HEMDIG(pt): construindo um framework para hemerotecas digitais em português

Eric Brasil

24 de abril de 2024

Resumo: O presente trabalho tem como objetivo apresentar e discutir o *HEMDIG(pt)*, um *framework* para coleta, organização e tratamento de acervos digitais de periódicos em língua portuguesa. Concebido como parte de uma pesquisa de pós-doutorado no Instituto de História Contemporânea da Universidade Nova de Lisboa, o *HEMDIG(pt)* busca facilitar o acesso a essas fontes históricas, promovendo, ao mesmo tempo, uma abordagem metodológica rigorosa e transparente na pesquisa digital em humanidades. A pesquisa está estruturada em cinco fases distintas. A fase 1, *Planejamento e Preparação*, foca no levantamento bibliográfico, preparação técnica básica e definição de licenças, estratégias de documentação e organização. A segunda fase, *Crítica dos Acervos*, analisa as características e escopo dos acervos digitais e suas interfaces gráficas, focando na estrutura e acessibilidade dos dados. A fase seguinte, *Coleta de Dados*, utiliza técnicas de web scraping e ferramentas específicas para extrair e organizar metadados e conteúdo textual dos repositórios. A quarta fase, *Tratamento dos Dados*, apresenta tutoriais, documentação e ferramentas para aplicação de técnicas de Reconhecimento Ótico de Caracteres (OCR). A última fase, *Revisão, Preservação e Publicação*, consiste na revisão da documentação e dos dados para publicação, garantindo a preservação digital e a acessibilidade a longo prazo. Tais fases estruturam o processo e fundamentam os resultados alcançados, demonstrando como uma abordagem metódica potencializa a qualidade e acessibilidade dos acervos digitais. Os resultados do *HEMDIG(pt)* incluem não apenas o *framework* funcional para hemerotecas digitais e ferramentas específicas para acervos em português, mas também reflexões críticas sobre o processo de digitalização e análise de documentos históricos. Essas reflexões abordam os desafios técnicos, como a qualidade da digitalização e seus impactos nos resultados do OCR, as características das interfaces gráficas e destacam a necessidade de investimento contínuo em tecnologias adaptadas às especificidades dos textos históricos em língua portuguesa. Todas as etapas foram organizadas em um *JupyterBook*, que documenta todo o processo e serve como um recurso educacional e metodológico para outros pesquisadores e está disponível online sob licença *Creative Commons BY-NC-SA 4.0*. Por conseguinte, o *HEMDIG(pt)* não só otimiza o acesso a fontes digitais importantes, mas também contribui significativamente para o campo

da história digital, oferecendo métodos e ferramentas que podem ser adaptadas para outros contextos de pesquisa histórica e digital e ressaltando a importância de uma abordagem metodológica rigorosa e transparente na pesquisa digital em humanidades.