IMPLEMENTAÇÃO DE INFRAESTRUTURA DE API PARA EXTRAÇÃO DE DADOS TABULARES A PARTIR DE DOCUMENTOS PDF

ARTHUR PEREIRA ROZADO[1], ANDREIWID SH. CORRÊA[2]

[1] Graduando em Tecnologia em Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, IFSP, Câmpus Campinas, arthurprozado@gmail.com
[2] Professor do Câmpus Campinas, andreiwid@ifsp.edu.br
Área de conhecimento (Tabela CNPq): Sistemas de Computação – 1.03.04.00-2

Apresentado no
7° Congresso de Iniciação Científica e Tecnológica do IFSP
29 de novembro a 02 de dezembro de 2016 - Matão-SP, Brasil

RESUMO: O movimento de dados abertos vem sendo consolidado nos últimos anos para definir requisitos para promover uso, a reutilização e redistribuição dos dados por qualquer e para qualquer propósito. A área governamental tomou frente neste movimento com iniciativas de divulgar informações de transparência atendendo aos requisitos de dados abertos com apoio das legislações específicas. O problema é que o atendimento dos requisitos de dados abertos é algo que demanda tempo e preparação dos agentes públicos. Com isso, tem-se informações sendo divulgadas, e não dados, o que compromete os benefícios pretendidos. Um dos principais formatos preferidos é o *Portable Document Format* (PDF) indicado somente para leitura humana. Este trabalho objetiva implementar uma infraestrutura composta de *Application Programming Interfaces* (APIs) para extração de dados tabulares e convertê-los em formatos compatíveis com dados abertos. Com os resultados deste trabalho, a comunidade poderá utilizar as interfaces disponibilizadas para utilização por outros sistemas sem limitações de linguagens e tecnologias.

PALAVRAS-CHAVE: Dados abertos; Tabula; CSV.

IMPLEMENTATION OF API INFRASTRUCTURE TO EXTRACT TABULAR DATA FROM PDF DOCUMENTS

ABSTRACT: The open data movement has been consolidated in the last years and aim to define requirements to promote use, reuse and redistribution of data by anyone and for any purpose. The government sector took part of this movement with initiatives to disclose information regarding transparency which meets the open data requirements and specific legislation. The problem is that the call of the open data requirements is something that takes time and preparation of public sector. Thus, there is information being disclosed, and not data, which undermines the intended benefits. One of the main preferred format is the Portable Document Format (PDF) suitable for human reading. This work aims to implement an infrastructure with the use of Application Programming Interfaces (APIs) for tabular data extraction and convert them into formats compatible with open data. With the results of this work, the community can use the interfaces available for use by other systems without limitations of languages and technologies.

KEYWORDS: Open data; Tabula; CSV.

INTRODUÇÃO

O movimento de dados abertos propõe uma série de requisitos para guiar a abertura de registros públicos com o uso de infraestrutura específica de software. Além da literatura disponível, existe toda uma legislação própria para tratar do assunto, como exemplo a Lei de Acesso à Informação (Lei nº 12.527/2011) em vigência no cenário brasileiro. Para que a disponibilização de dados seja compatível com dados abertos, deve-se seguir requisitos previamente definidos para que os dados (provenientes da transparência pública) possam ser livremente usados, reutilizados e redistribuídos, por qualquer um, para qualquer propósito (OPEN KNOWLEDGE FOUNDATION, 2012; TAUBERER, 2014).

Porém, apesar dos potenciais benefícios com a disponibilização de dados abertos, Coelho *et al.* (2015) revelam que a publicação de dados governamentais compatíveis com dados abertos no Brasil ainda é incipiente. Percebe-se a proliferação de websites de transparência na contramão dos dados abertos. Estes websites mostram-se verdadeiros repositórios de documentos semelhantes a relatórios impressos, normalmente viabilizados através de PDF e HTML (CORRÊA; CORRÊA; DA SILVA, 2014).

Este trabalho objetiva implementar uma infraestrutura composta de APIs (*Application Programming Interfaces*) que forneçam serviços de extração de dados tabulares contidos em documentos PDFs, que é um dos principais formatos preferidos para disponibilização de documentos. Os dados extraídos são convertidos para o formato CSV (*Comma-Separated Values*), amplamente conhecido, universal e compatível com dados abertos. Acredita-se que a utilização do CSV é um passo para que os dados possam ser livremente usados e redistribuídos para qualquer propósito.

MATERIAL E MÉTODOS

Este trabalho de iniciação científica tem como abordagem a implementação da infraestrutura de APIs a partir das diretrizes propostas por Corrêa, Corrêa e da Silva (2015). Neste os autores definem uma arquitetura de referência organizada em camadas para promover a estruturação de dados a partir de uma abordagem colaborativa, que á a possibilidade de qualquer pessoa fazer a extração e conversão de dados. A Figura 1 apresenta esta arquitetura cuja ênfase é dada para a Camada de Estruturação, onde se encontram os serviços de extração dos dados, conversão para CSV e interação com a Camada de Apresentação.

