Relatório Final da Pesquisa

HEMDIG(pt) Framework - Métodos, ferramentas e hemerotecas digitais em português

Eric Brasil

2023-10-11

Conteúdo

0.1 Introdução 2
0.2 Cronograma 3
0.3 Metodologia 4
0.3.1 Git e Github 4
0.3.2 Etapas da pesquisa 5
0.4 Resultados 6
0.4.1 Dados dos acervos 6
0.4.2 Dados das interfaces 7
0.4.3 Gráficos, visualizações e imagens 7
0.4.4 Imagens utilizadas no livro 8
0.4.5 Ferramentas 8
0.4.6 Tutoriais e testes de OCR 8

0.1 Introdução

Nessa pesquisa, realizada em estágio de pós-doutoramento junto ao Instituto de história Contemporânea da Universidade Nova de Lisboa, busquei enfrentar o problema comum a todos/as que executam a operação historiográfica: o processo de selecionar, recolher, organizar as fontes primárias e realizar sua crítica – a chamada heurística das fontes. Agora, com um aspecto diferente, uma heurística em ambientes digitais, através de ferramentas e dados digitais. O foco aqui está na análise de repositórios e interfaces gráficas que permitem o acesso a periódicos em língua portuguesa digitalizados.

Tomo como objeto de estudo a Hemeroteca Digital Brasileira, da Fundação Biblioteca Nacional, a coleção de periódicos da Biblioteca Nacional Digital de Portugal, e a Hemeroteca Digital da Hemeroteca Municipal de Lisboa. Analiso tanto as interfaces gráficas e possibilidades de acesso aos dados quanto seus acervos disponibilizados. Busco subsidiar reflexões sobre as relações entre teoria, metodologia e epistemologia da História. Essa tarefa foi realizada através da aproximação entre "os saberes fundamentais da pesquisa em história com conhecimentos técnicos de programação, atuando em zonas de troca, interdisciplinares e colaborativas", avançado em pesquisas que venho desenvolvendo e publicando nos anos recentes. Na prática, busquei exercer uma hermenêutica digital, como definidas por Fickers e Clavert:

> Tornar explícito como a produção de conhecimento histórico através de ferramentas e tecnologias digitais é o resultado de um processo complexo de interação humano-máquina, de co-construção do 'objeto epistêmico' da inquirição e investigação histórica. (FICKERS, CLAVERT, 2021, parágrafo 6, tradução minha)

Como afirmei anteriormente,

> as novas formas de realizar pesquisas em repositórios e interfaces digitais de buscas impactam, mediam e direcionam tanto a coleta e seleção das fontes quanto sua análise. Diante disso, é fundamental que o método histórico leve em consideração a aplicação de práticas de heurística digital coerentes tanto com as características das ferramentas e métodos utilizados, das fontes e dados trabalhados quanto com as reflexões teóricas básicas da disciplina histórica. (BRASIL, 2022, p. 189)

A partir dessas análises e reflexões, produzi um *framework* para pesquisa nos referidos repositórios, buscando encadear um fluxo de trabalho que atenda para os aspectos metodológicos específicos da disciplina História, contemplando desde a preparação computacional inicial, criação de um plano de gerenciamento dos dados, a pesquisa e análise dos acervos e interfaces gráficas, a coleta e organização dos dados, o tratamento e preservação dos dados.

O *HEMDIG(pt) framework* é um enquadramento dos processos de pesquisa composto por uma biblioteca de referências bibliográficas sobre o temas – seus aspectos técnicos e teóricos –, a lista de repositórios

de periódicos históricos em português digitalizados, uma documentação acerca de cada um dos repositórios e suas interfaces gráficas (apresentando as principais características, parâmetros possíveis de busca, opções de acesso e resultados); um conjunto de *Jupyter Notebooks* com ferramentas para registro metodológico das pesquisas e resultados, e ferramentas de coletas dos dados; indicação e manual de uso de programas de Reconhecimento Ótico de Caracteres (OCR) e análise de layout (OLR); indicações de estratégias e ferramentas para preservar, documentar e compartilhar os dados e resultados da pesquisa.

Tudo isso está organizado e disponível online em um livro digital, publicado em formato de Jupyter Book, disponível em https://ericbrasiln.github.io/hemdig-framework/, com licença Creative Commons Atribuição-NãoComercial-CompartilhaIgual 4.0 Internacional.

