INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS FLORIANO
CURSO ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

PEDRO HENRIQUE DE SOUSA CAVALCANTE

MÉTODO PARA IMPLEMENTAÇÃO DE CONVERSÃO DE PDF PARA OFX

FLORIANO
2024

PEDRO HENRIQUE DE SOUSA CAVALCANTE

MÉTODO PARA IMPLEMENTAÇÃO DE CONVERSÃO DE PDF PARA OFX

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal do Piauí, Campus Floriano .

Orientador(a): Prof. Dr. Rafael Ângelo Santos Leite

FLORIANO
2024

MÉTODO PARA IMPLEMENTAÇÃO DE CONVERSÃO DE PDF PARA OFX

Pedro Henrique de Sousa Cavalcante[1]

Rafael Angelo Santos Leite[2]

RESUMO

A gestão financeira digital enfrenta desafios na automação de processos devido à utilização de formatos como PDF, inadequados para integração com softwares contábeis. Este trabalho apresenta o desenvolvimento de uma plataforma gratuita para converter arquivos PDF em OFX, um formato amplamente utilizado na contabilidade digital. A proposta combina técnicas avançadas de extração de dados, como expressões regulares (Regex) e bibliotecas especializadas como *Poppler*, para garantir precisão e eficiência na conversão, independentemente do layout do documento. Além disso, a aplicação integra um *backend* robusto e uma interface intuitiva, proporcionando praticidade e acessibilidade ao usuário. O projeto busca reduzir erros manuais, economizar tempo e atender à crescente demanda por soluções de automação no setor financeiro. Como contribuição, oferece uma ferramenta inovadora que aprimora a integração de dados, fortalecendo a transformação digital na contabilidade.

Palavras-chave: conversão de PDF; formato OFX; automação contábil; extração de dados; transformação digital.

ABSTRACT

Digital financial management faces automation challenges due to the use of formats like PDF, which are unsuitable for integration with accounting software. This study presents the development of a free platform to convert PDF files into OFX, a format widely used in digital accounting. The proposal combines advanced data extraction techniques, such as regular expressions (Regex) and specialized libraries like Poppler, to ensure accuracy and efficiency in conversion, regardless of document layout. Additionally, the application integrates a robust backend and an intuitive interface, providing users with accessibility and convenience. The project aims to reduce manual errors, save time, and address the growing demand for automation solutions in the financial sector. As a contribution, it delivers an innovative tool that enhances data integration and strengthens digital transformation in accounting.

Keywords: PDF conversion; OFX format; accounting automation; data extraction; digital transformation.

Data de aprovação: 18/12/2024 (data de apresentação do TCC)

[1] Graduando do curso de Tecnologia em Análise e Desenvolvimento de Sistemas. IFPI Campus Floriano. pedrohscavalcante00@gmail.com.
[2] Doutor em Ciência da Propriedade Intelectual pela UFS e docente do eixo Gestão e Negócios. IFPI Campus Floriano. rafaelangelo@ifpi.edu.br.

1 INTRODUÇÃO

A era digital trouxe uma série de benefícios para o armazenamento e a gestão de informações financeiras, facilitando o acesso, o compartilhamento de dados e a automação de processos contábeis. Cada vez mais empresas e indivíduos utilizam sistemas de gestão financeira digital para acompanhar transações, elaborar relatórios e manter suas finanças organizadas. No entanto, muitos desses sistemas dependem de formatos de arquivo específicos para realizar a integração automática de dados com softwares de contabilidade. Um dos formatos mais amplamente utilizados para esse fim é o Open Financial Exchange (OFX), desenvolvido para padronizar a troca de dados financeiros entre instituições e sistemas contábeis (Vovchenko et al., 2018).

Contudo, embora a popularidade do formato OFX seja significativa, muitas instituições financeiras ainda fornecem extratos e relatórios financeiros em formato PDF. Este formato, amplamente acessível e portátil, é ideal para visualização e armazenamento de informações, mas não foi projetado para a integração de dados estruturados. A falta de estrutura interna do PDF dificulta sua interpretação automática por sistemas contábeis, gerando um problema recorrente para profissionais que desejam automatizar suas rotinas financeiras.

Estudos anteriores têm explorado métodos para conversão de PDFs para formatos estruturados. Por exemplo, o sistema PDFX automatiza a conversão de PDFs para XML, extraindo estruturas lógicas como tabelas e metadados de maneira eficiente (Constantin; Pettifer; Voronkov, 2013). Outra abordagem é a aplicação de OCR (Optical Character Recognition), como discutido por Kruttschnitt (2022), para reconhecer dados tabulares e contextuais em documentos técnicos. Ainda assim, estas soluções não abordam plenamente as especificidades do formato OFX, deixando um espaço significativo para inovação no desenvolvimento de ferramentas de conversão de dados financeiros.

Tecnologias como **OFX Fácil** (Ofx facil, 2024) e **Importefile** (Importfile, 2024) são exemplos de solução para conversão de arquivos para o formato OFX disponíveis no mercado. Apesar de realizarem a conversão, funcionam com base em modelos predefinidos, exigindo que o usuário selecione o modelo correspondente ao banco ou instituição financeira. Isso mostra que ainda há espaço para soluções que superam essas limitações. Assim, esse estudo visa garantir maior autonomia no processamento dos dados financeiros, permitindo que a conversão para OFX seja realizada de maneira mais eficiente, independente de padrões ou modelos específicos.

Diante desse contexto, este estudo tem como objetivo desenvolver uma solução que permita a conversão eficiente e precisa de PDFs financeiros para o formato OFX, preservando a integridade dos dados e minimizando a ocorrência de erros. Assim, a questão de pesquisa deste trabalho é: *Como converter de forma eficiente e precisa documentos financeiros do tipo extratos bancários no formato PDFs para arquivos em formato OFX, preservando a integridade dos dados e minimizando a ocorrência de erros?*

Para responder a essa questão, o presente trabalho propõe um método que realiza a conversão automática de arquivos PDF para OFX. A aplicação busca oferecer uma solução prática e acessível que facilita a integração de dados financeiros com softwares contábeis, promovendo a contabilidade digital e a automação de processos. Para isso, foram explorados métodos de extração de dados de PDFs, incluindo técnicas de reconhecimento de padrões e análise de dados não estruturados, além de formas de estruturar essas informações no formato OFX de maneira compatível com os principais softwares contábeis.

