RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

Tratamento de texto extraído de livros digitais para a indexação em mecanismo de busca

Glauber José Vaz
Embrapa Agricultura Digital, Campinas, SP, Brasil
glauber.vaz@embrapa.br

Pedro Henrique Rodrigues da Cunha da Veiga
IZagro, Franca, SP, Brasil
pedro@izagro.com.br

Rafael Gomes Caldas
IZagro, Franca, SP, Brasil
rafael.gcaldas01@gmail.com

Wyviane Carlos Lima Vidal
Embrapa Agroenergia, Brasília, DF, Brasil
wyviane.vidal@embrapa.br

Cristiane Pereira de Assis
Embrapa Sede, Superintendência de Comunicação, Brasília, DF, Brasil
cristiane.assis@embrapa.br

ARTIGOS

Jorge Luiz Correa
Embrapa Agricultura Digital, Campinas, SP, Brasil
jorge.l.correa@embrapa.br

Maria Fernanda Moura
Embrapa Agricultura Digital, Campinas, SP, Brasil
maria-fernanda.moura@embrapa.br

DOI: https://doi.org/10.26512/rici.v16.n2.2023.42740

Recebido/Recibido/Received: 2022-10-25
Aceitado/Aceptado/Accepted: 2023-07-13

Resumo
Este trabalho apresenta uma metodologia de tratamento dos textos extraídos dos livros digitais da Coleção 500 Perguntas 500 Respostas da Embrapa a fim de que seu conteúdo possa ser indexado e acessado via um mecanismo de busca específico. A metodologia envolve a extração dos elementos essenciais dos livros, como imagens e arquivos HTML, o pré-processamento desses elementos, sua análise e edição, e a construção de componentes adequados para sua indexação. Além de um intenso trabalho de análise humana, são consideradas tecnologias como o formato Epub para livros digitais, o editor Sigil, scriptspara processamento de texto, padrões web de representação e Elasticsearch. Experimentos mostram que a metodologia viabiliza a disponibilização de textos bem formatados para sua indexação e seu uso em mecanismos de busca, propiciando uma rica experiência ao usuário, além de possibilitar a construção de novas soluções digitais. Nesse contexto, a curadoria digital é fundamental para agregar valor aos recursos digitais e atender às necessidades específicas de seus usuários.
Palavras-chave: Curadoria digital. Recuperação da informação. Processamento de texto. Disseminação da informação. Indexação. Livros Digitais.

Treatment of text extracted from digital books for search engine indexing
Abstract
This article presents a methodology for treating texts extracted from digital books from Embrapa's 500 Questions 500 Answers Collection to index their content and to allow its access via a search engine. The methodology involves extracting the essential elements of the books, such as images and HTML files; pre-processing them; analyzing and editing them; and building suitable components for their indexing. In addition to a large amount of human analysis, the technologies used are Epub format for digital books, the Sigil editor, scripts for text processing, web representation standards, and Elasticsearch. The results show that this method can provide well-formatted texts for indexing and use in search engines, giving a rich user experience and enabling the construction of new digital solutions. Therefore, such a digital curation is essential for adding value to digital resources and meeting specific user needs.
Keywords: Digital curation. Information retrieval. Text processing. Information dissemination. Indexing. Digital books.

Tratamiento del texto extraído de los libros digitales para su indexación en los motores de búsqueda
Resumen
Este trabajo presenta una metodología para el tratamiento de los textos extraídos de los libros digitales "500 Preguntas 500 Respuestas" de Embrapa, para que su contenido pueda ser indexado y accedido a través de un motor de búsqueda específico. La metodología presentada implica la extracción de elementos esenciales del libro (como, por ejemplo, imágenes y archivos HTML), el preprocesamiento de estos elementos, su análisis y edición, y por último, la construcción de componentes adecuados para su indexación. Además de un exhaustivo trabajo de análisis humano, se tuvieron en cuenta tecnologías como el formato Epub para libros digitales, el editor Sigil, scripts para el tratamiento de textos, estándares de representación web y Elasticsearch. Los resultados obtenidos muestran que la metodología permite disponer de textos viables para su indexación y su utilización en los motores de búsqueda, proporcionando al usuario una experiencia rica, además de permitir la construcción de nuevas soluciones digitales. En este contexto, la curación digital es fundamental para añadir valor a los recursos digitales y satisfacer las necesidades específicas de los usuarios.
Palabras-clave: Curación digital. Recuperación de información. Tratamiento de textos. Difusión de información. Indexación. Libros digitales.

1 Introdução

Uma grande quantidade de livros digitais vem sendo produzida nas diferentes áreas do conhecimento. No entanto, segundo Martins (2016), as oportunidades oferecidas pela sua natureza digital são inibidas devido à maneira como são tratados, semelhante ao tratamento dispensado a documentos impressos. Para o autor, os leitores não são acolhidos em suas perspectivas de consumo e é necessário aperfeiçoar o arcabouço que envolve o gerenciamento da informação digital. Entre os principais obstáculos para a expansão dos livros digitais na sociedade brasileira, ele cita a utilização de modelos de negócios inviáveis na contemporaneidade.

Devido ao dinamismo em seu uso, as informações digitais passaram a ser usadas em contextos que não foram previstos no momento em que foram coletadas ou criadas (NationalResearchCouncil, 2015). Soluções digitais baseadas no conteúdo de livros, por exemplo, podem facilitar o acesso e oferecer melhor experiência aos leitores**,** especialmente aquelas que incluem mecanismos de busca. Assim, os livros digitais podem ser explorados de maneiras alternativas, inclusive por meio de modelos de negócios mais atuais, como os que

envolvem APIs (*ApplicationProgramming Interface*). Fundamentais no cenário atual de transformação digital, as APIs possibilitam a comunicação automática entre diferentes aplicações.

Teixeira e Valentim (2017) defendem que é possível criar produtos e serviços informacionais distintos e baseados na necessidade do público-alvo, e que maior inovação e geração de ideias são viabilizadas pelos fluxos de informação. A curadoria digital, de acordo com Bax e Resende (2020), busca justamente desenvolver estratégias para resolver problemas do fluxo da informação digital, de maneira a agregar valor aos repositórios digitais para uso presente e futuro, reduzindo a obsolescência digital e mantendo as informações acessíveis aos usuários. A curadoria digital trata da gestão atuante e do aprimoramento de ativos de informação digital para uso atual e futuro, envolvendo muitas vezes a representação dos dados para satisfazer necessidades específicas e auxiliar nos desafios relacionados a interoperabilidade e acessibilidade (National Research Council, 2015).

