HemDig(pt): A Framework for Collecting, Organizing, and Processing Digitized Newspaper Pages in Portuguese

Eric Brasil

May 31, 2024

Introduction

- **Title:** HemDig(pt): Methods, tools and Portuguese Newspaper Digital libraries.
- **Objective:** To present a framework for the collection, organization, and processing of digitized newspaper pages in Portuguese.
- **Research Importance:** Digitized newspapers are valuable historical sources that need efficient handling for research purposes.

Motivation

- **Historical Newspapers:** Rich sources of historical data.
- **Challenges:** Volume, variety, and the need for accurate processing.
- **Current Gap:** Lack of analyses, tools, and models for handling digitized historical newspapers in Portuguese.
- **Objective:** Develop a robust framework to handle these challenges effectively, specifically targeting Portuguese-language newspapers.

Research Background: History, Versioning, and Repository

- **History:**
  - HemDig(pt) evolved from the need to systematically handle large collections of digitized historical newspapers in Portuguese.
  - The framework has undergone several iterations to improve its functionality and efficiency.
- **Versioning:**

  - Implemented version control to manage updates and changes.
  - Ensures the framework remains adaptable and can integrate new features seamlessly.
- **Repository:**
  - Utilizes GitHub for code hosting and version control.
  - Provides transparency and accessibility for researchers and developers.
  - Encourages collaboration and contributions from the research community.
- **License:**
  - All research utilizes free and open-source tools.
  - Data, code, and visualizations are licensed under Creative Commons Attribution 4.0 International (CC BY-NC-SA 4.0).
  - **Terms:**
    * **Attribution:** Give appropriate credit, provide a link to the license, and indicate if changes were made. Do this in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
    * **Non-Commercial:** You may not use the material for commercial purposes.
    * **ShareAlike:** If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
    * **No Additional Restrictions:** You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.
- **Learn More:** Link to License

HemDig(pt) Framework Phases

- **Five Phases of the HemDig(pt) Framework:**
  1. **Planning and Preparation:** Establishing objectives, resources, and methodologies.
  2. **Critique of Collections:** Evaluating the quality and relevance of newspaper archives.
  3. **Data Collection:** Gathering digitized newspaper pages from various sources.
  4. **Data Processing:** Cleaning, OCR (Optical Character Recognition), and metadata extraction.
  5. **Review, Preservation, and Publication:** Ensuring data accuracy, long-term preservation, and sharing findings.

Phase 1: Planning and Preparation

Summary: 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.

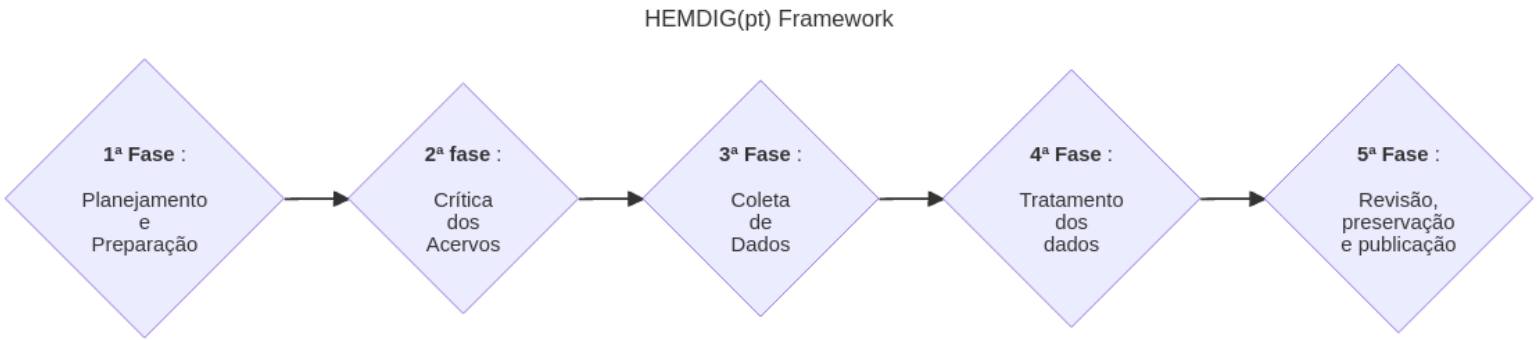


Figure 1: Framework Diagram

1. **Literature Review:**
   - Initial literature review.
   - Tools: Zotero, Scopus database search, and a public reference library.
2. **Basic Technical Preparation:**
   - Introduction to relevant lessons from Programming Historian.
   - Topics: Basic computational knowledge, data management, writing, and publishing.
3. **Research Planning:**
   - Suggestions for overall research planning.
   - Strategies: Documentation, data organization, and usage licenses.