A relevância deste estudo reside na crescente demanda por soluções que otimizem processos financeiros automatizados, especialmente em um mercado onde tempo e precisão são recursos valiosos. A partir do desenvolvimento dessa plataforma, espera-se contribuir para o avanço da automação contábil, oferecendo uma ferramenta que possa ser utilizada por profissionais de diferentes áreas financeiras, com segurança e confiabilidade.

2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica sustenta o desenvolvimento deste trabalho ao reunir conceitos, padrões e ferramentas que viabilizam a extração, a transformação e a integração de dados financeiros. Neste capítulo, são contextualizados os principais formatos e tecnologias envolvidos na solução proposta — com ênfase no PDF (formato orientado à apresentação), no OFX (padrão de intercâmbio financeiro) e no CSV (estrutura tabular para processamento), além de técnicas de reconhecimento de padrões baseadas em expressões regulares e bibliotecas especializadas para leitura de documentos digitais.

Também são discutidas arquiteturas e tecnologias web que dão suporte à construção de APIs e interfaces escaláveis, compondo o arcabouço necessário para automatizar o fluxo de conversão de extratos bancários em arquivos OFX. Ao integrar fundamentos conceituais e práticos, este referencial estabelece as bases para compreender os desafios de precisão, robustez e interoperabilidade na conversão de documentos heterogêneos, orientando as decisões de projeto e os critérios de avaliação do método proposto.

2.1 FORMATO PDF

O Portable Document Format (PDF) foi criado pela Adobe Systems na década de 1990 com o objetivo de preservar a apresentação visual dos documentos, independentemente do dispositivo ou sistema operacional utilizado. Apesar de sua popularidade e aceitação como padrão universal para comunicação e arquivamento de informações, o formato PDF apresenta desafios significativos para a extração de dados estruturados devido à ausência de semântica explícita e à sua estrutura interna fragmentada (Maggi; Vicario, 2023). Em contextos empresariais e acadêmicos, a conversão de elementos como tabelas e gráficos em dados utilizáveis é uma tarefa complexa que exige algoritmos avançados para identificar e estruturar as informações (Shere; Chittimalli; Naik, 2022). Ferramentas como o *PDFDataExtractor*, projetadas para a interpretação de metadados e a reconstrução da estrutura lógica de artigos científicos, têm sido empregadas para melhorar a precisão e eficiência no processamento de documentos PDF, especialmente em domínios como a química e a descoberta de materiais (Zhu; Cole, 2022).

2.2 FORMATO OFX

O Open Financial Exchange (OFX) foi introduzido em 1997 por um consórcio formado pela Microsoft, Intuit e CheckFree, com o objetivo de criar um padrão unificado para a troca eletrônica de dados financeiros. Desenvolvido para substituir protocolos proprietários como o Open Financial Connectivity (OFC), o OFX inicialmente utilizava SGML e, posteriormente, adotou XML para organizar informações financeiras de maneira estruturada, promovendo interoperabilidade entre instituições financeiras e softwares de gestão (Chepakov, 2020).

O protocolo oferece diversas vantagens, como a automação de transações e a redução da necessidade de entrada manual de dados, permitindo que informações sejam integradas automaticamente em softwares financeiros, como Quicken e Microsoft Money(Almehrej; Freitas; Modesti, 2020). A versão 2.0 do OFX, compatível com XML, ampliou as funcionalidades do padrão, incluindo suporte a dados fiscais e gestão de aposentadorias, melhorando a integração de serviços financeiros. Essa evolução reflete a tendência de digitalização no setor financeiro, onde protocolos padronizados como OFX e ISO 20022 são fundamentais para aumentar a eficiência e segurança nas operações financeiras (Nikolić, 2023). Além disso, a adoção do XML tem sido fortalecida pelo desenvolvimento de técnicas avançadas para garantir a segurança de transações, como a criptografia baseada em classificações fuzzy, que melhora o processamento e a proteção de dados financeiros (Ammari, 2013).

2.3 FORMATO CSV

O *Comma-Separated Values* (CSV) é um dos formatos mais simples e amplamente utilizados para armazenamento e intercâmbio de dados tabulares. Sua estrutura baseada em linhas e colunas, separadas por delimitadores como vírgula ou ponto e vírgula, torna o formato de fácil leitura tanto por humanos quanto por máquinas.

Em estudos sobre ciência de dados, observa-se que arquivos CSV frequentemente necessitam de pré-processamento significativo devido a inconsistências em delimitadores, formatos de dados ou estruturas de colunas (Van Den Burg; Nazabal; Sutton, 2018). Outro trabalho relevante apresenta uma extensão do CSV com metadados, o formato CSVM, que permite uma melhor interoperabilidade e documentação de dados tabulares para usos científicos (Beyries; Rodriguez, 2012).

2.4 EXPRESSÕES REGULARES (REGEX)

As expressões regulares (Regex) são ferramentas essenciais para manipular padrões textuais, especialmente em grandes volumes de dados não estruturados. Sua utilização é crucial em áreas como a extração de dados financeiros, permitindo identificar informações-chave como valores monetários e datas, mesmo em textos com layouts inconsistentes (Murtaza; Nie; Lyn, 2023). Por exemplo, o padrão \b\d{1,3}(?:\.\d{3})*(?:,\d{2})\b localiza valores monetários no formato "1.234,56", enquanto \b\d{1,2}/\d{1,2}/\d{2,4}\b identifica datas como "12/05/2023".

No contexto de documentos financeiros do tipo extratos bancários, Regex é amplamente utilizada para localizar padrões em extratos bancários e relatórios, permitindo automação em processos que antes dependiam de modelos específicos. Trabalhos recentes destacam como o uso combinado de Regex com técnicas avançadas, como aprendizado de

máquina, melhora a precisão e reduz a complexidade na extração de eventos financeiros (Malik et al., 2011). Além disso, a aplicação de Regex em sistemas de classificação e extração de dados demonstra resultados superiores quando integrados a modelos de linguagem pré-treinados (Murtaza; Nie; Lyn, 2023).