A gestão dos objetos e dados digitais, segundo Brayner (2018), ”é um processo que demanda um acompanhamento constante do desenvolvimento tecnológico e revisão das políticas de gestão digital, de modo a garantir a acessibilidade e a utilização dos conteúdos eletrônicos no futuro”. Rusbridge *et al*. (2005) também apontam a necessidade de intervenções regulares e planejadas para amanutenção, a usabilidade e a sobrevivência dos recursos digitais. Estes, de acordo com Higgins (2008), são suscetíveis a mudanças tecnológicas desde o momento de sua criação.

Oliver e Harvey (2016) consideram que a curadoria digital começa antes mesmo da criação dos objetos digitais, com o estabelecimento de padrões para planejar a coleta dos dados que resulta nos objetos digitais que estarão na melhor condição possível de assegurar sua manutenção e seu uso futuro. Para os autores, a curadoria digital foca na agregação de valor aos conjuntos de dados e objetos digitais de maneira a inserir novos metadados ou anotações para serem reusados. Os benefícios da curadoria envolvem melhoria no acesso, na qualidade e na proteção dos dados, além do estímulo ao compartilhamento e ao reúso.

Oliver e Harvey (2016) ainda caracterizam a curadoria digital por um conjunto de processos aplicados aos objetos digitais desde sua criação, por uma preocupação com a reprodutibilidade dos dados e pela agregação de valor aos objetos digitais para que sejam reutilizados, inclusive para auxiliar na descoberta, na gestão e na recuperação dos dados. Ainda apontam o envolvimento dos *stakeholders*, o forte interesse em soluções *open-source* e a forte ligação entre a pesquisa e a prática.

No contexto da agricultura tropical, a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) é uma das maiores produtoras de conhecimento técnico-científico no mundo. Suas

obras normalmente são acessadas por meio de publicações completas disponíveis em papel ou em formatos eletrônicos, que podem ser encontradas por meio de buscas em ferramentas digitais de uso mais amplo, como o Google Acadêmico (https://scholar.google.com.br/), ou ferramentas disponibilizadas pela própria empresa, como Alice (http://www.embrapa.br/alice), BDPA (http://www.embrapa.br/bdpa), Infoteca (http://www.embrapa.br/infoteca) e o próprio Portal Embrapa (http://www.embrapa.br/biblioteca). No entanto, essas ferramentas indexam apenas os metadados das obras, o que torna impossível uma busca diretamente por seu conteúdo.

A Coleção 500 Perguntas 500 Respostas, editada pela Embrapa em parceria com outras instituições, é uma das principais coleções da linha editorial de transferência de tecnologia. Os temas dessa coleção são relacionados à agricultura e à pecuária e os títulos são elaborados a partir de perguntas formuladas por produtores, associações de produtores, cooperativas, etc., e respondidas pelos pesquisadores da Embrapa. O público-alvo da coleção são produtores rurais, técnicos da extensão rural, estudantes de escolas agrotécnicas e sociedade em geral. Em 2013, a Embrapa converteu a coleção para *e-book*. O *site* da Coleção 500 Perguntas 500 Respostas (EMBRAPA, 2022) foi desenvolvido para disponibilizar, gratuitamente, os *e-books* nos formatos Epub e PDF, que podem ser lidos em microcomputadores e dispositivos móveis, como *smartphones*, *tablets* e leitores de *e-book*.

Embora essa coleção seja direcionada também aos pequenos agricultores, com linguagem mais simples e acessível, é necessário haver um nível de habilidade mínimo que possibilite o usuário acessar ambientes virtuais para poder explorar o conteúdo dessas obras por meio de mecanismos de busca. Segundo Moreira *et. al.* (2017), muitas da necessidades informacionais dos pequenos produtores não são atendidas devido às condições de conectividade e à falta de conteúdo informacional compatível com a cultura desses produtores.

A digitalização das bibliotecas melhora o acesso a suas obras e possibilita maior interoperabilidade entre os sistemas de informação, mas demanda novas habilidades para a gestão dos acervos digitais (Cunha, 2022), pois os serviços de informação e o trabalho dos profissionais da área já estão sendo fortemente impactados pela transformação digital e pela inteligência artificial (Gomes, 2022).

A qualidade do acesso ao conteúdo digital depende muito dos processos de indexação. Porém, Teixeira e Spiassi (2022), por exemplo, verificaram que mesmo o resumo das obras é pouco empregado na recuperação de informação em bibliotecas mantidas pelas instituições federais brasileiras de ensino superior. Tartarotti e Dal'Evedove (2021) avaliaram a indexação dos repositórios institucionais das principais universidades públicas paulistas, por meio de índices que revelam o nível de consistência entre indexações distintas. O estudo revelou baixos

índices de consistência entre as indexações realizadas nos repositórios e sugerem o direcionamento de esforços e recursos para a melhoria da indexação em repositórios institucionais.

Em grande medida, uma melhor indexação depende da curadoria realizada nos recursos digitais. Porém, embora haja oportunidades significativas nas organizações para a adoção das práticas de curadoria digital, seus benefícios ainda não são muito bem compreendidos (National Research Council, 2015). Por isso, são importantes o compartilhamento e a difusão de experiências como a deste trabalho, que relatam práticas de curadoria digital e mostram seus benefícios.

O objetivo deste trabalho é apresentar uma metodologia de tratamento de textos extraídos de livros digitais disponibilizados em formato Epub. Assim, é possível indexar seu conteúdo e torná-lo acessível de forma facilitada por meio de um mecanismo de busca que explora padrões consolidados para a troca de informações. As obras consideradas neste trabalho são da Coleção 500 Perguntas 500 Respostas da Embrapa.

2 Material e Métodos

A Coleção 500 Perguntas 500 Respostas conta com dezenas de publicações que são estruturadas da mesma maneira, exibindo perguntas numeradas de 1 a 500 seguidas pelas suas respostas. Estes livros são disponibilizados em formatos PDF e Epub. O primeiro não é muito adequado para processamento computacional, embora haja trabalhos que o considerem, como o de Kasmani, Maniyar e Narvekar (2020), mas o Epub sim. Este, que tem sido amplamente utilizado para livros digitais (*e-books*), é um formato de distribuição e intercâmbio para publicações e documentos digitais que fornece um meio de representar, empacotar e codificar conteúdo web para distribuição em um arquivo contêiner único (W3C EPUB 3 Community Group, 2019a). A versão corrente do Epub consiste de uma família de especificações que envolvem requisitos tanto para as publicações digitais quanto para as aplicações que consomem e apresentam o conteúdo dessas publicações. A especificação para as publicações define um formato baseado nos padrões web e XML, com a versão mais recente do HTML (W3C EPUB 3 Community Group, 2019b), que constitui a linguagem de marcação utilizada nas páginas da web e interpretada pelos navegadores.

