Curadoria de conteúdo agrícola disponível em livros digitais

Felipe Ramos de Campos[1,2], Maria Eduarda Nogueira Rodrigues [1,3], Glauber José Vaz[1]

[1]Embrapa Agricultura Digital, Campinas, SP – Brazil

[2]Graduando em Engenharia Agrícola, Unicamp, Campinas, SP – Brazil

[3]Graduanda em Ciências Sociais, Unicamp, Campinas, SP – Brazil

{felipe.campos,maria.rodrigues}@colaborador.embrapa.br,
glauber.vaz@embrapa.br

Abstract. *This article shows the importance of digital curation in making open data available and in the dissemination of technical information, dealing especially with the analysis and editing phase of digital books content. The text treatment carried out in the books of the 500 Questions 500 Answers collection for their availability in data repositories illustrates how human work is essential in digital curation in order to achieve better quality of data that are the foundation of digital solutions.*

Resumo. *Este artigo mostra a importância da curadoria digital na disponibilização de dados abertos e na disseminação de informações técnicas tratando especialmente da fase de análise e edição do conteúdo de livros digitais. O tratamento de texto realizado nas obras da coleção 500 Perguntas 500 Respostas para sua disponibilização em repositórios de dados ilustra como o trabalho humano é essencial na curadoria digital a fim de se alcançar maior qualidade de dados que são a base de soluções digitais.*

1. Introdução

A coleção “500 Perguntas 500 Respostas: o produtor pergunta a Embrapa responde” é uma série de publicações da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) voltada para a disseminação de informações técnicas relacionadas à agricultura e pecuária, e conta com dezenas de livros organizados na forma de perguntas e respostas.

Com as tecnologias digitais, a difusão de conhecimento tem potencial de alcance muito maior, e auxilia na solução de uma vasta gama de necessidades sociais [Serpa and Ferreira 2019]. Da integração entre o ciberespaço e o espaço físico, buscam-se soluções para as necessidades humanas em equilíbrio com os avanços econômicos na chamada

Sociedade 5.0 [Japan Cabinet Office 2023].

Muitas das tecnologias digitais dependem do acesso a uma quantidade razoável de dados com boa qualidade. Uma maneira de se obter dados confiáveis é por meio de plataformas construídas para disponibilizá-los de forma aberta, como, por exemplo, o Repositório de Dados de Pesquisa da Embrapa (Redape) [Embrapa 2023], que facilita a busca por dados produzidos pela empresa.

Gerar dados com boa qualidade requer atividades de curadoria, envolvendo uma série de tarefas que visam garantir, além da qualidade, a acessibilidade e a disponibilidade contínua dos ativos digitais [Rusbridge et al. 2005]. Segundo Abbott (2008), a curadoria digital relaciona-se à prática de gestão de dados e à garantia de que esses dados permanecerão disponíveis no futuro. A curadoria empregada neste trabalho requer capacidades analíticas inerentemente humanas, como interpretação textual e análise de figuras.

O objetivo deste trabalho é mostrar a importância da curadoria na disponibilização de ativos digitais em plataformas de dados abertos, considerando a oferta no Redape do conteúdo dos livros da coleção 500 Perguntas 500 Respostas da Embrapa. Essa iniciativa visa facilitar a disseminação desse conteúdo e garantir um nível de qualidade de excelência no apoio ao desenvolvimento de tecnologias baseadas nesses dados.

2. Metodologia

Este trabalho faz uso da metodologia apresentada por Vaz et al. (2023), que foi elaborada para o tratamento dos livros digitais da coleção 500 Perguntas 500 Respostas a fim de indexar seu conteúdo e possibilitar seu acesso por meio de um mecanismo de busca. Essa metodologia envolve quatro etapas: a extração de elementos essenciais das obras, como as imagens e os arquivos HTML com o texto das perguntas e das respostas, o pré-processamento desses arquivos para a automatização de algumas tarefas, a análise e a edição do conteúdo, e a sua preparação para a indexação em mecanismos de busca. Este artigo envolve a fase de análise e edição do conteúdo dos livros, que requer trabalho humano de curadoria para sua execução.

Os autores da metodologia ainda enumeram exemplos de tarefas realizadas durante a curadoria, entre as quais: cópia de tabelas em todas as questões em que é citada, inclusão do significado das siglas utilizadas no texto, cópia das referências bibliográficas citadas em uma determinada resposta, edição de respostas que citam outras perguntas e transformação de imagens para o formato Base64, que possibilita sua inclusão diretamente em um arquivo HTML em formato textual. Nesta etapa, algumas perguntas sofrem alterações para que ganhem sentido completo e as imagens que não se relacionam às suas respectivas perguntas são removidas.

Para a edição do HTML, utiliza-se o editor de livros digitais Sigil. Na Figura 1, observa-se sua interface após a abertura do livro sobre feijão [Gonzaga 2014]. À esquerda, visualiza-se a estrutura do livro, com pastas e arquivos específicos para cada componente, inclusive os textuais, como capítulos, notas e referências. Ao centro, é exibido o corpo de texto em HTML, representado pelo arquivo ‘c5.xhtml’ da pasta ‘Text’. O texto é exibido com o recurso de destaque da sintaxe HTML, em que são

usadas cores distintas para representar cada tipo de elemento. À direita, é exibida a pré-visualização do conteúdo, que contribui na etapa de edição. Para comparação e análise do conteúdo são utilizados os arquivos PDF da obra enquanto a edição é realizada.

