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

Arthur Pereira Rozado ; Andreiwid Sheffer Corrêa;
IFSP – *Campus* Campinas;

Objetivo

O projeto objetiva utilizar APIs (*Application Programming Interfaces*) já disponiveis para converter PDFs, que é um dos principais formatos utilizados para disponibilização de documentos. Os dados extraídos são convertidos para CSV (*Comma-Separated Values*), que é um formato amplamente conhecido, universal e compatível com dados abertos.

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. 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, pesquisas revelam que a publicação de dados governamentais compatíveis com dados abertos no Brasil ainda é incipiente. 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.

Materiais e Métodos

A infraestrutura de APIs implementada tem como base o trabalho iniciado por Corrêa, Corrêa e da Silva (2015). No projeto, é difinido uma arquitetura 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, e desta forma, permitir a propagação de dados de forma livre.

Resultados preliminares

A função de extração foi implementada utilizando a biblioteca Tabula, e retorna um CSV da conversão. Também, já possívelel enviar dados ao CKAN, podendo assim já ser utilizado em ambiente de teste. Como próximo objetivo o desenvolvimento de uma interface gráfica responsável por facilitar a utilização do serviço pelo consumidor final indispensável.

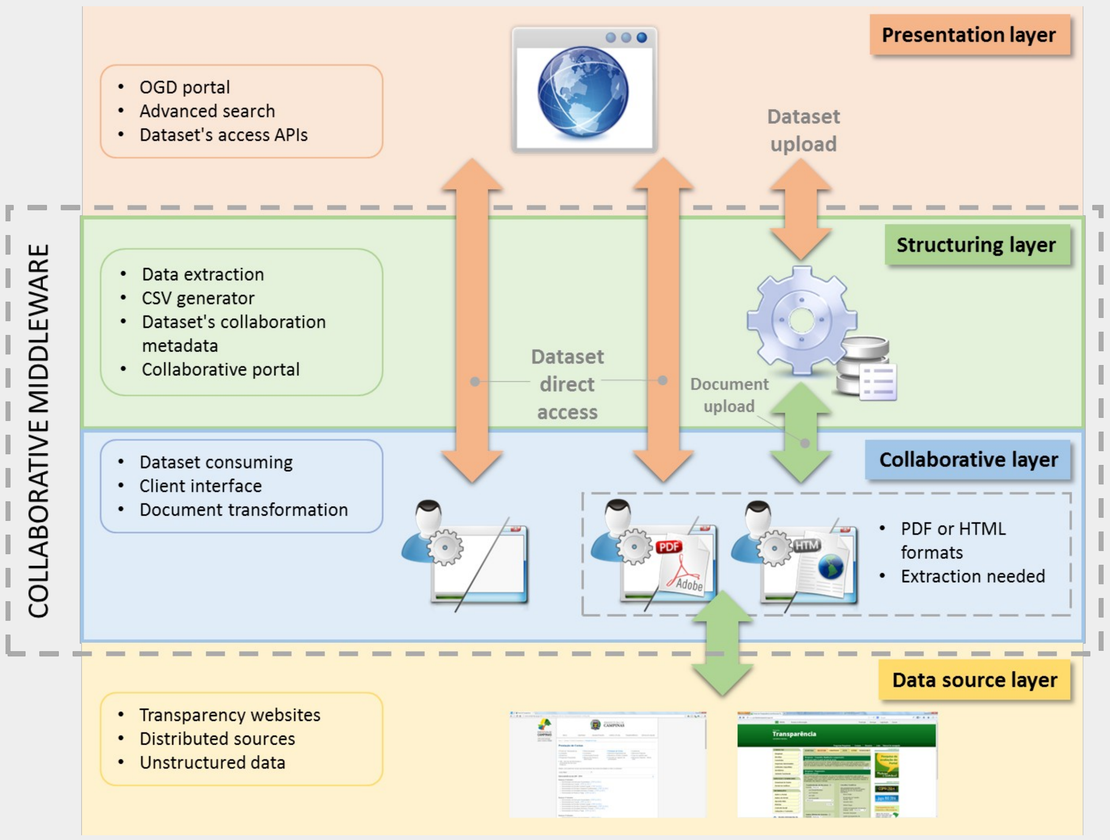

Figura 1: Arquitetura proposta por Corrêa, Corrêa e da Silva (2015).

Conclusão

A implementação desta infraestrutura de APIs objetivou viabilizar ferramentas específicas para extração de dados tabulares com base em uma arquitetura em camadas e com funcionamento colaborativo. A partir dos testes preliminares, foi possível extrair conteúdos em HTML e PDF e torná-los acessíveis sem as restrições de processamento inerentes aos formatos não compatíveis com dados abertos.

Referências

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