Editores de Epub podem ser utilizados para manipular publicações neste formato. No presente trabalho, foi utilizado o Sigil, editor sob licença livre GNU GPL. Na Figura 1, é exibida uma tela do editor após a abertura do livro sobre Integração Lavoura-Pecuária-Floresta (ILPF) (Cordeiro *et al*., 2015). Na janela à esquerda, exibe-se a estrutura do livro, com pastas específicas para diferentes tipos de elementos, como texto, imagens, fontes utilizadas, áudios e outros. Os

livros considerados neste trabalho apresentam uma estrutura com três principais pastas: (i) ‘Text’: contém arquivos em HTML com textos das perguntas e respostas separados por capítulos, além de arquivos para componentes do livro, como a introdução, a apresentação e a capa; (ii) ‘Styles’: com arquivos de estilo para apresentação no formato CSS (*CascadingStyleSheets,* ou Folhas de Estilo em Cascata), utilizado para especificar a aparência das páginas na web; e (iii) ‘Images’: com as figuras e ilustrações. Embora o livro seja distribuído como um único arquivo digital com extensão ‘.epub’, ele é formado por vários arquivos empacotados. Na janela à direita da Figura 1, exibe-se parte do conteúdo do quinto capítulo do livro sobre ILPF, presente no arquivo ‘c5.html’ da pasta ‘Text’. Seu conteúdo é descrito em HTML e há referências a elementos do tipo ‘class’, como ‘grupo’, ‘pergunta’, ‘italico’, ‘separador’ e ‘sobrescrito’. Esses elementos são definidos nos arquivos CSS com instruções para a forma de apresentação visual.

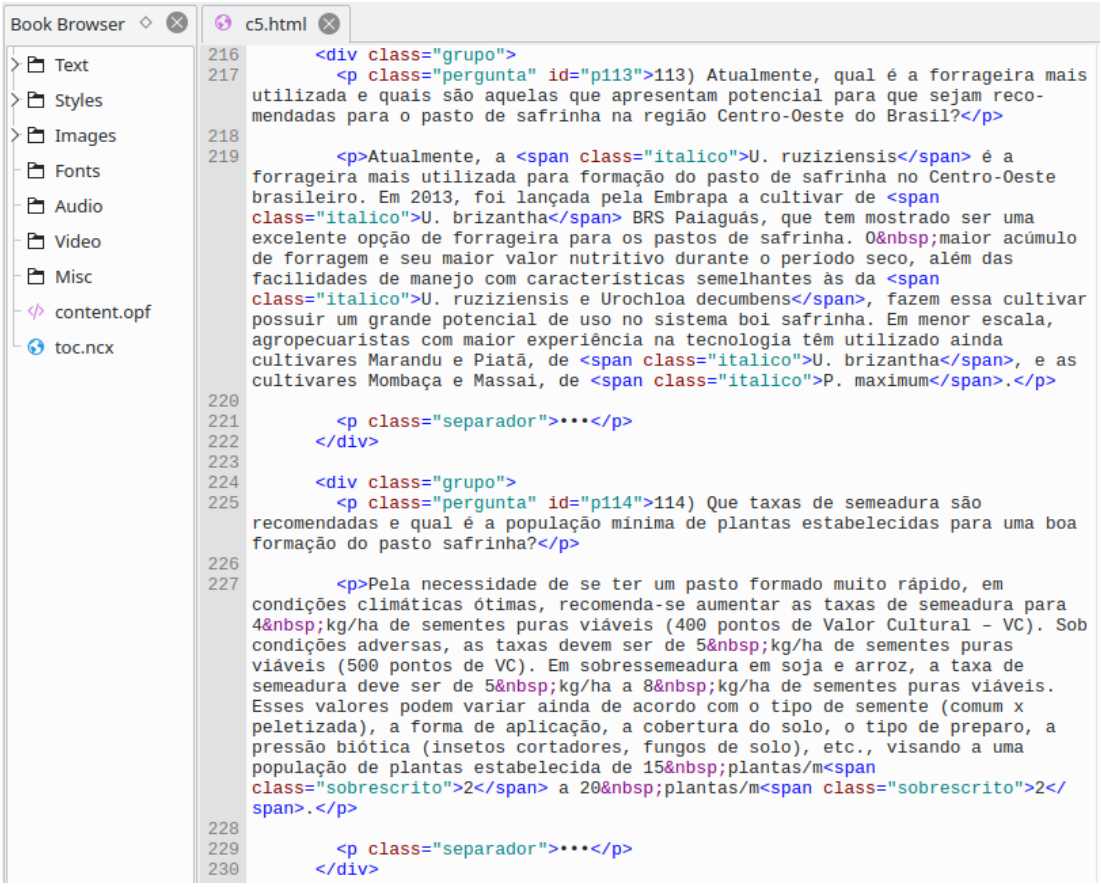

Figura 1: Tela do editor Sigil, após aberto o capítulo 5 do livro sobre ILPF da Coleção 500 Perguntas 500 Respostas.

Este trabalho visa à preparação dos textos para a sua recuperação por meio de um mecanismo de busca, que foi implementado com Elasticsearch. Esta tecnologia oferece

armazenamento e indexação eficientes de dados, viabilizando buscas rápidas em diferentes tipos de aplicações (ELASTIC, 2022). Também oferece uma interface simples via API, de maneira que aplicações de terceiros possam embutir o sistema em suas próprias soluções.

Além das tecnologias mencionadas, exige-se um intenso trabalho de pessoas na análise dos textos. De acordo com Rehm *et. al.* (2020), processos ineficientes aumentam ainda mais o esforço de processamento manual. Por isso, criaram uma plataforma para simplificar e agilizar a curadoria de conteúdo digital. Embora não vise à criação de uma plataforma, este trabalho estabelece um processo para o tratamento de textos.