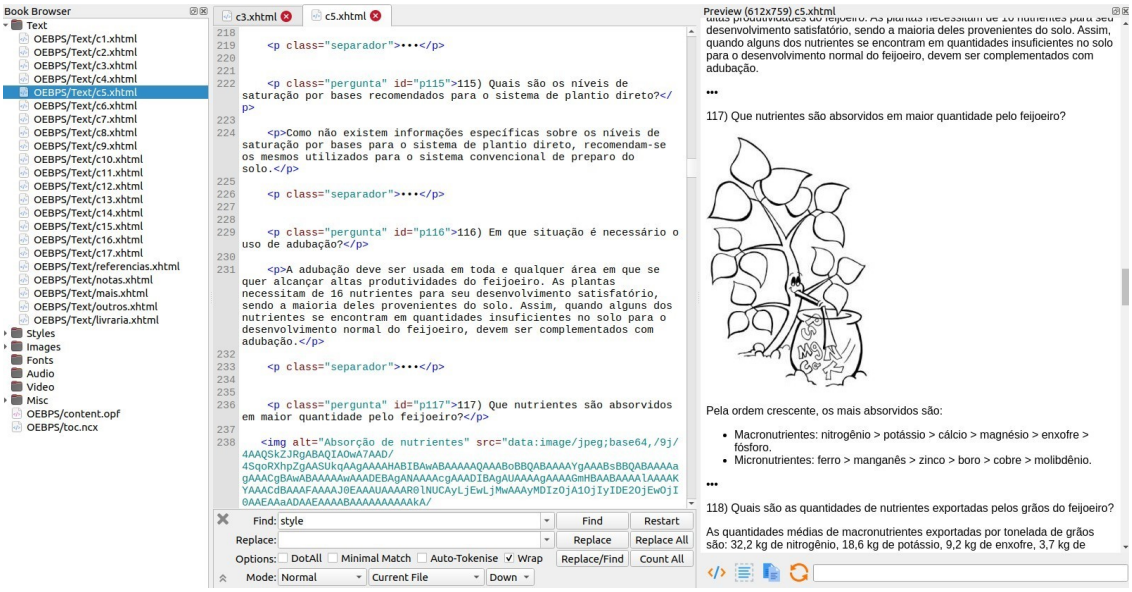


Figura 1. Tela do editor Sigil.

Durante o trabalho de curadoria, algumas imagens são removidas porque não contribuem para a compreensão do texto. As imagens mantidas recebem devido tratamento, realizado através do editor de imagens GIMP, visando manter a sua qualidade e diminuir seu tamanho. Assim, são transformadas para escala de cinza, redimensionadas para um tamanho padrão e exportadas para o formato jpeg com 60% de qualidade. Depois são convertidas para o formato Base64 para serem incluídas no corpo do texto codificado em HTML.

A curadoria de cada livro da coleção gera como resultado um único arquivo no formato HTML, a partir da concatenação dos arquivos correspondentes a todos os capítulos. Então, esse arquivo é incluído em uma base de dados que foi construída para disponibilizar esse conteúdo de maneira aberta e que já está disponível no Redape [Vaz et al., 2022].

3. Resultados e Discussão

O Redape conta com a primeira versão dos arquivos com o conteúdo dos livros da coleção 500 Perguntas 500 Respostas tratados e indexados. Ela inclui dados dos livros sobre algodão, integração lavoura-pecuária-floresta (ILFP), feijão-caupi, produção orgânica de hortaliças, e sobre pêssego, nectarina e ameixa. Novas versões serão disponibilizadas no futuro, com a inclusão de novos livros que abordam temas como mandioca, feijão, uva, manga e suínos. Alguns exemplos de alterações que estão sendo realizadas nessas obras podem ser visualizadas nos Quadros 1 e 2.

245) Quais são as principais doenças causadas por fungos que sobrevivem no solo e que atacam o feijoeiro-comum?

As principais doenças causadas por fungos que sobrevivem no solo são: podridão-radicular, causada por *Rhizoctonia solani* e *Fusarium solani* f. sp. *phaseoli*; murcha-de-fusário, causada por

Fusarium oxysporum f. sp. *phaseoli*; mofo-branco, causado por *Sclerotinia sclerotiorum*; podridãocinzenta-da-haste, causada por *Macrophomina phaseolina*; murcha-de-esclerócio, causada por *Sclerotium rolfsii*; e mela ou murcha-de-teiamicélica, causada por *Thanatephorus cucumeris* (forma de reprodução sexuada de *R. solani*).

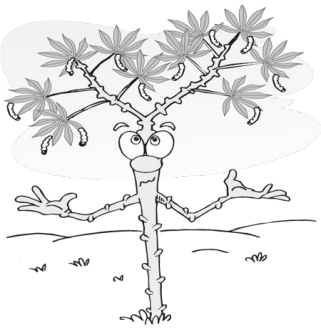
246) Como **estes patógenos** sobrevivem no solo?