2.5 LEITURA DE PDF PARA TEXTO COM *POPPLER*

A biblioteca Poppler é amplamente reconhecida por sua eficácia na manipulação de arquivos PDF, proporcionando funcionalidades robustas como extração de texto, imagens e metadados. Derivada do projeto Xpdf, ela se destaca por lidar eficientemente com documentos complexos, sendo uma ferramenta essencial para aplicações que exigem precisão na interpretação de PDFs. Pesquisas têm demonstrado o uso da Poppler em áreas como a análise de documentos científicos, permitindo a conversão confiável de PDFs em dados utilizáveis (Tiedemann, 2014).

Por exemplo, na tarefa de extrair transações financeiras de um extrato bancário digitalizado, a Poppler pode ser utilizada para converter o conteúdo em texto estruturado. Combinada com técnicas como Regex, é possível identificar padrões específicos, como valores monetários e datas, facilitando o processamento automático de grandes volumes de dados (Liu; Bai; Gao, 2011). Além disso, a Poppler tem sido utilizada em projetos que exigem extração de componentes lógicos de PDFs, como tabelas e gráficos, para aprimorar processos de mineração de dados e visualização (Heinz; Ilyushin; Markovich, 2004). Sua integração com sistemas de pré-processamento também demonstra alta eficiência na detecção de fronteiras lógicas em documentos digitais (Liu; Bai; Gao, 2011).

2.6 TECNOLOGIAS WEB PARA DESENVOLVIMENTO DE APIS E INTERFACES

O uso de tecnologias web modernas é essencial para a construção de APIs e interfaces robustas e escaláveis. No front-end, o React, um dos frameworks JavaScript mais populares, possibilita a criação de interfaces dinâmicas e responsivas, promovendo modularidade por meio de componentes reutilizáveis. Esse framework utiliza um modelo de Document Object Model (DOM) virtual, que otimiza a renderização de elementos dinâmicos, garantindo alta performance e interatividade em aplicações web complexas (Komperla et al., 2022). No backend, o Flask, um microframework amplamente utilizado em Python, desempenha um papel fundamental ao viabilizar a criação de servidores web leves e eficientes, facilitando a comunicação entre o servidor e as aplicações. Essa abordagem é especialmente eficaz em arquiteturas que processam dados financeiros de alta complexidade, como a conversão de documentos PDF para formatos padronizados como OFX (Skrzypiec; Plechawska-wójcik, 2023). Essa integração harmoniosa entre front-end e backend assegura a escalabilidade e a eficiência do sistema.

Os proxies reversos, como o NGINX, são componentes essenciais em arquiteturas modernas, atuando como intermediários entre os clientes e os servidores. Diferentemente de um proxy tradicional, que representa o cliente ao acessar recursos externos, um proxy reverso recebe todas as solicitações enviadas pelos clientes e as redireciona para os servidores internos apropriados. Entre suas principais funções estão o balanceamento de carga, que distribui o tráfego entre múltiplos servidores para evitar sobrecargas, e a implementação de cache, que reduz o tempo de resposta ao armazenar recursos acessados frequentemente. Além disso, o NGINX é amplamente utilizado para reforçar a segurança, fornecendo uma camada adicional de proteção contra ataques, como DDoS, e gerenciando certificados SSL para

criptografar os dados transmitidos (Madsen; Lhotak; Tip, 2020). Essas funcionalidades garantem maior resiliência e desempenho às aplicações, além de oferecer uma experiência mais fluida e segura aos usuários finais.

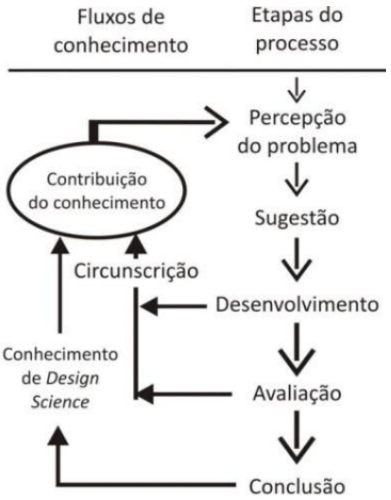
3 METODOLOGIA

Este estudo foi conduzido com a abordagem **Design Science Research (DSR)**, amplamente reconhecida por sua capacidade de resolver problemas complexos por meio da criação e validação de artefatos tecnológicos (Van aken, 2004). A DSR é uma metodologia que combina rigor acadêmico e aplicabilidade prática, sendo especialmente eficaz no desenvolvimento de sistemas para manipulação de dados, onde a precisão, eficiência e confiabilidade são fundamentais (Romme, 2003).

A abordagem segue um processo estruturado, que abrange desde a definição do problema até a avaliação do artefato, assegurando que as soluções sejam inovadoras e capazes de atender a demandas específicas (Hevner; Chatterjee, 2010). Conforme preceituado por Antony et al. (2023) em sua abordagem estruturada para resolver problemas complexos mal estruturados, esta pesquisa resultou num sistema modular que lida com a manipulação de dados de extratos bancários, incluindo a conversão de arquivos em formatos amplamente utilizados no setor financeiro, como o OFX, promovendo escalabilidade e consistência, conforme conceitos.

A Figura 1 apresenta o fluxo do Design Science Research (DSR), representando as etapas do processo e a retroalimentação do conhecimento ao longo do desenvolvimento.

Figura 1 – Fluxo do processo estruturado do Design Science Research (DSR)



Fonte: Adaptado de Rioga e Baracho (2017)

O processo inicia com a Percepção do Problema, na qual a necessidade é identificada e delimitada. Em seguida, ocorre a Sugestão de soluções potenciais, que são implementadas durante a fase de Desenvolvimento. Na etapa de Avaliação, o artefato é testado quanto à sua eficácia e aderência às necessidades identificadas, culminando na Conclusão, onde os resultados são consolidados e documentados.

Além disso, conforme demonstrado na Figura 1, o fluxo do conhecimento é retroalimentado por meio da contribuição do conhecimento, que conecta as etapas de Desenvolvimento e Avaliação. Essa dinâmica destaca a importância da circunscrição para o refinamento contínuo do artefato, permitindo a incorporação de melhorias com base nos aprendizados gerados ao longo do processo. O conhecimento de Design Science serve como base teórica e prática, assegurando que a solução proposta seja inovadora, consistente e escalável para atender às demandas específicas do setor financeiro.