Phase 2: Critique of Collections

Summary: 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.

Organization:

1. **Case Studies:**
   - **Repositories:** Brazilian Digital Newspaper Library, National Digital Library of Portugal, and Municipal Newspaper Library of Lisbon.
   - **Chapter 4:** In-depth evaluation of the data from these repositories.
2. **Analysis of Graphical Interfaces:**
   - **Chapter 5:** Evaluation of the graphical interfaces of the selected repositories.
   - **Method:** Utilizes the method developed by the Impresso project team for graphical interface evaluation.
3. **Comparative Effort:**
   - **Chapter 6:** Elements to support the comparison of collections and interfaces.
   - Presents data and a series of visualizations and general comparisons.

Note: - This chapter provides data and visualizations for general comparison. - In-depth analyses are beyond the scope of this research and will be used for future articles and publications. - Data is available under the CC-BY-SA 4.0 license for use by other researchers.

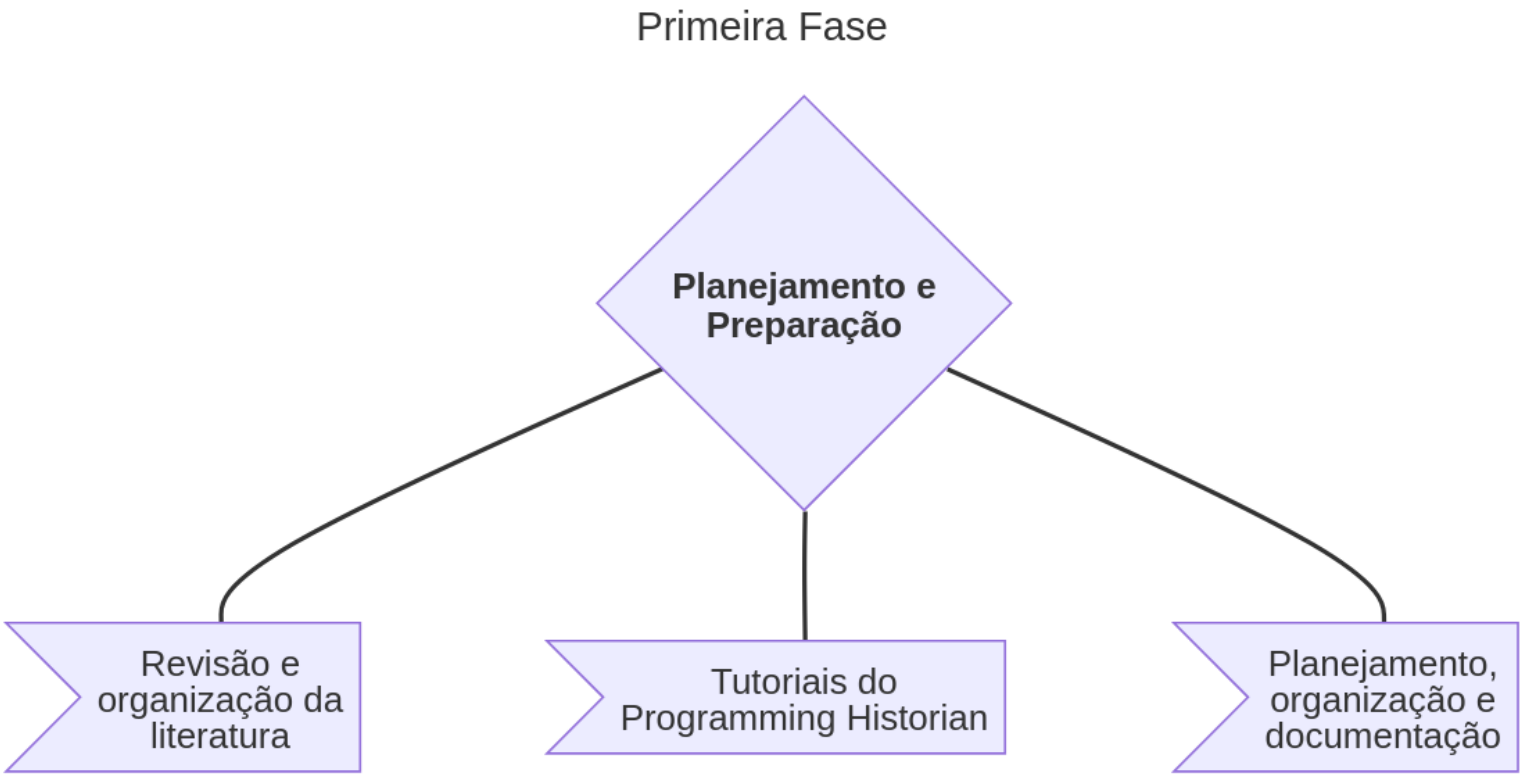


Figure 2: Phase 1 Diagram

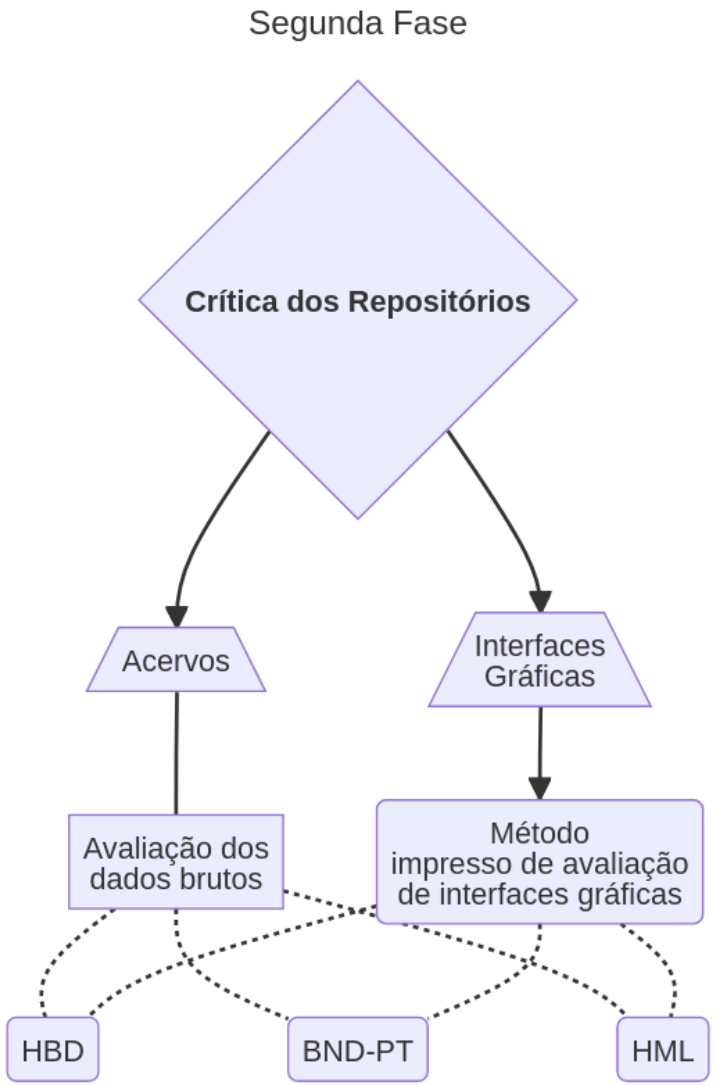


Figure 3: Phase 2 Diagram

Phase 3: Data Collection

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

Organization:

1. **Reports and Documentation of Searches:**
   - **Chapter 7:** Brief justification for the importance of consistent registration and methodological documentation strategies.
   - **Section 7.1:** Introduction and explanation of a methodological registration tool developed to assist research in digital repositories.
2. **Data Scrapers:**
   - **Chapter 8:** Presentation of the strategies and tools used for metadata and data collection in the three repositories utilized.
3. **Organization of Datasets:**
   - **Chapter 9:** Presentation of possibilities for the general organization of the collected data.

Phase 4: Data Processing

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

Organization:

1. **Testing and Evaluation:**
   - 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.
   - Evaluated open-source and free tools, and listed proprietary options useful for newspaper OCR.
2. **CLI Tools:**
   - **Command Line Interface Tools:** Tesseract, OCR-D, and Kraken.
   - Presented tutorials, documentation, tests, and results obtained for each tool.
3. **GUI Tools:**
   - **Graphical User Interface Tools:** gImageReader.
   - Presented tutorials, documentation, tests, and results obtained.
4. **Programming Historian Lessons:**
   - Listed relevant lessons for those seeking to perform OCR on newspapers.

Objective: - More than providing step-by-step usage guides, the goal is to present key features, advantages, and disadvantages of each tool. - Encourage critical reflection on the use of these tools. - Support researchers with knowledge and materials to consciously and critically develop their workflows.