246*) Como **os patógenos que atacam o feijoeiro-comum** sobrevivem no solo?

Quadro 1. Perguntas 245 e 246 do livro de feijão.

303) Qual a melhor forma de controle para grandes plantios?

Para grandes plantios recomenda-se a aplicação do *Baculovirus weinnyis* ou de produtos formulados à base de *Bacillus thuringiensis*.



304) Em que ínstares do mandarová a aplicação **desses produtos** é mais eficiente?

A aplicação desses produtos é mais eficiente no período compreendido entre o 1º e 3º ínstar.

304*) Em que ínstares do mandarová a aplicação **do *Baculovirus erinnyis* ou de produtos formulados à base de *Bacillus thuringiensis*** é mais eficiente?

Quadro 2. Perguntas 303 e 304 do livro de mandioca.

A leitura sequencial tradicionalmente atribuída a livros físicos não se aplica ao seu uso em ferramentas digitais. A dinâmica de leitura é diferente, uma vez que as perguntas e as respostas podem ser acessadas de maneira direta e individualizada, sem que seja necessário acessar perguntas anteriores. Entretanto, em alguns casos, as perguntas foram elaboradas de maneira a depender de outras perguntas. Essa situação é explicitada tanto no Quadro 1, onde a pergunta 246 do livro sobre feijão [Gonzaga 2014] requer elementos da questão 245 para que seja compreendida, quanto no Quadro 2, em que a mesma situação ocorre com as perguntas 303 e 304 do livro sobre mandioca [Mattos et al. 2006]. Em ambos os casos, modificam-se os textos para tornar as perguntas completas.

No Quadro 1, o trecho 'estes patógenos' da questão 246 refere-se a 'os patógenos que atacam o feijoeiro-comum', conforme indica a questão anterior. Portanto, a nova redação da questão, indicada por asterisco, é alterada para incluir o trecho completo e permitir sua compreensão independentemente de outras questões. Situação semelhante ocorre no Quadro 2.

A decisão sobre manter ou excluir imagens também é efetuada conforme análise realizada pelo curador. A imagem do Quadro 1, por exemplo, foi removida porque não contribui para a compreensão da questão, enquanto a imagem do Quadro 2 foi mantida, por auxiliar na compreensão do conteúdo do texto. O trabalho executado exige habilidades analíticas que normalmente são associadas a humanos e que podem ser inviáveis de ser implementadas em software.

4. Conclusão

Este artigo apresentou a fase de análise e edição de conteúdo relacionada à curadoria digital dos livros da coleção 500 Perguntas 500 Respostas da Embrapa. O trabalho humano envolvido nesse processo garante a qualidade dos dados que são disponibilizados abertamente e que podem ser utilizados no desenvolvimento de diferentes soluções digitais.

O trabalho realizado apresenta grande potencial na contribuição de avanços sociais e tecnológicos, pois a partir do desenvolvimento tecnológico associado ao trabalho humano, caminha-se rumo à Sociedade 5.0, contribuindo-se para o atendimento a necessidades sociais.

Referências

Abbott, D. (2008). "What is digital curation?". In DCC Briefing Paper, Edimburgo, Sld. Digital Curation Centre.

Embrapa (2023). Repositório de Dados de Pesquisa da Embrapa (Redape). https://www.redape.dados.embrapa.br/. Acessado:14-07-2023.

Gonzaga, A.C.O. (2014). (Ed.). Feijão: o produtor pergunta, a Embrapa responde. Brasília, DF: Embrapa. (Coleção 500 perguntas 500 respostas).

Japan Cabinet Office (2023). "What is society 5.0?", https://www8.cao.go.jp/cstp/english/society5_0/index.html. Acessado: 14-07-2023.

Mattos, P.L.P., Farias, A.R.N., and Filho, J.R.F. (2006). (Ed.). Mandioca: o produtor

pergunta, a Embrapa responde. Brasília, DF: Embrapa. (Coleção 500 perguntas 500 respostas).

Rusbridge, C. et al. (2005). The digital curation centre: a vision for digital curation. IEEE International. Symposium on Mass Storage Systems and Technology. Proceedings. IEEE. p. 31-41.

Serpa, S., and Ferreira, C. (2019). “Society 5.0 and sustainability digital innovation: a social process”. Journal of Organizational Culture, Communications and Conflicts, Azores, v-3, issue 1, pages 1-14.

Vaz, G. J., Veiga, P. H. R. C., Caldas, R. G., Vidal W. C. L., Assis, C. P., Correa, J. L., and Moura, M. F. (2023). Tratamento de texto extraído de livros digitais para a indexação em mecanismo de busca. Aceito para publicação na Revista Ibero-Americana de Ciência da Informação.

Vaz, G. J., Veiga, P. H. R. C., and Moura, M. F. (2022). Content from the books of Embrapa's 500 Questions 500 Answers Collection (Coleção 500 Perguntas 500 Respostas) treated to be used in digital solutions. https://doi.org/10.48432/YIGNPF