Dessa forma, a pesquisa exemplifica a aplicação prática de um processo rigoroso e estruturado, resultando em um sistema modular eficiente, capaz de lidar com a manipulação e conversão de dados financeiros, como arquivos OFX, em consonância com as abordagens de Hevner e Chatterjee (2010) e Antony et al. (2023).

O quadro 1 apresenta o fluxo metodológico seguido neste estudo.

Quadro 1 - Percurso Metodológico

Passos Metodológicos	Descrição
(1) Tipo de textos usados.	Informações contidas em extratos bancários no formato PDF
(2) Fonte da qual esses dados foram extraídos.	Sistemas bancários ou exportados por usuários finais.
(3) Quantidade de dados analisados.	30 tipos de documentos diferentes. Ex: Só o Banco Brasil contém 5 tipos de modelos de extratos bancários diferentes.
(4) Software usado para captura, pré-processamento e análise.	Conversor e Regulador em C++; e Extrator em Python 3.12.
(5) Técnicas que foram usadas para capturar e pré-processar os dados.	Expressões Regulares (Regex) e Pandas (CSV) para capturar e pré-processar os dados.
(6) Técnicas que foram usadas para analisar os dados.	Expressões Regulares (Regex).

Fonte: Autores (2024)

O quadro 1 resume o percurso metodológico do estudo foi planejado para garantir o rigor e a eficácia do sistema proposto. Inicialmente, foram selecionados 30 tipos de documentos financeiros do tipo extrato bancário em PDF, abrangendo uma ampla variedade de formatos, desde arquivos gerados automaticamente por sistemas bancários até digitalizações feitas por usuários finais. Essa diversidade foi intencional para assegurar que o sistema pudesse lidar com diferentes estruturas de dados e características específicas de cada documento.

A Engenharia de Software foi incorporada como fundamento para estruturar o desenvolvimento do artefato, assegurando qualidade, consistência e escalabilidade desde as fases iniciais. Foram aplicados princípios de engenharia de requisitos para identificar

funcionalidades essenciais e restrições do sistema. Entre os requisitos funcionais, destacaram-se a necessidade de extração de informações financeiras de diferentes formatos documentais, a padronização dos dados em estruturas consistentes e a geração de arquivos em um padrão amplamente aceito pelo setor financeiro. Já os requisitos não funcionais abrangeram aspectos de desempenho, garantindo tempos de processamento reduzidos mesmo em grandes volumes de dados, segurança, assegurando a integridade das informações tratadas, e manutenibilidade, possibilitando que o sistema evoluísse de forma modular e controlada. Além disso, foram adotadas boas práticas de modelagem e arquitetura orientadas à modularidade, o que favoreceu o baixo acoplamento e a alta coesão entre os componentes, permitindo maior flexibilidade para ajustes futuros. O processo de desenvolvimento seguiu uma abordagem incremental, possibilitando evolução contínua e integração progressiva das soluções. Para garantir a confiabilidade, foram considerados princípios de qualidade de software, apoiados por versionamento de código e práticas de validação sistemática. Dessa forma, a Engenharia de Software serviu como alicerce metodológico e técnico, complementando a abordagem Design Science Research e contribuindo para que o artefato fosse construído de forma estruturada e alinhada a padrões reconhecidos da área.

3.1 OS DADOS COLETADOS E O PIPELINE DE PROCESSAMENTO

Os dados coletados foram submetidos a uma sequência de etapas, com cada módulo desempenhando um papel crítico no pipeline de processamento. O **Conversor de PDF para CSV** iniciou o processo, garantindo que informações textuais fossem extraídas com precisão. Em seguida, o **Regulador** ajustou os dados, corrigindo formatações e consolidando informações em uma estrutura uniforme, essencial para evitar erros em análises posteriores. Por fim, o **Extrator** aplicou Regex para localizar e organizar padrões específicos, como atributos de transações financeiras, convertendo-os em um formato OFX estruturado e padronizado. Isso pode ser ilustrado na Figura 2.

Figura 2 - Exemplo de Regex para Localizar Valores Monetários e Datas

Fonte: Autores (2024)

A Figura 2 ilustra alguns dos exemplos das Regex implementadas, detalhando padrões monetários e de datas, ao analisar os dados regulados, percebe-se que os valores monetários e datas apresentam padrões específicos. A Regex \b\d{1,3}(?:\.\d{3})*(?:,\d{2})\b captura valores como "1.234,56", enquanto \b\d{1,2}/\d{1,2}/\d{2,4}\b identifica formatos de datas

como "12/05/2023". Ambas foram implementadas no módulo Extrator, permitindo a transformação precisa das informações no formato exigido pelo padrão OFX.

Essa abordagem modular foi integrada em uma API robusta, desenvolvida com Flask e utilizando o Nginx como proxy reverso, conforme descrito em sua documentação oficial (Nginx, 2024). Essa arquitetura garantiu que o sistema fosse seguro, escalável e capaz de lidar com grandes volumes de dados financeiros. O percurso metodológico demonstrou como etapas interdependentes e bem definidas podem gerar soluções tecnológicas adaptadas às necessidades específicas do mercado financeiro