3 Resultados

Na Figura 2, é apresentada a metodologia desenvolvida neste trabalho para o tratamento dos textos contidos nas obras da Coleção 500 Perguntas 500 Respostas, a fim de se tornar possível o acesso ao conteúdo dessas obras via um mecanismo de busca específico e de fácil acesso. A partir do livro em formato Epub, obtêm-se arquivos preparados para a indexação de acordo com as etapas a seguir. Um conjunto inicial desses arquivos está disponível no Repositório de Dados de Pesquisa da Embrapa (Rodapé) (Vaz; Veiga; Moura, 2022).

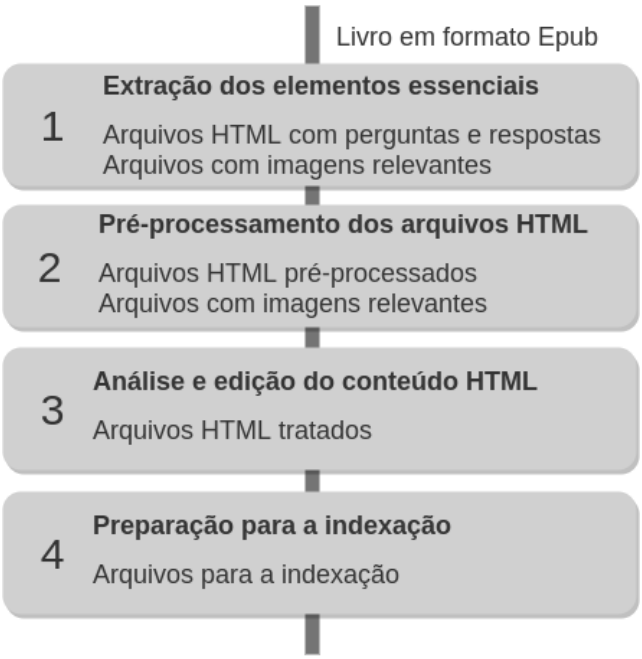

Figura 2: Metodologia para o tratamento de textos de livros digitais.

a. Extração dos elementos essenciais

A partir dos livros em formato Epub, são extraídos os arquivos HTML com o conteúdo

das perguntas e respostas, e as imagens que auxiliam na compreensão das respostas. São ignorados arquivos com elementos como introdução, apresentação, autores e capa, e também aqueles referentes ao estilo de apresentação em formato CSS, que é utilizado para descrever como os elementos do documento devem ser apresentados, uma vez que a apresentação deve ser determinada pelos desenvolvedores de soluções que utilizarem a API resultante deste trabalho. Esta etapa pode ser realizada com o auxílio de editores de livros digitais como o Sigil.

Em uma abordagem inicial, para manter os textos organizados da maneira mais simples possível, procurou-se manter uma estrutura envolvendo apenas texto, em que as perguntas eram identificadas por números ordenados de 1 a 500 e descritas em linhas únicas. As respostas correspondentes eram encontradas nas linhas seguintes até o surgimento de uma nova pergunta. Na Figura 3, é ilustrado um exemplo para esta abordagem, considerando-se o mesmo conteúdo exibido na Figura 1.

113) Atualmente, qual é a forrageira mais utilizada e quais são aquelas que apresentam potencial para que sejam recomendadas para o pasto de safrinha na região Centro-Oeste do Brasil?
Atualmente, a U. ruziziensis é a forrageira mais utilizada para formação do pasto de safrinha no Centro-Oeste brasileiro. Em 2013, foi lançada pela Embrapa a cultivar de U. brizantha BRS Paiaguás, que tem mostrado ser uma excelente opção de forrageira para os pastos de safrinha. O maior acúmulo de forragem e seu maior valor nutritivo durante o período seco, além das facilidades de manejo com características semelhantes às da U. ruziziensis e Urochloa decumbens, fazem essa cultivar possuir um grande potencial de uso no sistema boi safrinha. Em menor escala, agropecuaristas com maior experiência na tecnologia têm utilizado ainda cultivares Marandu e Piatã, de U. brizantha, e as cultivares Mombaça e Massai, de P. maximum.

114) Que taxas de semeadura são recomendadas e qual é a população mínima de plantas estabelecidas para uma boa formação do pasto safrinha?
Pela necessidade de se ter um pasto formado muito rápido, em condições climáticas ótimas, recomenda-se aumentar as taxas de semeadura para 4 kg/ha de sementes puras viáveis (400 pontos de Valor Cultural – VC). Sob condições adversas, as taxas devem ser de 5 kg/ha de sementes puras viáveis (500 pontos de VC). Em sobressemeadura em soja e arroz, a taxa de semeadura deve ser de 5 kg/ha a 8 kg/ha de sementes puras viáveis. Esses valores podem variar ainda de acordo com o tipo de semente (comum x peletizada), a forma de aplicação, a cobertura do solo, o tipo de preparo, a pressão biótica (insetos cortadores, fungos de solo), etc., visando a uma população de plantas estabelecida de 15 plantas/m2 a 20 plantas/m2.

Figura 3: Primeira proposta para organização dos textos

No entanto, esse padrão não é suficiente para promover uma experiência de leitura satisfatória. Alguns exemplos de casos que não são bem representados por textos puros e que são encontrados nas respostas das perguntas são enumerados a seguir:

- Tabelas.
- Referências a figuras.
- Textos que precisam de representação de caracteres sobrescritos ou subscritos, como em h^{-1}, P_2O_5, Al^{3+} e $mmol_c$.
- Símbolos que não são muito comuns, como '±' e '≤'.
- Lista de tópicos.

- Termos que normalmente demandam destaque em itálico para melhor leitura, como nomes científicos e nomes estrangeiros.

Para representar mais adequadamente esse tipo de conteúdo e outros presentes em livros digitais, os textos foram mantidos em HTML. Apesar de possibilitar uma representação mais completa de conteúdo, ela demanda apenas caracteres de codificações padrões. O formato Epub já atende às especificações do HTML, mas aceita também outros recursos, como, por exemplo, o CSS. Neste trabalho, foi adotada a representação dos textos apenas com uso de HTML sem explorar outros padrões web, para tornar mais simples o processamento computacional e porque, neste contexto, elementos complementares são desnecessários.

b. Pré-processamento dos arquivos HTML

Para facilitar a análise e a edição do conteúdo por humanos, os textos foram processados de maneira a realizar tarefas que pudessem ser feitas automaticamente, sem necessidade de intervenção humana. Nesta etapa, scripts foram utilizados para, por exemplo, descartar os trechos dos capítulos referentes aos sumários, eliminar *tags* HTML desnecessárias e substituir trechos por outros. Uma alternativa a essa abordagem é o uso do próprio editor Sigil para a realização dessas tarefas.

