ABHD
Revista Brasileira de Humanidades Digitais RBHD

Uma análise do uso das práticas de crowdsourcing em projetos de transcrição

Ana Ligia Medeiros, Luziana Jordão Lessa Trézze, Vitor Silveira Pereira

A pesquisa em acervos manuscritos, além de relevante para pesquisadores e para a construção da memória, se torna custosa, árdua e morosa, tanto para o usuário quanto para as instituições, visto o tempo despendido com os recursos suscitados pela dificuldade na legibilidade do texto, a condição física do material, o tipo de letra e o conhecimento das técnicas de transcrição. Biasi (2010), considera que a seleção e a organização de documentos manuscritos contribuem consideravelmente com as pesquisas futuras, porém muitos pesquisadores ficam desestimulados com o tamanho e a dificuldade da incumbência de organização, classificação e decifração dos manuscritos custodiados em acervos brasileiros e estrangeiros.

Como solução a este problema, inicialmente, as instituições responsáveis pela preservação do patrimônio e da memória coletiva empenharam seus esforços na digitalização de acervos para preservar, disseminar e facilitar o acesso aos usuários. Lévy (2007) afirma que a principal razão para a digitalização é que a mesma permite um tipo de tratamento de informações eficaz e complexo, impossível de ser executado por outras vias.

Todavia, percebeu-se a necessidade de utilizar novos aparatos tecnológicos no sentido de reduzir custos, bem como possibilitar a melhoria do uso das fontes de informação para a pesquisa, em especial dos acervos manuscritos. Nesse contexto, o conceito das Humanidades Digitais (HDs) tem sido aplicado pelas instituições por meio de projetos inovadores de crowdsourcing, como por exemplo, para a transcrição dos mesmos.

No crowdsourcing, termo desenvolvido em 2006 pelo jornalista e pesquisador estadunidense Jeff Howe, são desenvolvidas iniciativas colaborativas, por uma gama de voluntários que se reúnem online, em torno da solução de problemas, tendo em vista construir ideias e encontrar estratégias a baixo custo.

> O crowdsourcing capitaliza a profunda natureza social da espécie humana. Ao contrário da visão distópica e agourenta de que a internet serve principalmente para isolar as pessoas, o crowdsourcing usa a tecnologia para

> incentivar níveis inéditos de colaboração e trocas significativas entre pessoas com as formações mais diversas, das mais distantes localizações geográficas. (HOWE, 2009)

Os projetos de crowdsourcing são, então, uma forma da comunidade se engajar ativamente em algum projeto de forma voluntária e coletiva.

> No âmbito das ciências sociais é importante determinar quais as motivações dos indivíduos quando se envolvem nos projetos. Quando falamos de uma atividade remunerada, as razões do envolvimento dos colaboradores são razões econômicas, mas nas atividades de aspecto social e cultural, constata-se que a correlação entre o envolvimento dos indivíduos e os projetos está relacionado com uma tradição consolidada de voluntariado e com motivações de carácter emocional. (VEIGA, 2015)

Normalmente o fluxo de trabalho, tratando-se de manuscritos e outros tipos de acervos semelhantes, especialistas, ou até mesmo membros da comunidade fazem uma primeira interpretação do texto, posteriormente encaminha-se para uma revisão e por fim, a validação do texto.

A seguir apresentaremos 5 projetos diferentes entre si indicando se o mesmo dispõe de software gratuito, se está em código aberto e se existe algum tutorial de utilização.

FromThePage é uma plataforma proprietária com valores entre $80 e $300 mensais, onde a instituição pode fazer o upload de documentos em PDF, em formato de imagem ou então utilizar páginas de publicação como Omeka e Internet Archive. O seu projeto de transcrição pode ser público para a comunidade FromThePage ou então para uma rede de colaboradores privada. No momento de escrita deste resumo eles contabilizavam 729.584 páginas transcritas.

Library of Congress by the People, como o nome indica, é uma iniciativa da biblioteca do congresso estadunidense. A aplicação roda em Concordia, um software em código aberto desenvolvido pela Library of Congress para impulsionar projetos de transcrição via crowdsourcing e está disponível gratuitamente no repositório Github.

Scribe Framework é uma plataforma de código aberto altamente configurável para projetos de transcrição em crowdsourcing. Assim como a Library of Congress by the People, o Scribe está disponível gratuitamente no Github.

Transkibus é uma iniciativa mantida pelo grupo de Digitalização e Preservação Digital Universidade de Innsbruck, financiada atualmente pela Horizon 2020 Research and Innovation da União Europeia. O software é gratuito e na sua maior parte é de código aberto, disponível no Github. O Transkribus oferece várias ferramentas para o processamento automatizado de documentos, como: Reconhecimento de texto manuscrito (HTR) e análise de layout, reconhecimento óptico de caracteres (OCR).

What's on the Menu? é uma iniciativa da Biblioteca Pública de Nova York (NYPL) com aproximadamente 45.000 menus que datam da década de 1840 até o presente, a coleção de menus de restaurantes da Biblioteca Pública de Nova York é uma das maiores do mundo, usada por historiadores, chefs, romancistas e entusiastas da comida. Eles utilizam uma API (conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na web) de código aberto desenvolvida pela NYPL e disponível no Github, mas o usuário interessado em fazer uso pode enviar um email com o assunto API ACCESS e uma descrição do seu projeto para a equipe do projeto.