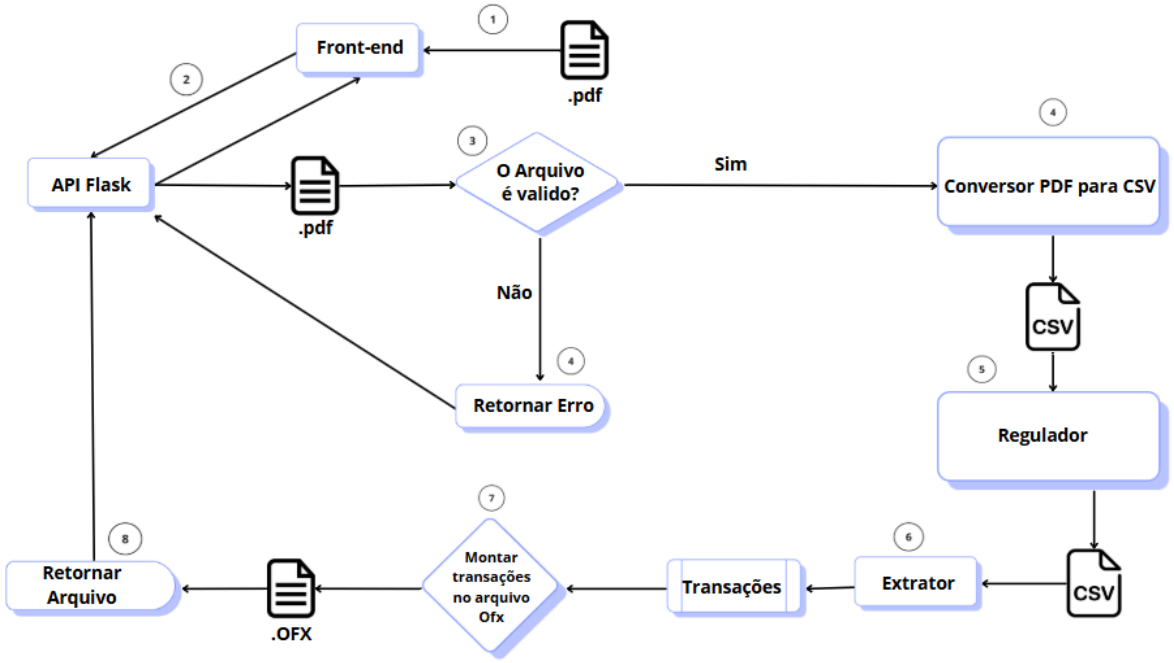
.

4 RESULTADOS

O resultado foi um método para extrair e estruturar informações financeiras de qualidade a partir de arquivos PDF, especificamente transações financeiras formatadas em OFX. A implementação deste método é realizada por meio de um algoritmo denominado **PDFtoOFX**, que utiliza 16 expressões regulares (Regex) diferentes para identificar e estruturar padrões como datas, valores monetários e descrições de transações.

A Figura 3 apresenta o fluxo do algoritmo, demonstrando as etapas de extração, padronização e conversão para o formato OFX, garantindo precisão e conformidade com os padrões do mercado financeiro.

Figura 3 - Fluxograma do Sistema PDFtoOFX

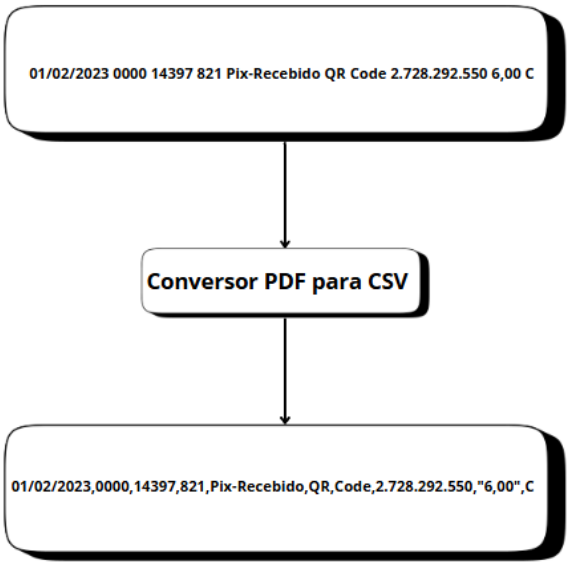


Fonte: Autores (2024)

A Figura 3 apresenta a arquitetura funcional dos módulos e o fluxo sequencial de processamento. O documento é inicialmente carregado no front-end e enviado, via API desenvolvida em Flask, para iniciar o processamento. O primeiro passo envolve a validação

do arquivo, garantindo que se trata de um documento PDF válido e sem erros. Após essa etapa, o pipeline avança para o primeiro módulo da estrutura o Conversor de PDF para CSV. Esse módulo é responsável por extrair os textos dos arquivos PDF e organizá-los em tabelas no formato CSV, marcando o início efetivo do processamento. O módulo Conversor de PDF para CSV foi o primeiro a ser acionado, iniciando o pipeline de processamento ao extrair os textos dos arquivos PDF e organizá-los em tabelas no formato CSV. Este módulo utilizou a biblioteca Poppler para interpretar os elementos textuais e preservar a estrutura original do documento, lidando com arquivos criptografados e digitalizados. Durante os testes, o Conversor apresentou desempenho consistente, processando arquivos grandes em menos de 20 segundos e mantendo a integridade dos dados extraídos. A Figura 4 ilustra uma linha de texto de um PDF processado pelo módulo, que converteu o conteúdo para o formato CSV, onde cada vírgula delimita uma coluna.

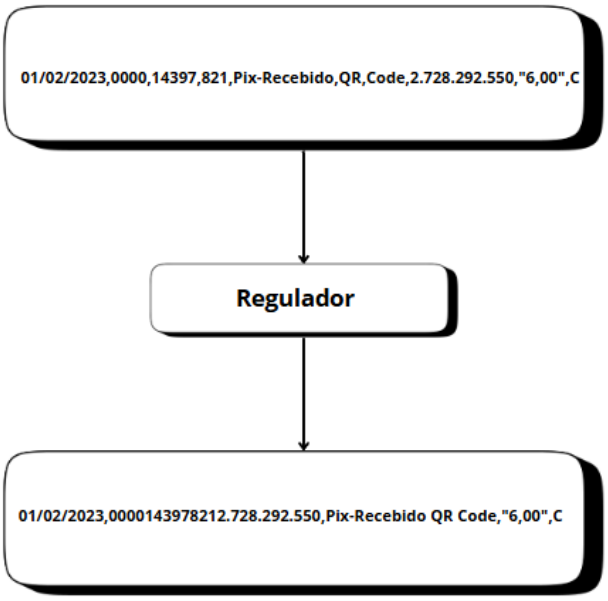
Figura 4 - Módulo conversor PDF para CSV



Fonte: Autores (2024)

O módulo Regulador, segunda etapa do pipeline, recebeu os arquivos CSV gerados pelo Conversor e aplicou técnicas avançadas de organização e padronização de dados. Ele foi projetado para identificar e corrigir formatações inconsistentes, consolidando informações dispersas em uma estrutura uniforme, essencial para evitar erros em análises posteriores. Durante os testes, o Regulador consolidou informações dispersas em células únicas e ajustou os cabeçalhos de maneira eficiente, resultando em arquivos organizados e prontos para a etapa final de extração. A Figura 5 ilustra o processamento da mesma linha apresentada na Figura 3, destacando como, no Regulador, as linhas do CSV são padronizadas e os tipos de dados realocados conforme as regras configuradas no módulo.