Na Figura 4 é apresentado o resultado do pré-processamento do texto mostrado na Figura 1. No tratamento, são removidas algumas referências a classes de CSS, uma vez que toda informação relevante deve estar contida na própria pergunta ou resposta. Então, informações de formatação não devem ser extraídas dos CSS, mas sim do próprio HTML. Por isso, *tags* como <spanclass="italico"> e <spanclass="sobrescrito">, que exploram CSS, são substituídas por *tags*HTML, como <i> e <sup>. Outras *tags*como <divclass="grupo"> também são removidas, pois são usadas apenas para auxiliar na apresentação da informação por meio de CSS. Referências a classes específicas como, por exemplo, ‘separador’ e ‘pergunta’ são mantidas, pois denotam a estrutura dos livros e auxiliam no processamento computacional.

```
        <p class="pergunta" id="p113">113) Atualmente, qual é a forrageira mais
utilizada e quais são aquelas que apresentam potencial para que sejam recomendadas
para o pasto de safrinha na região Centro-Oeste do Brasil?</p>

        <p>Atualmente, a <i>U. ruziziensis</i> é a forrageira mais utilizada para
formação do pasto de safrinha no Centro-Oeste brasileiro. Em 2013, foi lançada
pela Embrapa a cultivar de <i>U. brizantha</i> BRS Paiaguás, que tem mostrado ser
uma excelente opção de forrageira para os pastos de safrinha. O maior acúmulo de
forragem e seu maior valor nutritivo durante o período seco, além das facilidades
de manejo com características semelhantes às da <i>U. ruziziensis e Urochloa
decumbens</i>, fazem essa cultivar possuir um grande potencial de uso no sistema
boi safrinha. Em menor escala, agropecuaristas com maior experiência na tecnologia
têm utilizado ainda cultivares Marandu e Piatã, de <i>U. brizantha</i>, e as
cultivares Mombaça e Massai, de <i>P. maximum</i>.</p>

        <p class="separador">•••</p>

        <p class="pergunta" id="p114">114) Que taxas de semeadura são recomendadas
e qual é a população mínima de plantas estabelecidas para uma boa formação do
pasto safrinha?</p>

        <p>Pela necessidade de se ter um pasto formado muito rápido, em condições
climáticas ótimas, recomenda-se aumentar as taxas de semeadura para 4 kg/ha
de sementes puras viáveis (400 pontos de Valor Cultural – VC). Sob condições
adversas, as taxas devem ser de 5 kg/ha de sementes puras viáveis (500 pontos
de VC). Em sobressemeadura em soja e arroz, a taxa de semeadura deve ser de
5 kg/ha a 8 kg/ha de sementes puras viáveis. Esses valores podem variar
ainda de acordo com o tipo de semente (comum x peletizada), a forma de aplicação,
a cobertura do solo, o tipo de preparo, a pressão biótica (insetos cortadores,
fungos de solo), etc., visando a uma população de plantas estabelecida de
15 plantas/m<sup>2</sup> a 20 plantas/m<sup>2</sup>.</p>

        <p class="separador">•••</p>
```

Figura 4: Trecho de arquivo HTML após pré-processamento

c. Análise e edição do conteúdo HTML

Livros normalmente são organizados de tal forma que seja esperada uma leitura sequencial. No caso desta coleção, já é esperada uma dinâmica diferente em sua leitura porque as perguntas e respostas são agrupadas por temas e possibilitam um acesso direto a questões específicas. No entanto, há vários casos em que o par de pergunta e resposta não apresenta todo o conteúdo necessário para a compreensão completa da questão, o que demanda tratamento do texto. Por exemplo:

- Uma tabela presente na resposta de uma questão é referenciada em resposta a outra pergunta. Neste caso, é necessário copiar a tabela na resposta de todas as perguntas que a referenciam.
- Termos e siglas que foram definidos em uma questão são usados em outra pergunta sem a definição correspondente nesta última. Nos casos em que a definição é essencial para a compreensão da resposta, as definições dos termos e siglas são adicionadas na resposta.
- As referências bibliográficas são enumeradas ao final de cada capítulo, mas precisam estar associadas diretamente a todas as perguntas que as citam. Então essas referências são copiadas em cada resposta que as citam.

- Respostas que fazem referências a outras perguntas do livro. Em muitos casos, a referência é desnecessária, então é apenas removida. Caso contrário, a resposta recebe tratamento conforme a necessidade específica.
- Notas de rodapé, que são informadas nas respostas e enumeradas ao final do capítulo, com *links*. Essas notas são copiadas nas respostas que as citam.

Outra tarefa executada é a transformação das imagens para o formato Base64 e sua inclusão direta no arquivo HTML. Essa codificação permite a representação de imagens como texto, o que possibilita o uso de apenas um arquivo para armazenar todo o conteúdo de um livro, sem a necessidade de arquivos adicionais para as imagens. Outra vantagem é que o conteúdo das imagens e dos textos são retornados em uma única requisição via web.

Embora algumas dessas tarefas pudessem ser executadas por sistemas automáticos, seria inviável tratar todos os casos identificados. Então, essas e outras tarefas são executadas por pessoas que também realizam a análise das obras. Assim, integrantes da equipe leram o conteúdo dos livros a fim de identificar trechos que pudessem gerar resultados de busca insatisfatórios e editaram os arquivos correspondentes para gerar uma saída padronizada.

A Figura 5 exibe um exemplo de conteúdo representado com o padrão que é usado por todas as obras que são indexadas neste trabalho. Trata-se de um arquivo HTML com um cabeçalho que apresenta informações relativas à obra, como os *links* para os livros digitais em pdf e Epub, o ano de publicação, o título e o identificador do livro. Além dessas informações necessárias para a indexação, outros metadados podem ser acrescentados como os nomes dos autores, dos editores e dos responsáveis pela curadoria. Os capítulos do livro são identificados pelas *tags*<h1>. Enquanto não surge uma nova *tag* deste tipo, as perguntas continuam sendo consideradas de um mesmo capítulo. Cada pergunta é apresentada dentro de um bloco de parágrafo delimitado por e , e seu texto sempre começa com um número inteiro seguido por parêntese. A resposta é composta por tudo o que está entre este bloco de pergunta e o próximo bloco ••• .