0.2 Cronograma

Atividades	10/22	11/22	12/22	01/23	02/23	03/23	04/23	05/23	06/23	07/23	08/23	09/23	10/23
Levantamento Bibliográfico e organização de biblioteca de referências no Zotero	X	X											
Mapeamento dos repositórios e interfaces gráficas		X	X										
Curso de Python Avançado				X	X	X	X						
Criação do formulário com os critérios de avaliação				X			X						
Pesquisa e avaliação de ferramentas de código aberto para OCR e OLR						X		X	X	X			
Criação de jupyter notebooks com passo a passo para criação de relatórios de registro metodológico						X	X						
Avaliação das interfaces e repositórios							X	X	X				
Workshop para o LAB_HD								X					

Atividades	10/22	11/22	12/22	01/23	02/23	03/23	04/23	05/23	06/23	07/23	08/23	09/23	10/23
Criação de jupyter notebooks com scripts para organização e tabulação de dados e metadados									X	X			
Escrita de artigo											X	X	X
Organização do Workflow e finalização do projeto											X	X	X

0.3 Metodologia

A pesquisa foi desenvolvida no Laboratório de Humanidades Digitais da Universidade Nova de Lisboa, e contou com a supervisão do professor Daniel Alves. A pesquisa foi organizada conforme o cronograma apresentado acima e suas etapas serão descritas a seguir. Mas antes, é importante ressalte que toda a pesquisa foi realizada utilizando ferramentas e programas de código aberto e livres e todos os dados, resultados e produtos foram disponibilizados online, com licença Creative Commons Atribuição-NãoComercial-CompartilhaIgual 4.0 Internacional.

0.3.1 Git e Github

Além disso todas as etapas de realização da pesquisa foram controladas e registradas através do sistema de controle de versões Git e utilizei o Github como repositório remoto para armazenamento do histórico de alterações e compartilhamento dos dados e resultados. É possível acessar todos os dados de alterações da pesquisa através do arquivo log_main.csv.

Para gerenciar as tarefas e atividades da pesquisa, utilizei a funcionalidade Projects do Github, criando quadros Kanban para organizar as tarefas em listas de pendências, em andamento e concluídas. É possível acessar a página da pesquisa através do link https://github.com/users/ericbrasiln/projects/6. A utilização do Github para gerenciar as tarefas da pesquisa possibilitou o registro de todas as atividades, o acompanhamento do progresso ao longo do tempo e visualizações gráficas do andamento da pesquisa. Foram criados *milestones* para etapas chave da pesquisa, que reuniram conjuntos de *issues* e *pull requests* relacionados a cada etapa. É possível acessar a lista de *milestones* aqui e todas as *issues* aqui. Para melhor organização das *issues*, foram criados *labels* para identificar as tarefas relacionadas a cada etapa da pesquisa. É possível acessar a lista de *labels* aqui.

Nesse repositório remoto no Github, https://github.com/ericbrasiln/hemdig-framework, portanto, estão disponíveis todos os dados, o histórico de mudanças e também a organização geral da pesquisa.

Nele também está armazenado o código fonte do Jupyter Book, que foi utilizado para publicar o livro digital da pesquisa, disponível em https://ericbrasiln.github.io/hemdig-framework/. Os arquivos necessários para publicação do livro no Github Pages estão disponíveis no branch `gh-pages` e o código fonte do Jupyter Book está disponível no branch `main`, na pasta `book/`.

0.3.2 Etapas da pesquisa

A) Organização geral da pesquisa: nessa etapa foi criada a estrutura geral da pesquisa e o plano de gerenciamento de dados, assim como a listagem de repositórios e ferramentas a serem utilizadas. Ver o *milestone* Organização geral do repo.

B) Levantamento bibliográfico: aqui foi desenvolvida a estratégia de levantamento bibliográfico,sua implementação, organização e tratamento dos dados. Também foi realizada a inserção dos dados na biblioteca de referências bibliográficas. Ver o *milestone* Levantamento bibliográfico.

C) Análise dos acervos dos repositórios: nessa etapa, foi empreendida uma detalhada análise dos conjuntos de dados disponibilizados nas plataformas dos acervos digitais. Ver o *milestone* Repositórios - Análise dos Dados.

D) Análise das interfaces gráficas dos acervos: a partir do método *impresso Review*, analisei amplamente as interfaces gráficas dos repositórios, buscando identificar as principais características, parâmetros possíveis de busca, opções de acesso e resultados. Ver o *milestone* Repositórios e Interfaces Gráficas.

E) Ferramentas de suporte metodológico: criação e teste de ferramentas de suporte metodológico. Foram criadas ferramentas de raspagem e ferramentas de geração de relatórios de pesquisas. Ver o *milestone* Ferramentas de suporte metodológico

F) Análise e teste de ferramentas de linha de comando para OCR: nessa etapa foram testadas e avaliadas ferramentas de reconhecimento ótico de caracteres (OCR) de linha de comando. Ver o *milestone* Análise de OCR e OLR.

G) Análise e teste de programa com interface gráfica para OCR: teste e avaliação do gImageReader, um programa com interface gráfica para reconhecimento ótico de caracteres. Ver o *milestone* GUI para OCR.

H) Revisão final e publicação: nessa etapa foram realizadas as revisões finais do livro e do repositório, e a publicação do livro digital. Ver o *milestone* Publicação.