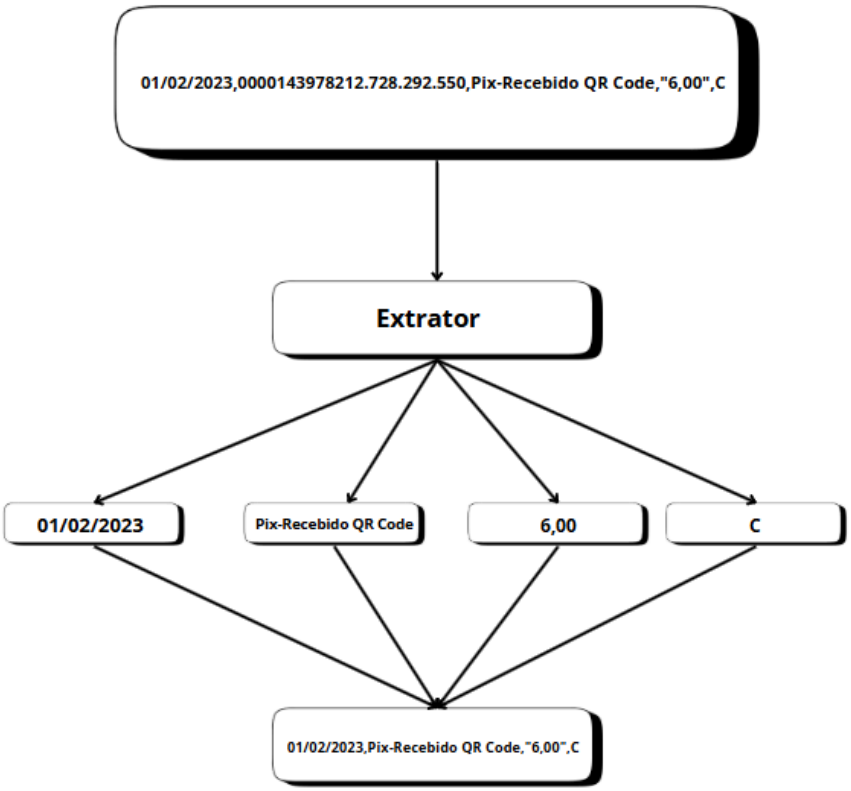
Figura 5 - Módulo Regulador



Fonte: Autores (2024)

Por fim, o módulo Extrator processou os dados regulados para convertê-los no formato OFX. Este módulo utilizou expressões regulares para localizar e capturar informações específicas, como datas em formatos variados (dd/mm/yyyy, dd-mm-yyyy), valores monetários (123,45+/-, 123.45C/D) e descrições textuais. Os dados foram organizados em transações financeiras no padrão OFX, cada uma contendo elementos como data, valor, descrição e tipo de transação. O Extrator foi testado em diferentes cenários e demonstrou alta precisão na identificação dos padrões, gerando arquivos OFX compatíveis com sistemas financeiros utilizados no mercado. A Figura 6 ilustra o funcionamento do módulo Extrator, mostrando como, a partir do CSV organizado pelo módulo anterior, ele extrai as informações necessárias para compor as transações no formato OFX, incluindo data, descrição, valor e tipo de transação.

Figura 6 - Módulo Extrator



Fonte: Autores (2024)

A integração dos módulos garantiu um fluxo sequencial eficiente, onde a saída de um módulo servia diretamente como entrada para o próximo. Essa arquitetura modular resultou em um tempo médio de processamento inferior a 40 segundos por arquivo, mesmo em casos de documentos volumosos com grandes quantidades de transações. Além disso, a modularidade do sistema facilitou a manutenção e permitiu escalabilidade, atendendo às demandas do mercado financeiro por soluções robustas e confiáveis (Gerardo; Allen; Botton, 2018). O sistema provou ser eficaz e escalável, com potencial para ser implementado em instituições financeiras para automação e padronização do processamento de dados financeiros.

4.1 DESEMPENHO DO MÉTODO

O sistema foi submetido a uma avaliação detalhada utilizando 400 arquivos de extratos bancários em PDF, organizados em 30 tipos diferentes. A distribuição dos arquivos incluiu: Banco do Brasil (25, 20, 15, 30 e 20 PDFs para os Tipos 1, 2, 3, 4 e 5, respectivamente), Caixa Econômica Federal (18, 12 e 10 PDFs para os Tipos 1, 2 e 3), Bradesco (22, 15 e 10 PDFs para os Tipos 1, 2 e 3), Itaú (20 e 15 PDFs para os Tipos 1 e 2) e Santander (18 e 12 PDFs para os Tipos 1 e 2). Além disso, foram analisados PDFs de Banco Inter (15), BTG Pactual (10), Banco Cora (8), PagSeguro (12), Mercado Pago (10 e 8 PDFs para os Tipos 1 e 2), Sicoob (14 e 11 PDFs para os Tipos 1 e 2) e Sicredi (12 e 10 PDFs para os Tipos 1 e 2). Outros bancos incluíram Banco do Nordeste (BNB) (10 PDFs), Banrisul (12 PDFs), Banco Original (10 PDFs), Banco Pan (8 PDFs) e Banco Votorantim (8 PDFs).

Desses, 368 PDFs foram convertidos com sucesso, enquanto os outros apresentaram inconsistências que afetaram a precisão da extração. As falhas ocorreram devido a dois

principais fatores: o reconhecimento de informações que não eram transações bancárias, como saldos diários sendo interpretados como transações, e problemas nos formatos das expressões regulares (Regex), que extraíram valores de cabeçalhos ou rodapés como parte dos dados financeiros.