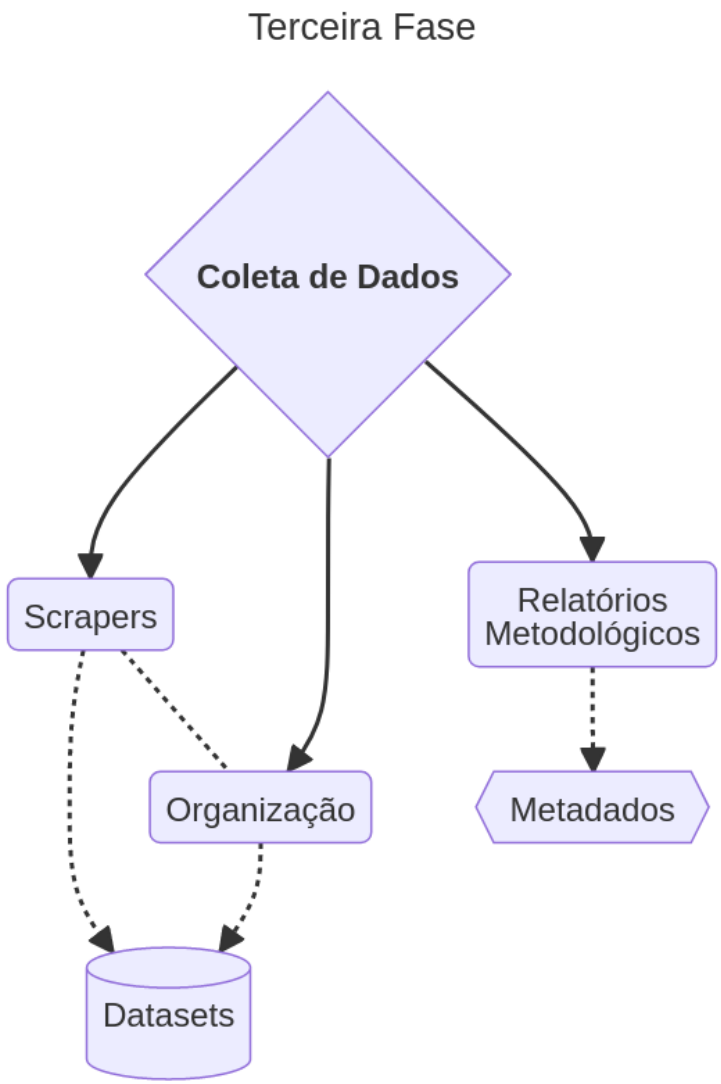


Figure 4: Phase 3 Diagram

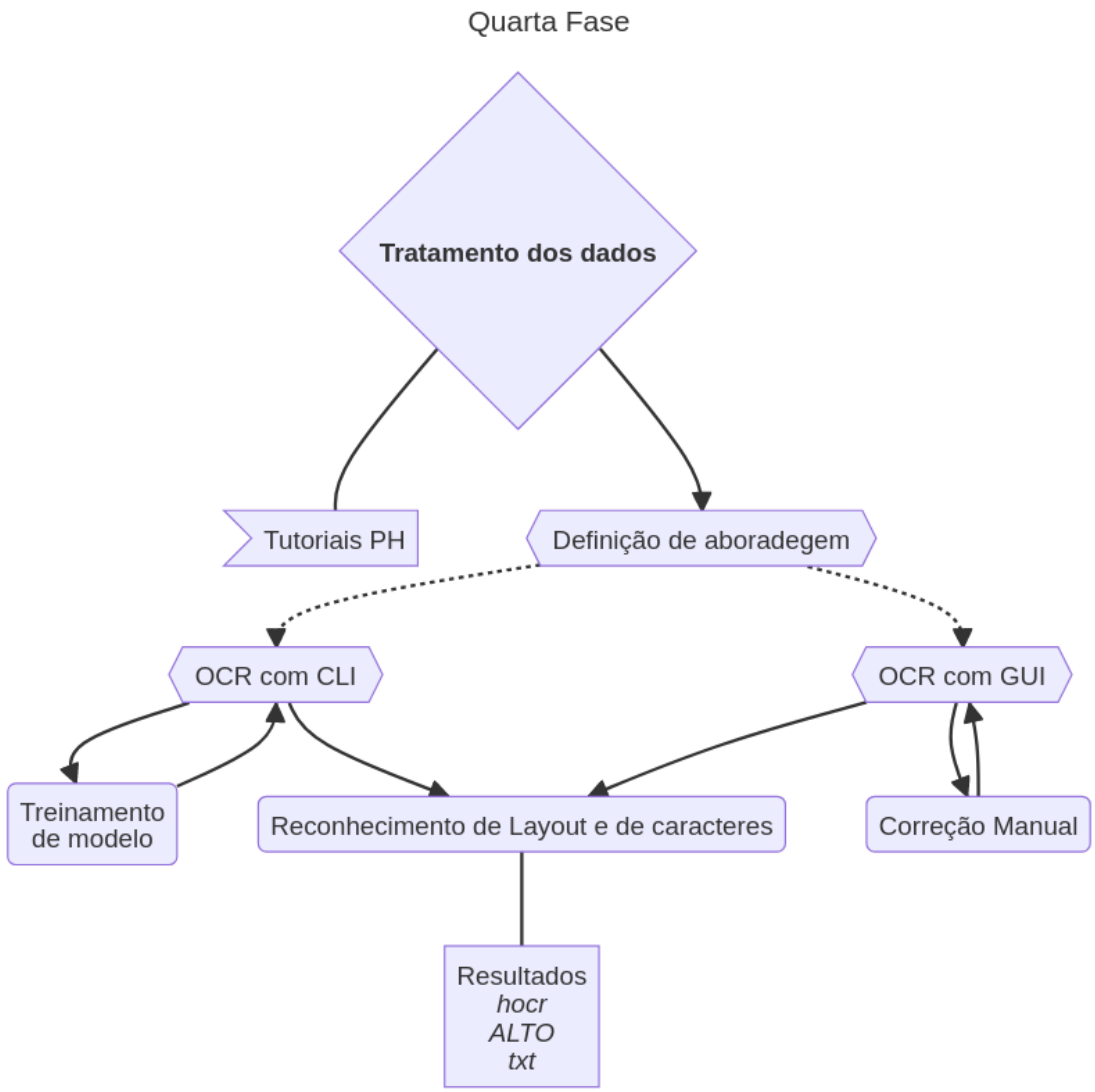


Figure 5: Phase 4 Diagram

Phase 5: Review, Preservation, and Publication

Summary: 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.

Key Points:

- **Review and Consolidation:**
  - Conduct a thorough review of the documentation and metadata.
  - Consolidate the data structure.
- **Continuous Process:**
  - Emphasize that this phase is a continuation and consolidation of a process that began in Phase 1.
  - Documentation and methodological recording should be carried out concurrently with data selection, collection, processing, and analysis.
  - This should be done transparently and in a standardized manner, adhering to clearly defined and publicized criteria.
- **Reference to Phase 1:**
  - Recommending revisiting Phase 1 for key aspects of documentation and methodological recording.
- **Chapters Outline:**
  - **Consolidation of Documentation:**
    - Ensure all documentation is comprehensive and well-organized.
  - **Review of Data and Metadata:**
    - Conduct a detailed review to ensure accuracy and completeness.
  - **Data Publication:**
    - Publish the research data, making it accessible for further research and analysis.

Detailed Examples

In this presentation, we will delve into two detailed examples to illustrate the application of the HemDig(pt) framework:

1. **Data Analysis of Newspapers from the National Digital Library of Portugal**
2. **Impresso Review Method for Interface Analysis**

These examples will highlight the practical implementation and the insights gained from using the framework.