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
  "http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">

<head>
  <meta charset="UTF-8">
  <meta name="identifier" content="ilpf">
  <meta name="pdf"
        content="https://mais500p500r.sct.embrapa.br/view/pdfs/90000033-ebook-pdf.pdf">
  <meta name="epub"
        content="https://mais500p500r.sct.embrapa.br/view/epubs/90000033/90000033-ebook-epub.epub">
  <meta name="year" content="2015">

  <meta name="author" content="Embrapa Cerrados">

  <meta name="editor" content="Nome do editor 1">
  <meta name="editor" content="Nome do editor 2">
  <meta name="editor" content="Nome do editor 3">
  <meta name="editor" content="Nome do editor 4">

  <meta name="curator" content="Nome do curador">

  <title>Integração Lavoura-Pecuária-Floresta</title>
</head>

<body>

  <h1>Conceitos e Modalidades da Estratégia de Integração Lavoura-Pecuária-Floresta</h1>

  <p class="pergunta" id="p1">1) O que é integração lavoura-pecuária-floresta (ILPF)?</p>

  <p>É um sistema de produção sustentável que integra atividades agrícolas, pecuárias e florestais,
  realizadas na mesma área, em cultivo consorciado, em sucessão ou em rotação, e busca efeitos
  sinérgicos entre os componentes do agroecossistema, contemplando a adequação ambiental, a
  valorização do homem e a viabilidade econômica da atividade agropecuária.</p>

  <p class="separador">•••</p>

  <p class="pergunta" id="p2">2) O que é integração lavoura-pecuária (ILP) ou sistema agropastoril?
  </p>

  <p>É o sistema de produção que integra os componentes agrícola e pecuário, em rotação, consórcio
  ou sucessão, na mesma área e no mesmo ano agrícola ou por múltiplos anos.</p>

  <p class="separador">•••</p>

  :