Os resultados indicam que o PDFtoOFX é robusto e confiável, apresentando uma taxa de sucesso de 92% na conversão de arquivos PDF em transações financeiras estruturadas. Apesar das inconsistências identificadas, em nenhum caso ocorreram omissões de transações legítimas; os erros se limitaram à inclusão de informações adicionais, como saldos diários ou valores de cabeçalhos e rodapés.

Esse comportamento demonstra que o sistema prioriza a completude dos dados, garantindo que todas as transações relevantes sejam capturadas, mesmo que acompanhadas de registros supérfluos. Ao mesmo tempo, evidencia oportunidades de aprimoramento, incluindo o refinamento das expressões regulares e a implementação de filtros mais específicos para reduzir a presença de informações redundantes.

Para contornar essas inconsistências, algumas estratégias podem ser adotadas:

1. **Validação cruzada de registros:** comparar os dados extraídos com bases de referência ou arquivos anteriores para identificar e remover registros supérfluos.

2. **Filtros adaptativos e inteligentes:** aplicar critérios dinâmicos que ajustem a captura de dados conforme padrões identificados nos extratos, evitando a inclusão de informações irrelevantes.

3. **Aprimoramento contínuo das Regex:** revisar e otimizar expressões regulares para reduzir a captura de dados não transacionais, mantendo a cobertura das transações legítimas.

4. **Auditoria periódica e aprendizado incremental:** monitorar falhas recorrentes e ajustar regras do sistema com base nos aprendizados obtidos.

5. **Registro e classificação de exceções:** manter logs detalhados dos casos de inconsistência para análise futura e refinamento do sistema.

A arquitetura modular adotada, composta pelos módulos Conversor, Regulador e Extrator, favorece ajustes pontuais e manutenção contínua, permitindo que essas estratégias sejam implementadas sem comprometer o fluxo completo do processamento. Dessa forma, o PDFtoOFX mantém alta confiabilidade, precisão e escalabilidade, tornando-se adequado para aplicações financeiras que exigem automação e padronização de dados.

Vantagem em relação a métodos existentes: Diferentemente de ferramentas como **OFX Fácil (OFX Fácil, 2024)** e **Importefile (Importfile, 2024)**, que dependem de templates pré-ajustados para cada tipo de extrato, o PDFtoOFX apresenta **capacidade adaptativa**, conseguindo converter arquivos variados sem a necessidade de modelos predefinidos, aumentando flexibilidade e aplicabilidade em cenários com múltiplos formatos de extratos.

5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo desenvolver uma solução que permita a conversão eficiente e precisa de PDFs financeiros para o formato OFX, para isso usamos a metodologia do DSR Este estudo aplicou a abordagem Design Science Research (DSR) para criar um artefato tecnológico que resolve problemas complexos relacionados ao processamento de dados financeiros, destacando a eficácia de uma metodologia estruturada e modular para alcançar resultados confiáveis e replicáveis.

O resultado foi o desenvolvimento e a implementação do sistema **PDFtoOFX** demonstraram ser uma solução robusta e escalável para a conversão automatizada de dados financeiros de arquivos PDF para o formato OFX, atendendo às demandas do mercado financeiro por precisão e padronização.

Os resultados dos testes revelaram uma taxa de sucesso de conversão de **92%**, com 368 arquivos PDF processados corretamente em um total de 400. Embora algumas inconsistências tenham sido identificadas – como a interpretação incorreta de saldos diários ou valores de cabeçalhos como transações – é relevante ressaltar que essas falhas nunca comprometeram a integridade dos dados essenciais. Todos os registros de transações legítimas foram mantidos, ainda que acompanhados de dados adicionais supérfluos. Esse comportamento garante que o sistema priorize a completude das informações, um aspecto essencial em aplicações financeiras.

A arquitetura modular do sistema, composta pelos módulos Conversor, Regulador e Extrator, foi essencial para garantir a eficiência do pipeline de processamento. Cada módulo desempenhou um papel crítico na transformação dos dados, desde a extração inicial dos textos até a conversão final em transações financeiras estruturadas no padrão OFX. O uso de expressões regulares (Regex) permitiu uma identificação precisa de padrões, enquanto a integração dos módulos em uma API robusta assegurou desempenho e escalabilidade em diferentes cenários de aplicação.

Apesar do sucesso, o estudo também revelou áreas de melhoria, como o refinamento das Regex e a inclusão de filtros mais específicos para evitar registros redundantes. Essas melhorias podem aumentar ainda mais a precisão do sistema, tornando-o ainda mais confiável para aplicações financeiras de grande escala. Além disso, o desempenho observado, com tempos médios de processamento inferiores a 40 segundos por arquivo, demonstra o potencial do sistema para ser adotado em instituições financeiras, oferecendo uma solução eficiente e automatizada para o tratamento de dados financeiros.

Por fim, o PDFtoOFX não apenas atingiu os objetivos propostos, mas também estabeleceu uma base sólida para futuras evoluções tecnológicas. Este sistema representa um passo significativo na automação e padronização do processamento de dados financeiros, promovendo maior eficiência operacional e confiabilidade na gestão de informações financeiras.

REFERÊNCIAS

ALMEHREJ, A.; FREITAS, L.; MODESTI, P. ***Account and Transaction Protocol of the Open Banking Standard***. 2020. Disponível em: https://doi.org/10.1007/978-3-030-48077-6_16. Acesso em: 2024.

AMMARI, F. T. ***Securing financial XML transactions using intelligent fuzzy classification techniques***. 2013. Disponível em: https://eprints.hud.ac.uk/id/eprint/19506/1/Faisal_Ammari_-_Final_Thesis.pdf. Acesso em: 2024.

ANTONY, J.; SONY, M.; LAMEIJER, B. A.; BHAT, S.; JAYARAMAN, R. ***Towards a design science research (DSR) methodology for operational excellence (OPEX) initiatives***. *The TQM Journal*, 2023. Disponível em: https://doi.org/10.1108/tqm-01-2023-0017. Acesso em: 2024.

BEYRIES, G.; RODRIGUEZ, F. ***Technical Report: CSVM format for scientific tabular data***. 2012. Disponível em: https://arxiv.org/abs/1207.5711. Acesso em: 2024.