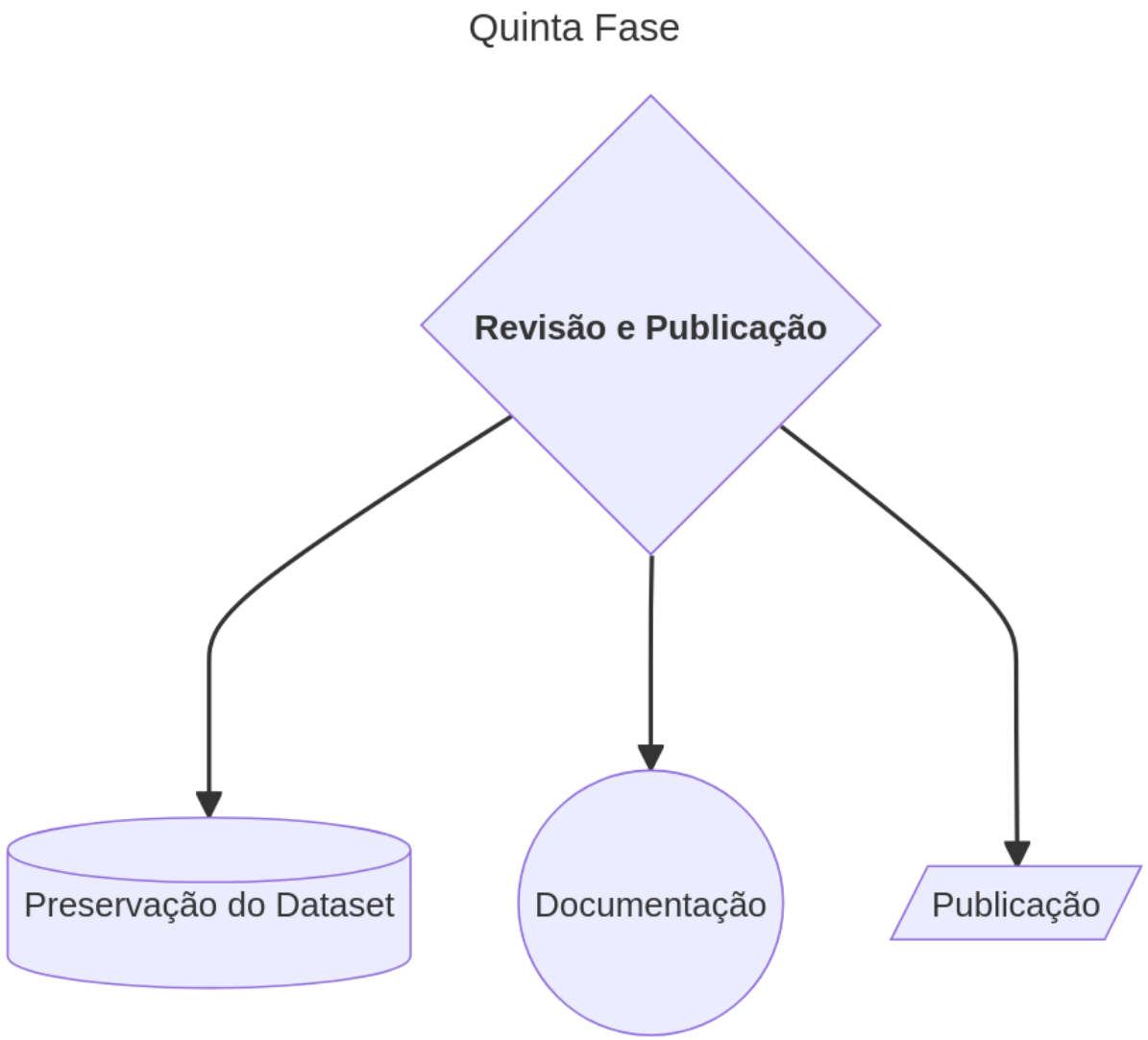


Figure 6: Phase 5 Diagram

Example 1: Data Analysis of Public Domain Newspapers from the National Digital Library of Portugal

Overview:

- **Objective:** Evaluate the quality and relevance of newspaper data from the National Digital Library of Portugal, focusing on public domain newspapers.
- **Data Collection:** Utilized scrapers to gather metadata and text data.

Analysis of Public Domain Newspapers:

- **Languages:**
  - Predominantly in Portuguese.
  - Some newspapers in other languages like French and English.
- **Publication Dates:**
  - Ranges from the 19th to early 20th century.
  - Earliest newspapers date back to the early 1800s.
  - Significant coverage during key historical periods, such as the Portuguese First Republic and World War I.

Example 2: Impresso Review Method for Interface Analysis

Overview:

- **Objective:** Assess the usability and effectiveness of graphical interfaces of digital newspaper repositories using the Impresso project's evaluation method.

Graphical Interface Analysis:

- **Criteria Evaluated:**
  - Browsing
  - Search
  - Result Sorting
  - Result Filtering
  - Result Display
  - Viewer
  - User Interaction
  - Enrichment
  - Info on Digitization
  - Connectivity
  - APIs