</body>
</html>
```

Figura 5: Padrão de arquivo tratado

d. Preparação para a indexação

A partir dos arquivos contendo o texto tratado em HTML, é possível fazer sua indexação para posteriores buscas. Os resultados dessas buscas devem trazer o texto conforme são apresentados no HTML, o que facilita a construção de interfaces de usuário baseadas nas tecnologias web. A indexação demanda um processo de conversão do texto presente no arquivo HTML em um formato com esses dados organizados em campos, de maneira que possam ser adequadamente analisados e indexados para buscas otimizadas.

A Figura 6 mostra um trecho do arquivo gerado a partir da entrada em HTML com a execução de um scriptconstruído especificamente para isso. Observam-se os identificadores dos

itens “ilpf_113” e “ilpf_114”, que correspondem às perguntas 113 e 114 do livro de ILPF, os números das questões em “question_number”, as próprias questões em “question”, suas respostas em “answer”, o capítulo do livro em que aparecem as perguntas (“chapter”), o nome do livro (“book”), o identificador para o livro (“book_id”), os *links* para os livros digitais nos formatos Epub (“epub”) e pdf (“pdf”), e o ano de publicação (“year”). Cada um destes campos é processado de uma maneira específica, o que é realizado de forma automática com o uso da tecnologia Elasticsearch.

A Figura 7 mostra uma interface de usuário utilizada como protótipo de validação da ferramenta, em que há essencialmente um campo para a busca e a exibição dos resultados da consulta. Esse tipo de interface só é possível devido ao trabalho de curadoria realizado de acordo com a metodologia apresentada. As informações exibidas na Figura 7 estão claramente relacionadas aos campos previamente analisados e mostrados na Figura 6 para a pergunta 113.

4 Discussão

O presente trabalho apresenta as características da curadoria digital apontadas por Oliver e Harvey (2016):

- Conjunto de processos aplicados aos objetos digitais desde sua criação. A metodologia proposta envolve um conjunto de processos aplicados aos livros digitais de maneira a promover uma recuperação das informações que propicia melhor experiência ao usuário. É importante salientar que o próprio processo de editoração da coleção considerada neste trabalho pode ser influenciado, assim como o de outras obras, uma vez que trata de inúmeras questões relacionadas à melhor exploração de recursos digitais para amplificar seu uso. A possibilidade de um objeto digital como um *e-book* ser a base de novas soluções digitais faz com que os processos envolvidos em sua criação passem a prever seu uso em outros contextos que não exclusivamente para leitura de livros digitais em equipamentos existentes, mas também para aplicações que sequer são consideradas na atualidade.

{**"index"**:{**"_id"**: "ilpf_113"}}
{**"question_number"**: 113, **"question"**:"Atualmente, qual é a forrageira mais utilizada e quais são aquelas que apresentam potencial para que sejam recomendadas para o pasto de safrinha na região Centro-Oeste do Brasil?", **"answer"**:" Atualmente, a U. ruziziensis é a forrageira mais utilizada para formação do pasto de safrinha no Centro-Oeste brasileiro. Em 2013, foi lançada pela Embrapa a cultivar de U. brizantha BRS Paiaguás, que tem mostrado ser uma excelente opção de forrageira para os pastos de safrinha. O maior acúmulo de forragem e seu maior valor nutritivo durante o período seco, além das facilidades de manejo com características semelhantes às da U. ruziziensis e Urochloa decumbens, fazem essa cultivar possuir um grande potencial de uso no sistema boi safrinha. Em menor escala, agropecuaristas com maior experiência na tecnologia têm utilizado ainda cultivares Marandu e Piatã, de U. brizantha, e as cultivares Mombaça e Massai, de P. maximum. ", **"chapter"**: "Práticas e

Manejo de Sistemas de Integração Lavoura-Pecuária na Safra e Safrinha para as Regiões Centro-Oeste e Sudeste", **"book"**: "Integração Lavoura-Pecuária-Floresta", **"book_id"**: "ilpf", **"epub"**: "https://mais500p500r.sct.embrapa.br/view/epubs/90000033/90000033-ebook-epub.epub", **"pdf"**: "https://mais500p500r.sct.embrapa.br/view/pdfs/90000033-ebook-pdf.pdf", **"year"**: 2015}
{**"index"**:{**"_id"**: "ilpf_114"}}
{**"question_number"**: 114, **"question"**:"Que taxas de semeadura são recomendadas e qual é a população mínima de plantas estabelecidas para uma boa formação do pasto safrinha?", **"answer"**:" Pela necessidade de se ter um pasto formado muito rápido, em condições climáticas ótimas, recomenda-se aumentar as taxas de semeadura para 4 kg/ha de sementes puras viáveis (400 pontos de Valor Cultural – VC). Sob condições adversas, as taxas devem ser de 5 kg/ha de sementes puras viáveis (500 pontos de VC). Em sobressemeadura em soja e arroz, a taxa de semeadura deve ser de 5 kg/ha a 8 kg/ha de sementes puras viáveis. Esses valores podem variar ainda de acordo com o tipo de semente (comum x peletizada), a forma de aplicação, a cobertura do solo, o tipo de preparo, a pressão biótica (insetos cortadores, fungos de solo), etc., visando a uma população de plantas estabelecida de 15 plantas/m² a 20 plantas/m². ", **"chapter"**: "Práticas e Manejo de Sistemas de Integração Lavoura-Pecuária na Safra e Safrinha para as Regiões Centro-Oeste e Sudeste", **"book"**: "Integração Lavoura-Pecuária-Floresta", **"book_id"**: "ilpf", **"epub"**: "https://mais500p500r.sct.embrapa.br/view/epubs/90000033/90000033-ebook-epub.epub", **"pdf"**: "https://mais500p500r.sct.embrapa.br/view/pdfs/90000033-ebook-pdf.pdf", **"year"**: 2015}

Figura 6: Trecho de arquivo gerado para a indexação

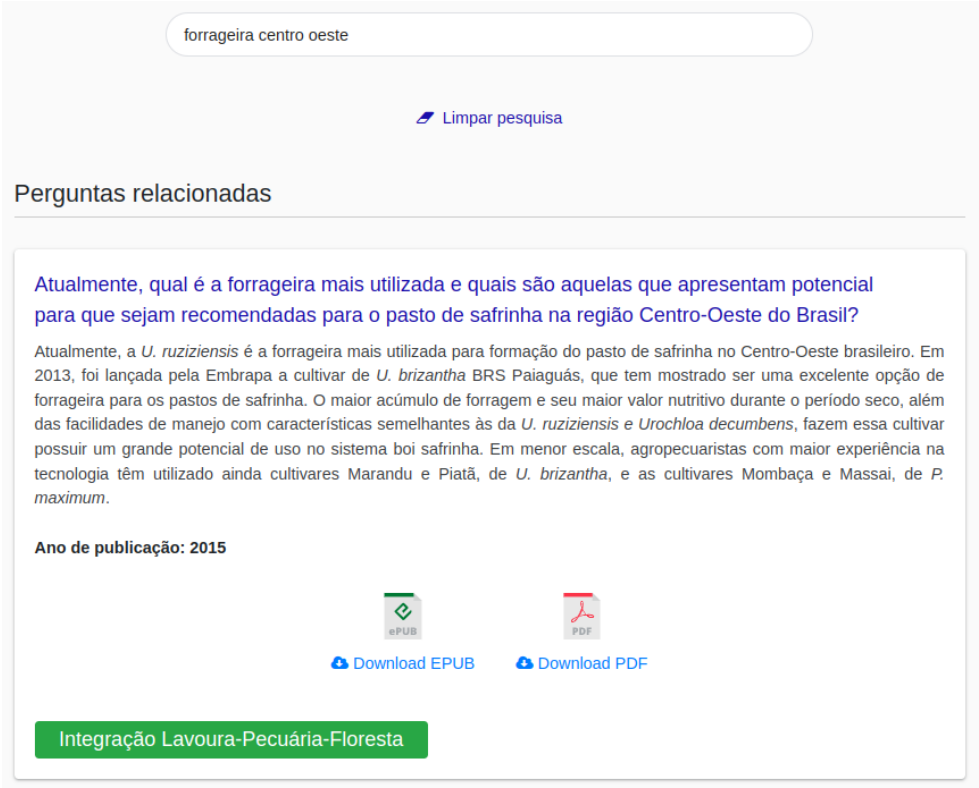

Figura 7: Interface de usuário para o protótipo do mecanismo de busca

- Preocupação com a reprodutibilidade dos dados. Os novos conjuntos de dados, já tratados, podem ser novamente gerados a partir da execução dos passos descritos. Tanto os arquivos HTML quanto os arquivos preparados para a indexação podem ser reutilizados. Estes podem ser usados em novas indexações dos livros com outras formas de análise do texto e aqueles

podem ser usados por novas aplicações baseadas em seu conteúdo.

- Agregação de valor aos objetos digitais para que sejam reutilizados, inclusive para auxiliar na descoberta, na gestão e na recuperação dos dados. A principal aplicação deste trabalho foi a própria recuperação dos textos por meio de mecanismo de busca específico. Os objetos digitais foram tratados de maneira a agregar valor com uma nova organização dos dados a fim de facilitar seu reuso e simplificar a busca do conteúdo.
- Envolvimento dos *stakeholders*. A metodologia foi construída com a participação de diversos atores. Ela foi co-desenvolvida pela instituição criadora do conteúdo e por uma startup que também busca oferecer conteúdo de qualidade a seu público-alvo ligado à produção rural. Houve envolvimento de diferentes equipes nas organizações, como as de desenvolvimento de software, infraestrutura computacional, editoração eletrônica, interface de usuário, transferência de tecnologia e pesquisadores da área. Outras instituições interessadas nesse tipo de solução, como aquelas relacionadas à extensão rural, também se envolveram na sua validação.