Quadro 1: Comparação entre projetos de transcrição

Nome do projeto	Software livre	Código aberto	Tutorial
FromThePage	não	não	sim
Library of Congress by the People	sim	sim	sim
Scribe Framework	sim	sim	sim
Transkribus	sim	em parte	sim
What's on the menu?	sim	sim	sim

Implantado em 2019 através da portaria nº 73, de 12 de julho de 2018 da Fundação Casa de Rui Barbosa / Ministério da Cultura, o Laboratório de Humanidade Digitais da FCRB (LabHD) tem como objetivo "atender às necessidades de acesso, preservação, disseminação, recuperação da informação, inovação e criação de novos conhecimentos a partir do fomento e da aplicação de métodos digitais de pesquisa ao campo das Humanidades". Por conseguinte, empenhou-se neste projeto, que demandou um estudo

prévio com o objetivo de examinar as iniciativas existentes em âmbito nacional e internacional (crowdsourcing transcription projects) de transcrição de manuscritos, objetivo dessa explanação.

Sob guarda do Centro de Memória e Informação (CMI), da Fundação Casa de Rui Barbosa (FCRB), encontram-se diversos documentos de natureza distintas, como os arquivos pessoais de interesse histórico e os arquivos pessoais de escritores brasileiros, que formam o ArquivoMuseu de Literatura Brasileira (AMLB). Entre o acervo do AMLB citamos o Arquivo Cruz e Sousa, o Arquivo Gonzaga Duque, o Arquivo Heitor Modesto, o Arquivo Lucio de Mendonça, o Arquivo Machado de Assis, o Arquivo Nestor Vitor, e o Arquivo Salvador de Mendonça, todos já digitalizados e de autores que estão em domínio público.

> Fundado em 1972, o Arquivo-Museu de Literatura Brasileira (AMLB) foi instalado na Fundação Casa de Rui Barbosa (FCRB) após um apelo feito por Carlos Drummond de Andrade em sua coluna do Jornal do Brasil de 11 de julho daquele ano. A crônica de Drummond era um lamento pela falta de um local onde ficasse guardada a memória literária nacional através das grandes obras da literatura do país. Em algumas linhas, o escritor fez também uma espécie de apelo para a criação de tal instituição: “mas falta o órgão especializado, o museu vivo que preserve a tradição escrita brasileira, constante não só de papéis como de objetos relacionados com a criação e a vida dos escritores [...] Será que a ficção, a poesia e o ensaio de nossos escritores não merecem possuí-lo?”.

Espera-se como resultado dessa análise, idealizar o projeto de transcrição de manuscritos na FCRB, bem como contribuir com as discussões em HDs, permitindo que as novas demandas sejam atendidas por meio de trabalhos colaborativos que visem compartilhar projetos e futuras melhorias desses aparatos tecnológicos. Por fim, acreditamos que as iniciativas em HDs e os trabalhos colaborativos facilitam o trabalho de transcrição dos documentos, agilizando tanto o processo técnico de tratamento quanto à pesquisa, permitindo, cada vez mais, que tesouros, até então escondidos, sejam conhecidos pela sociedade e gerem novas pesquisas.

Referências

BIASI, Pierre-Marc de. **A genética dos textos**. Tradução de Marie-Hélène Paret Passos. Porto Alegre: EDIPUCRS, 2010.

FROM THE PAGE. **FromThePage**. Disponível em: https://fromthepage.com/. Acesso em: 20 jun. 2020.

FUNDAÇÃO CASA DE RUI BARBOSA. Conheça mais sobre o Arquivo-Museu de Literatura Brasileira. Disponível em: http://www.casaruibarbosa.gov.br/interna.php?ID_S=9&ID_M=4416. Acesso em: 20 jun. 2020.

FUNDAÇÃO CASA DE RUI BARBOSA. **Portaria nº 73, de 12 de julho de 2018**. Institui o Laboratório de Humanidades Digitais (LabHD) da Fundação Casa de Rui Barbosa. 2018.

HOWE, Jeff. **O poder das multidões**: por que a força da coletividade está remodelando o futuro dos negócios. Rio de Janeiro: Elsevier, 2009.

LÉVY, Pierre. **Cibercultura**. São Paulo: Editora 34, 2007.

LIBRARY OF CONGRESS. **Library of Congress By the People**. Disponível em: https://crowd.loc.gov. Acesso em: 20 jun. 2020.

NEW YORK PUBLIC LIBRARY. **What's on the menu?**. Disponível em: http://menus.nypl.org. Acesso em: 20 jun. 2020.

NEW YORK PUBLIC LIBRARY LABS; ZOONIVERSE. **Scribe Framework**. Disponível em: https://scribeproject.github.io. Acesso em: 20 jun. 2020.

UNIVERSITY OF INNSBRUCK. **Transkribus**. Disponível em: https://transkribus.eu/. Acesso em: 20 jun. 2020.

VEIGA, Maria Alexandra de Figueiredo Araújo Leça da. **O recurso ao crowdsourcing como modelo válido para a recuperação da informação e construção de memória colectiva:** o projecto Memórias da I Guerra Mundial 1914-1918, Os Dias da Memória. 116 f. Dissertação (Mestrado) - Curso de Mestrado em Ciências da Informação e Documentação, Especialização em Arquivística, Universidade Nova de Lisboa, Lisboa, 2015. Disponível em: http://hdl.handle.net/10362/18264. Acesso em: 20 jun. 2020.