  - Newspaper Metadata

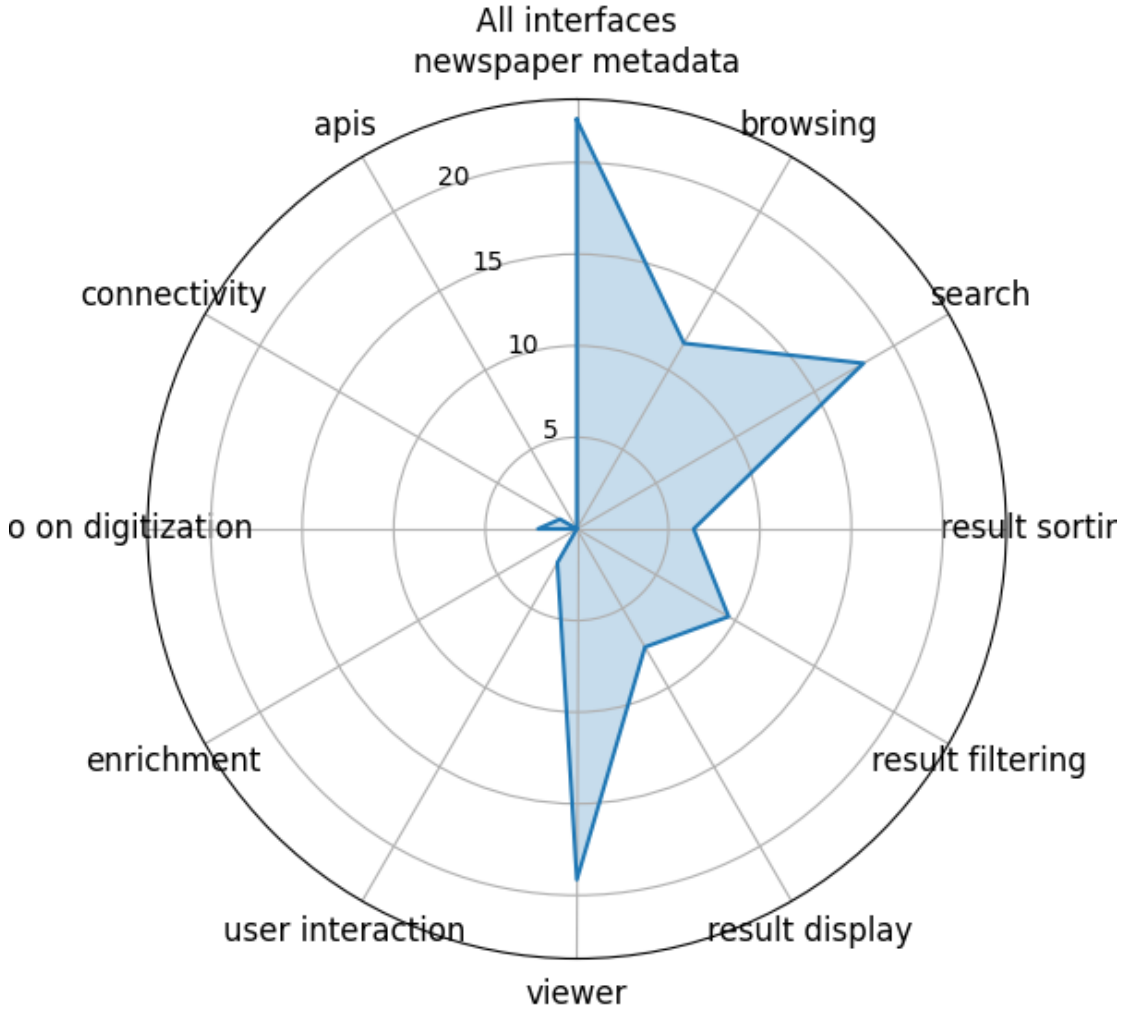
Key Insights from the Graph:

- **Strengths:**
  - **Browsing and Search:** High scores indicate effective navigation and search functionalities.
  - **Newspaper Metadata:** Well-structured and accessible metadata.
- **Weaknesses:**
  - **Result Display and Filtering:** Lower scores suggest the need for improvement in how search results are displayed and filtered.
  - **Viewer and User Interaction:** Indicates areas for enhancing the user interface for better interaction and viewing experience.