0.4 Resultados

Todos os resultados da pesquisa estão organizados, documentados e disponibilizados online. O livro digital da pesquisa está disponível em https://ericbrasiln.github.io/hemdig-framework/. O código fonte do livro está disponível no repositório remoto no Github, https://github.com/ericbrasiln/hemdig-framework/tree/main/book.

O *HEMDIG(pt) framework* representa o resultado mais explícito da pesquisa e reúne um conjunto de dados, tutoriais, indicações e reflexões que podem subsidiar inúmeros pesquisa futuras. As análises dos dados dos acervos da Hemeroteca Digital Brasileira e da coleção de periódicos da Biblioteca Digital Nacional de Portugal, assim como as analises das suas respectivas interfaces gráficas, abrem possibilidades muito amplas para novas pesquisas e publicações.

Além disso, os testes e comparações entre diferentes ferramentas de OCR, tanto de linhas de comando quanto de interface gráfica, geraram resultados substanciais para a elaboração de novas publicações.

O *HEMFIG(pt)* também oferece ferramentas de coleta de dados e geradores de relatórios de busca que favorecerão a realização de pesquisa mais eficientes e com maior qualidade metodológica.

Por fim, a pesquisa também gerou uma série de indicações sobre registro e preservação dos dados assim como a criação de uma bibliografia especializada bastante atualizada e bem organizada para acesso público.

Esse repositório foi armazenado no Zenodo e um DOI foi gerado: 10.5281/zenodo.8397782.

Assim, o produto final da pesquisa se mostra rico para todos e todas que tem interesse em pesquisar jornais históricos digitalizadas em língua portuguesa. Mas não apenas isso, pois possibilita reflexões sobre o próprio caráter da pesquisa histórica em ambientes digitais, suas possibilidades e limites.

Nos aspectos profissionais, esse período de estágio de pós-doutoramento, possibilitou a ampliação de um rede de colaborações no campo d história Digital, permitiu que eu avançasse na formação técnica, sobretudo com a linguagem de programação Python; realizasse eventos, palestra e oficinas e gerasse novas publicações.

0.4.1 Dados dos acervos

Lista com os dados dos acervos utilizados no projeto (com links para os diretórios).

0.4.1.1 Hemeroteca Digital Brasileira

- XML dos dados brutos
- CSV com dados de periódicos digitalizados em 2019 - também disponível na página da BND-BR.

- CSV com dados de periódicos digitalizados em 2020 - também disponível na página da BND-BR.
- CSV dos dados filtrados, contendo as colunas “title”, “subtitle”, “place”, “period”, “publisher”, “periodicity”, “language”
- CSV do dataframe final tratado
- CSV de periódicos por décadas

0.4.1.2 Biblioteca Nacional Digital de Portugal

- CSV bruto com dados gerais - também disponível na página da BND-PT.
- CSV bruto apenas com dados de obras em domínio público - também disponível na página da BND-PT.

0.4.1.3 Hemeroteca Digital da Hemeroteca Municipal de Lisboa

- CSV bruto com os dados coletados da HML
- CSV com dados de localidades bruto
- CSV com dados de localidades tratado

0.4.2 Dados das interfaces

Lista com os dados das interfaces dos acervos utilizados no projeto (com links para os diretórios).

- Formulário geral de dados básicos
- Formulário de revisão das interfaces - modelo *impresso Review*
- Formulário de mapeamento geral com base dos critérios de alto nível - modelo *impresso Review*

0.4.3 Gráficos, visualizações e imagens

Lista dos gráficos e visualizações gerados no projeto.

0.4.3.1 Hemeroteca Digital Brasileira Acesse aqui o diretório

0.4.3.2 Biblioteca Nacional Digital de Portugal Acesse aqui o diretório.

0.4.3.3 Visualizações das análises das interfaces gráficas - método *impresso Review* Acesse aqui o diretório

0.4.3.4 Imagens utilizadas nos testes de OCR Acesse aqui o diretório

0.4.4 Imagens utilizadas no livro

0.4.4.1 Diagramas

- Diagrama geral
- Diagrama da 1ª fase
- Diagrama da 2ª fase
- Diagrama da 3ª fase
- Diagrama da 4ª fase
- Diagrama da 5ª fase

0.4.4.2 Logos

- PNG, Grande, fundo branco
- PNG, Grande, fundo preto
- PNG, Pequeno, fundo branco
- PNG, Pequeno, fundo preto
- SVG, Grande, fundo transparente
- SVG, Pequeno, fundo transparente

0.4.4.3 Outras imagens Acesse aqui o diretório

0.4.5 Ferramentas

- Relatórios Metodológicos para Pesquisa em Interfaces Gráficas de Periódicos Digitalizados
- Scraper para a Biblioteca Digital de Portugal..
- Scraper para o site da Hemeroteca Digital de Lisboa.
- pyHDB - Coleta de metadados e acervos da HDB.

0.4.6 Tutoriais e testes de OCR

- OCR-D.
- Kraken.
- gImageReader.
- Tutoriais do Programming Historian.