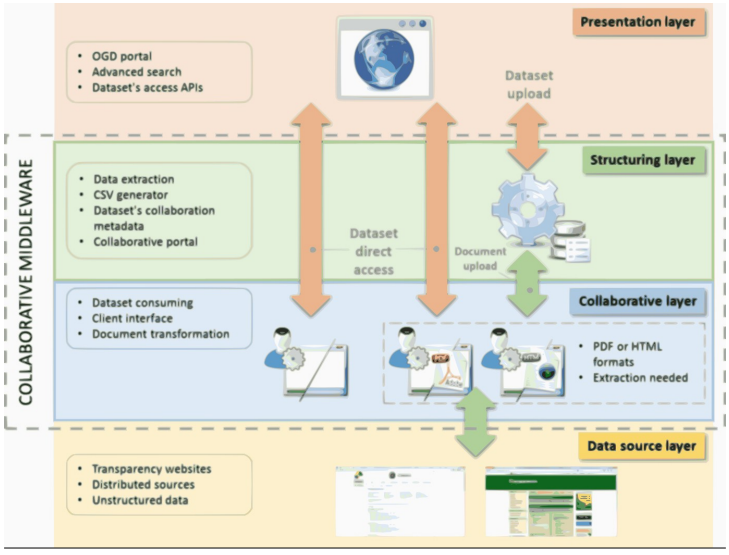

FIGURA 1. Arquitetura para estruturação de dados proposta por Corrêa, Corrêa e da Silva (2015).

Os principais materiais utilizados foram os seguintes:

- Ambiente de desenvolvimento baseado em Java composto pelo servidor web Apache Tomcat para Java Servlet, pela IDE Eclipse e pelo banco de dados MySQL, todos configurados no ambiente local do desenvolvedor.
- Instância do software livre e de código aberto CKAN (http://ckan.org) para servir de repositório na Camada de Apresentação
- Componente livre e de código aberto Tabula PDF (https://github.com/tabulapdf/tabula) para extração de dados de arquivos PDFs.

- Bibliotecas para manipulação de arquivos JSON.

A Figura 2 a seguir descreve o método utilizado para funcionamento das APIs evidenciando o fluxo de informações dentro do contexto dos serviços:

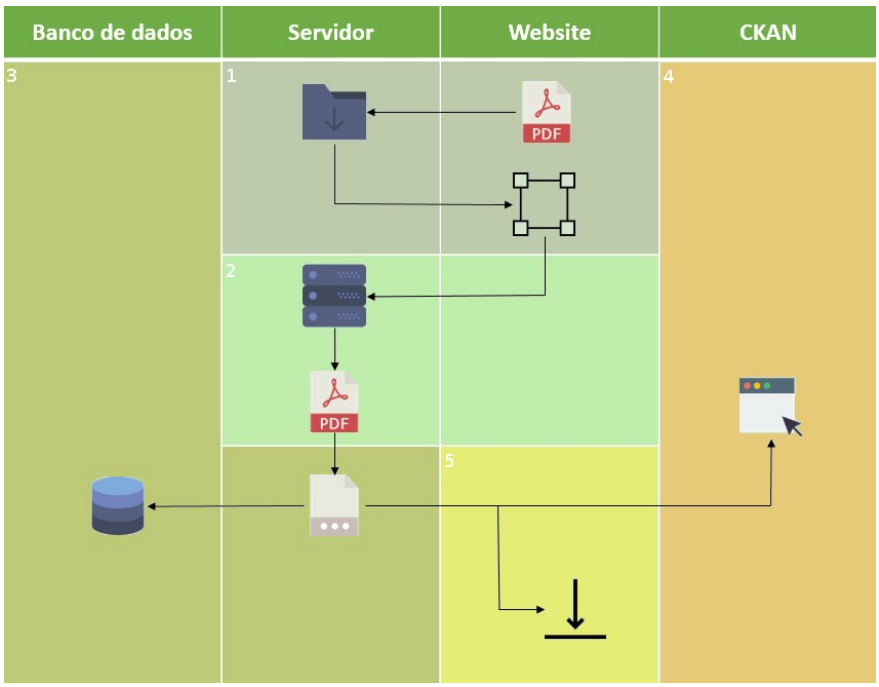

FIGURA 2. Fluxo de funcionamento das APIs.