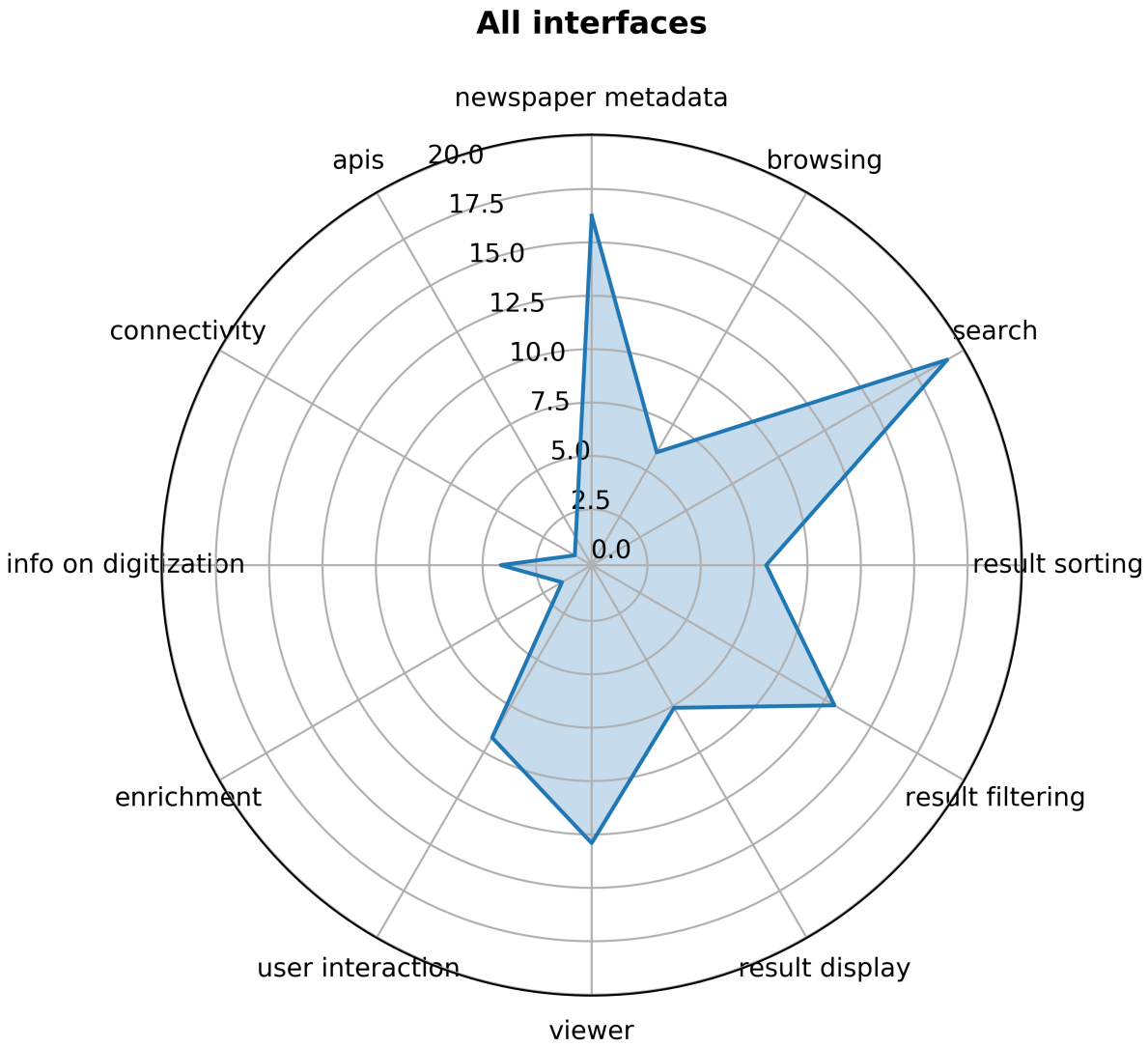
Comparison with Original Research (Ehrmann, Bunout, and Düring [2019]):

- **Similarities:**
  - Both studies highlight strong performance in browsing and search capabilities.
  - Metadata organization is a common strength.
- **Differences:**
  - The current study indicates weaker performance in result display and filtering compared to the original research.
  - Viewer and user interaction scores are notably lower in the current study, suggesting areas for improvement.

Visual Representations:



- **Current Study:**



- **Original Research:**

Conclusion

- **Deep Dive into Digitized Newspaper Data:**
  - Analyzed key collections from the National Libraries of Brazil and Portugal.
  - Generated visualizations and organized data on digitized newspapers.
- **Impactful Findings:**
  - Enabled numerous potential future research and publications.
  - Provided a comprehensive analysis of graphical interface characteristics.
  - Offered robust insights on the epistemological and methodological impacts of these technical choices on research.
- **Open and Accessible Data:**
  - Organized, documented, and made available in an open, multiplatform format.
  - Data can be accessed, reused, expanded, and critiqued.
- **OCR Tools and Methodologies:**
  - Conducted tests and studies on OCR tools, offering implementation guidance.
  - Created tutorials and supported materials for tools like gImageReader, OCR-D, Kraken, Zotero, and Tropy.
- **Scripts and Methodological Tools:**
  - Developed scripts for data extraction from graphical interfaces, enabling new visualizations and analyses.
  - Created a methodological support tool for generating research reports.
- **Educational Framework:**
  - All data, results, recommendations, and reflections were systematically organized into a framework.
  - Published online as a JupyterBook under a Creative Commons license.
- **Future Potential:**
  - Significant potential for advancing epistemological reflections in the field of history.
  - Emphasizes the growing relationship between humanities and technology.