gImageReader.
- Programming Historian Lessons

For each of these topics, we present a set of tutorials, documentation, tests, and results obtained.

Our goal, more than providing a step-by-step guide for using each tool, is to present the main features, advantages, and disadvantages, and promote critical reflection on their use.

Like the entire framework, we aim to provide researchers with knowledge and materials to consciously and critically develop their workflows.

Phase 5: Review, Preservation, and Publication

In this final phase of the framework, the objective is to conduct a comprehensive review of the documentation, metadata, and data structure, as well as to publish the research data.