CHEPAKOV, D. A. ***Standardization of financial operations: ISO 20022 and operational aspects of its practical implementation***. 2020. Disponível em: https://doi.org/10.25198/2077-7175-2020-1-51. Acesso em: 2024.

CONSTANTIN, A.; PETTIFER, S.; VORONKOV, A. ***PDFX: Fully-automated PDF-to-XML conversion of scientific literature***. *Document Engineering*, 2013. Disponível em: https://doi.org/10.1145/2494266.2494271. Acesso em: 2024.

GERARDO, U.; ALLEN, R. I.; BOTTON, N. M. ***How to Design a Financial Management Information System: A Modular Approach***. IMF Fiscal Affairs Department, International Monetary Fund, 2019. Disponível em: https://ideas.repec.org/p/imf/imfhtn/2019-003.html. Acesso em: 2024.

HEINZ, O.; ILYUSHIN, B.; MARKOVICH, D. M. ***Application of a PDF method for the statistical processing of experimental data***. 2004. Disponível em: https://doi.org/10.1016/J.IJHEATFLUIDFLOW.2004.05.009. Acesso em: 2024.

HEVNER, A. R.; CHATTERJEE, S. ***Introduction to Design Science Research.*** 2010. Disponível em: https://doi.org/10.1007/978-1-4419-5653-8_1. Acesso em: 2024.

IMPORTFILE. ***Solução para conversão de arquivos para o formato OFX***. [S. l.: s. n.], 2024. Disponível em: https://importefile.com.br/. Acesso em: 2024.

KOMPERLA, V.; PRATIBA, D.; GHULI, P.; PATTAR, R. ***React: A detailed survey. Indonesian Journal of Electrical Engineering and Computer Science***, v. 26, n. 3, p. 1710–1717, 2022. Disponível em: https://doi.org/10.11591/ijeecs.v26.i3.pp1710-1717. Acesso em: 2024.

KRUTTSCHNITT, J. ***Automatic conversion of table contents from PDF technical specification documents into database using AI Optical Character Recognition (OCR)***. 2022. Disponível em: https://doi.org/10.1007/978-981-19-3951-8_22. Acesso em: 2024.

LIU, Y.; BAI, K.; GAO, L. ***An efficient pre-processing method to identify logical components from PDF documents***. 2011. Disponível em: https://doi.org/10.1007/978-3-642-20841-6_41. Acesso em: 2024.

MAGGI, S.; VICARIO, S. ***Automated Extraction of Bioclimatic Time Series from PDF Tables***. 2023. Disponível em: https://doi.org/10.5194/egusphere-egu23-15072. Acesso em: 2024.

MADSEN, M.; LHOTAK, O.; TIP, F. ***A semantics for the essence of React. European Conference on Object-Oriented Programming***, 2020. Disponível em: https://par.nsf.gov/servlets/purl/10157540. Acesso em: 2024.

MALIK, H. H.; BHARDWAJ, V.; FIORLETTA, H. ***Accurate information extraction for quantitative financial events***. 2011. Disponível em: https://doi.org/10.1145/2063576.2064001. Acesso em: 2024.

MURTAZA, S. S.; NIE, Y.; LIN, J. ***Regex-augmented Domain Transfer Topic Classification based on a Pre-trained Language Model: An application in Financial Domain***. 2023. Disponível em: https://doi.org/10.48550/arXiv.2305.18324. Acesso em: 2024.

NGINX. ***Proxy reverso NGINX | Documentação NGINX***. [S. l.: s. n.], 2024. Disponível em: https://docs.nginx.com/nginx/admin-guide/web-server/reverse-proxy/. Acesso em: 2024.

NIKOLIĆ, L. ***Challenges of digitalization of financial transactions***. 2023. Disponível em: https://doi.org/10.5937/zrpfn1-44231. Acesso em: 2024.

OFX FÁCIL. ***Conversão de arquivos para o formato OFX***. [S. l.: s. n.], 2024. Disponível em: https://www.ofxfacil.com.br/. Acesso em: 11 dez. 2024.

RIOGA, D.; BARACHO, R. M. A. ***A gestão da informação aplicada ao processo de internacionalização universitária: um estudo de caso da UFMG***. *Pesquisa Brasileira em Ciência da Informação e Biblioteconomia*, João Pessoa, v. 12, n. 1, p. 1–15, abr. 2017. Disponível em: https://www.researchgate.net/publication/332409515. Acesso em: 2024.

ROMME, A. G. L. ***Making a Difference: Organization as Design***. *Organization Science*, v. 14, n. 5, p. 558–573, 2003. Acesso em: 2024.

SHERE, R.; CHITTIMALLI, P. K.; NAIK, R. L. ***Identifying and Extracting Hierarchical Information from Business PDF Documents***. 2022. Disponível em: https://doi.org/10.1145/3511430.3511440. Acesso em: 2024.

SKRZYPIEC, S.; PLECHAWSKA-WÓJCIK, M. ***Comparative analysis of Angular and React development frameworks***. *Journal of Computer Sciences Institute*, 2023. Disponível em: https://doi.org/10.35784/jcsi.3724. Acesso em: 2024.

TIEDEMANN, J. ***Improved Text Extraction from PDF Documents for Large-Scale Natural Language Processing***. 2014. Disponível em: https://doi.org/10.1007/978-3-642-54906-9_9. Acesso em: 2024.

VAN AKEN, J. E. ***Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules***. *The Journal of Management Studies*, v. 41, n. 2, p. 219–246, 2004. Acesso em: 2024.

VAN DEN BURG, G. J. J.; NAZABAL, A.; SUTTON, C. ***Wrangling Messy CSV Files by Detecting Row and Type Patterns***. 2018. Disponível em: https://arxiv.org/abs/1811.11242. Acesso em: 2024.

VOVCHENKO, N. et al. ***Information and financial technologies in a system of Russian banks' digitalization: A competency-based approach***. 2018. Disponível em: https://dx.doi.org/10.1108/S1569-375920180000100004. Acesso em: 2024.

ZHU, M.; COLE, J. M. ***PDFDataExtractor: A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format***. 2022. Disponível em: https://doi.org/10.1021/acs.jcim.1c01198. Acesso em: 2024.