1. O usuário passa como parâmetros pelo site o arquivo PDF e as coordenadas das áreas tabulares desejadas. O servidor armazena o arquivo em uma pasta temporária para continuidade do processo.
2. O sistema faz uma chamada do componente Tabula e retransmite os parâmetros de área tabulada e documento PDF passados pelo usuário ao componente, o qual acessa o arquivo PDF na pasta temporária.
3. O Tabula extrai os dados se baseando na área indicada pelo usuário e devolve um arquivo CSV como saída, que contem as informações da área tabulada do PDF e que posteriormente é armazenado em uma instância do banco de dados MySQL juntamente com demais metadados de interesse.
4. O arquivo CSV é então serializado em JSON e enviado para a instância do CKAN via API, onde lá é externado publicamente de forma compatível com dados abertos.
5. Com o processo finalizado, é possível ter acesso ao identificador único gerado pelo CKAN que identificará o dataset na Camada de Apresentação.

RESULTADOS E DISCUSSÃO

A partir dos materiais e métodos, foi implementado o sistema de conversão com a utilização do Tabula, armazenamento em banco de dados via MySQL e o envio ao CKAN utilizando a API CKAN. A Tabela 1 a seguir lista as APIs que foram implementadas, os parâmetros de execução e breve descrição de seu funcionamento.

TABELA 1. Resumo das APIs implementadas.

API	Parâmetros de entrada	Saída	Descrição
Extração dos dados e conversão para CSV	File pdf, int top, int left, int bottom, int right, int[] pages	File csv	API para converter o arquivo PDF em CSV. Os parâmetros utilizados são o pdf, a área tabulada do pdf, e as páginas das quais a tabela faz parte.

Connector com MySQL	File csv	String iddb	API para fazer a conexão entre a instância do MySQL e o software. Armazena os arquivos convertidos no banco de dados para acessos futuros.
Interação com a Camada de Apresentação	File csv String iddb	File json String idckan	API para enviar os arquivos convertidos para o CKAN. O arquivo JSON é enviado para o CKAN e disponibilizado online para acesso do usuário.

O código-fonte desenvolvido foi disponibilizado à comunidade pelo GitHub através do endereço https://github.com/ArthurPRozado/ifsp_open_data.

CONCLUSÕES

Ferramentas de fácil uso para extração de dados a partir de PDFs podem ser caras e normalmente funcionam de modo genérico, pois em sua maior parte, são direcionadas para todo tipo de aplicação. Ferramentas como “Nitro PDT to Excel” tem um custo alto, e outras como “Tabula” oferecem dificuldades ao usuário para a utilização. Neste sentido, a extração específica de dados tabulares contidos em PDF não é uma tarefa trivial de ser alcançada por qualquer usuário.

A implementação desta infraestrutura de APIs objetivou viabilizar ferramenta específica para extração de dados tabulares que tem seu funcionamento direcionado à comunidade de dados abertos. Com isso, qualquer usuário poderá valer-se dos serviços oferecidos pelas APIs de modo simples e rápido. Ademais, com o fornecimento de APIs, a comunidade poderá integrar outros sistemas com os serviços disponibilizados, promovendo assim o uso de qualquer linguagem e tecnologia.

Por fim, o uso desta infraestrutura de APIs é esperado também pela área governamental que ainda tem dificuldades em abrir seus dados e se desvencilhar dos PDFs. Dessa forma, poderá servir de transição até que os processos de geração de dados sejam compatíveis com dados abertos desde a geração.

AGRADECIMENTOS

Os autores agradecem ao IFSP por apoiar este projeto e por conceder bolsa de iniciação científica ao discente envolvido.

REFERÊNCIAS

COELHO, T. R. et al. **Transparência governamental nos estados e grandes municípios brasileiros: uma “dança dos sete véus” incompleta?** . In: XXXIX ENANPAD 2015. Belo Horizonte, MG: 2015

CORRÊA, A. S.; CORRÊA, P. L. P.; DA SILVA, F. S. C. **Transparency Portals Versus Open Government Data: An Assessment of Openness in Brazilian Municipalities**. Proceedings of the 15th Annual International Conference on Digital Government Research. **Anais**...: dg.o ’14.New York, NY, USA: ACM, 2014Disponível em: <http://doi.acm.org/10.1145/2612733.2612760>. Acesso em: 10 out. 2014

CORRÊA, A. S.; CORRÊA, P. L. P.; DA SILVA, F. S. C. **A Collaborative-oriented Middleware for Structuring Information to Open Government Data**. Proceedings of the 16th Annual International Conference on Digital Government Research. **Anais**...: dg.o ’15.New York, NY, USA: ACM, 2015Disponível em: <http://doi.acm.org/10.1145/2757401.2757409>. Acesso em: 11 jun. 2015

OPEN KNOWLEDGE FOUNDATION. **Open Data Handbook Documentation**, 14 nov. 2012. Disponível em: <http://opendatahandbook.org/>. Acesso em: 18 nov. 2014

TAUBERER, J. Open Government Data: The Book - Second Edition, 2014. Disponível em: <https://opengovdata.io/>. Acesso em: 18 nov. 2014