- Forte interesse em soluções *open-source*. O mecanismo de busca está sendo disponibilizado via uma API justamente para que terceiros possam construir soluções digitais com base nos resultados alcançados.
- Forte ligação entre a pesquisa e a prática. Esta metodologia e demais ativos foram desenvolvidos no âmbito de um projeto de pesquisa e desenvolvimento conduzido na instituição responsável pelo conteúdo em cooperação com *startup* do setor. A coleção de livros envolvida é construída com linguagem mais simples e acessível justamente para que os resultados obtidos pela pesquisa agropecuária possam alcançar os produtores rurais.

5 Conclusões

A metodologia proposta envolve a extração dos elementos essenciais dos livros digitais, o pré-processamento desses elementos, sua análise e edição, e a construção de componentes adequados para sua indexação. A criação dessa metodologia foi um processo que demandou, até a sua consolidação, muitos testes com diferentes ferramentas, padrões de dados e procedimentos, a fim de se gerar de maneira eficiente resultados que melhor atendam a aplicações de recuperação da informação e outras soluções digitais.

Os experimentos realizados mostram que a metodologia é componente fundamental da curadoria digital da Coleção 500 Perguntas 500 Respostas e viabiliza a geração de ativos que possibilitam uma melhor experiência dos usuários de mecanismos de busca baseados nos livros. Esses ativos também podem ser utilizados para a construção de novas soluções, e a metodologia pode ser adaptada para o tratamento de outras obras e em diferentes aplicações.

Soluções baseadas em livros digitais, explorando novos modelos de negócio, representam oportunidades de se agregar valor a recursos digitais e de se atender a necessidades específicas de usuários. Este trabalho mostra que a curadoria digital é fundamental nesse contexto.

Ainda assim, essa etapa é frequentemente negligenciada, uma vez que requer, além de um ferramental tecnológico e o estabelecimento de metodologias e padrões, a execução de atividades humanas que podem ser trabalhosas, demoradas e rotineiras. Um desafio, portanto, para esse tipo de curadoria é a motivação para que as tarefas sejam executadas com excelência.

Bases de texto curadas com o nível de qualidade apresentado neste trabalho serão cada vez mais demandadas, devido à crescente busca por ferramentas baseadas em inteligência artificial que utilizam amplas bases de texto para o treinamento de modelos.

Recentemente, houve um enorme avanço na construção de tecnologias que são capazes de analisar textos em língua natural e prover resultados de grande complexidade e coerência. No entanto, permanecem os desafios relacionados à veracidade das informações fornecidas. Por isso, a combinação das tecnologias envolvidas nessas ferramentas com aquelas presentes nos mecanismos de busca é uma das áreas que devem testemunhar grandes avanços nos próximos anos.

Para a construção das soluções do futuro, portanto, a disponibilidade de bases de texto confiáveis é fundamental. Isso depende de metodologias de curadoria como a apresentada neste trabalho.

Agradecimentos

À Embrapa e à IZagro, pelo suporte à pesquisa [SEG 30.21.00.016.00.00; SEG 20.22.10.019.00.00].

Referências

Bax, M. P.; Resende, L. C. A Curadoria Digital de Dados Científicos no Campo da Ciência da Informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 25, n. especial, p. 233-251, 2020.

Brayner, A. A. Curadoria digital: novos modelos de participação pública na descrição de conteúdos em instituições culturais. **Revista Ibero-Americana de Ciência da Informação**, Brasília, v. 12, n. 1, p. 53–65, 2018.

Cordeiro, L. A. M.; Vilela, L.; Kluthcouski, J.; Marchão, R. L. (Ed.). **Integração lavoura-pecuária-floresta**: o produtor pergunta, a Embrapa responde. Brasília, DF: Embrapa,

2015. (Coleção 500 perguntas, 500 respostas).

Cunha, M .B. da. Digitalização: meta urgente para as bibliotecas. **Revista Ibero-Americana de Ciência da Informação**, Brasília, v. 15, n. 1, p. 1–5, 2022.

Elastic. Elasticsearch Guide: what is Elasticsearch, 2022. Disponível em: <https://www.elastic.co/guide/en/elasticsearch/reference/current/elasticsearch-intro.html>. Acesso em 29 mar. 2022.

EMBRAPA. **Coleção 500 perguntas 500 respostas**: Você pergunta, a Embrapa responde. Disponível em: https://mais500p500r.sct.embrapa.br/view/index.php. Acesso em 29 mar. 2022.

Gomes, L. I. E. Transformação digital e Inteligência Artificial nos serviços de informação: inovação e perspectivas para a Ciência da Informação no mundo pós-pandemia. **Revista Ibero-Americana de Ciência da Informação**, Brasília, v. 15, n. 1, p. 148–166, 2022.

Higgins, S. The DCC Curation Lifecycle Model. **International Journal of Digital Curation**, v. 3, n. 1, p. 134-140, 2008.

KasmanI, F.; Maniyar, R.; Narvekar, M. Content based search engine for e-books. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). **Proceedings...**, IEEE, 2020. p. 528-533.

Martins, R. D. Obstáculos para expansão do uso dos e-books na sociedade brasileira. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 14, n. 2, p. 279-297, 2016.

Moreira, F. M. *et al.* Metadados para descrição de datasets e recursos informacionais do "Portal Brasileiro de Dados Abertos". **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 22, n. 3, p. 158-185, 2017.

National Research Council. **Preparing the workforce for digital curation**. Washington, DC: National Academies Press, 2015.

Oliver, G.; Harvey, R. **Digital curation**. Chicago: American Library Association, 2016.

Rehm, G. *et al.* QURATOR: innovative technologies for content and data curation. In: CONFERENCE ON DIGITAL CURATION TECHNOLOGIES (Qurator 2020), Berlin, Germany, 20-21 Jan. 2020. **Proceedings...**, 2020.

Rusbridge, C. *et al.* The digital curation centre: a vision for digital curation. In: IEEE INTERNATIONAL SYMPOSIUM ON MASS STORAGE SYSTEMS AND TECHNOLOGY, 2005.**Proceedings...** IEEE, 2005. p. 31-41.

Tartarotti, R. C .D.; Dal'Evedove, P. R. Avaliação da indexação em repositórios institucionais brasileiros: uma análise comparada entre USP, UNESP e UNICAMP. **Revista Ibero-Americana de Ciência da Informação**, Brasília, v. 14, n. 2, p. 583–599, 2021.

Teixeira, M. V.; Spiassi, A. O resumo como instrumento de recuperação da informação nos catálogos de bibliotecas. **Revista Ibero-Americana de Ciência da Informação**, Brasília, v. 15, n. 1, p. 76–88, 2022.

Teixeira, T. M .C.; Valentim, M. L. P. Processo de busca e recuperação de informação em ambientes organizacionais: uma reflexão teórica sobre a subjetividade da informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 22, p. 82-97, 2017.

Vaz, G. J.; Veiga, P. H. R.; Moura, M. F. Content from the books of Embrapa's 500 Questions 500 Answers Collection (Coleção 500 Perguntas 500 Respostas) treated to be used in digital solutions, **Redape**, v. 1, 2022. Disponível em: <https://doi.org/10.48432/YIGNPF>. Acesso em 20 dez. 2022.

W3C EPUB 3 CommunityGroup. **Epub 3.2: Final Community Group Specification 08 May 2019**, 2019a. Disponível em: <https://www.w3.org/publishing/epub32/epub-spec.html>. Acesso em 29 mar. 2022.

W3C EPUB 3 Community Group. **Epub Content Documents 3.2: Final Community Group Specification 08 May 2019**, 2019b. Disponível em: <https://www.w3.org/publishing/epub32/epub-contentdocs.html>. Acesso em 29 mar. 2022.