FERRAMENTA PARA EXTRAÇÃO DE INFORMAÇÕES DE PORTAIS DE TRANSPARÊNCIA PÚBLICA DISPOSTAS EM FORMATO NÃO ABERTO

DEVELOPMENT OF A TOOL EXTRACTION OF TRANSPARENCY WEB SITES INFORMATION AVALIABLE IN NON-OPEN FORMAT

Evandro Henrique Couto de Paula
Análise e Desenvolvimento de Sistemas
Análise e Desenvolvimento de Sistemas
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo

Paulo Henrique Pereira Cardoso
Análise e Desenvolvimento de Sistemas
Análise e Desenvolvimento de Sistemas
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo

Andreiwid Sheffer Corrêa
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo
Análise e Desenvolvimento de Sistemas
Análise e Desenvolvimento de Sistemas

O trabalho desenvolvido busca criar uma ferramenta que seja capaz de coletar as informações e documentos dos portais de transparência dos municípios brasileiros que se encontram em formatos não abertos ou não manipuláveis por máquinas, especialmente o HTML e PDF, formatos estes muito utilizados pelos municípios pela facilidade encontrada para se trabalhar com os mesmos, porém dificultando muito a manipulação e reutilização dos dados presentes nos documentos. A ferramenta tem como objetivo manipular e extrair dados abertos governamentais encontrados em formatos PDF e HTML, com foco em dados tabulares, fazendo uso da linguagem de programação Java e bibliotecas externas JSOUP e PDFBox, para que desta forma as informações possam ser manipuladas e extraídas dos documentos e devolvidas em formato CSV, contrastando com os formatos antes não abertos.

Palavras-chave: Dados abertos. Extração. Java

This work describe the process of development of a tool that collect information and documents of municipalities transparency web sites that are found in non-open or non-machine readable formats in especial HTML and PDF. These formats are common used because of the convenience for the municipal administration to generate them, however this practice impose a barrier to work this documents. This tool has the objective of manipulate and extract the data present in PDF and HTML documents witch focus only table data. Using Java language and the external libraries JSoup and PDFBox, enabling the application to manipulate and extract data also generate a CSV file for the user.
Keywords: Open Data. Extraction. Java.

1 INTRODUÇÃO

Os dados públicos são de domínio da sociedade e seu acesso é garantido por lei. A lei 12.527/11 também conhecida como Lei de Acesso à Informação que estabelece padrões para que os governos municipal, estadual e federal façam a publicidade de suas informações e garantam a transparência da gestão e possibilite que população possa conhecer melhor seu governo.
A lei prevê que as informações estejam a disposição à população, e que o meio preferencial para a exposição destes dados é a internet através dos portais de transparência dos órgãos envolvidos. A imaturidade dos municípios em relação aos requisitos da Lei, e a exigências dos princípios de dados abertos, no entanto, não garante a transparência dos governos (CORRÊA, A; CORRÊA, P; SILVA,2014).
Os dados abertos são aqueles que são dispostos em formato aberto e bruto, compreensíveis por máquina e que não possuam restrições de uso, e que podem ser reutilizados pela sociedade nas mais diversas aplicações (W3C, 2010).
Segundo W3C (2011), atualmente nos portais de transparência dos municípios muitas informações são dispostas em formatos não abertos como o PDF ou proprietários como o xls. Estes formatos não atendem os princípios propostos de dados abertos uma vez que podem causar limitações tecnológicas de acesso à informação.
Neste contexto existem outros formatos indicados como ideais por W3C (2011) como por exemplo o CSV e RDF. O foco deste trabalho é a extração de dados tabulares e sua conversão em formato CSV.
A obtenção destes dados em formatos utilizáveis não é tão simples, quando procurados esbarra-se em limitações tecnológicas e a extração da informação acontece através do trabalho manual da pessoa interessada. Quando o interessado busca uma alternativa para automatizar a obtenção da informação descobre-se que existem ferramentas que realizam estas atividades, muitas destas, porém são ferramentas pagas ou que possuem alguma restrição de utilização, não facilitando desta forma a obtenção dos dados.
O presente trabalho busca detalhar o processo de desenvolvimento de uma aplicação para coleta de informações e extração de dados abertos dos portais de transparência pública utilizando a linguagem de programação java e bibliotecas externas PDFBox e Jsoup.
Para a extração dos dados presentes em documentos PDF a ferramenta utilizará um webservice, neste serviço estará funcionando a ferramenta Tabula, ferramenta gratuita e de código aberto criada pelos jornalistas Manuel Aristarán, Mike Tigas e Jeremy B. Merrill e com suporte de várias organizações, que tem como objetivo a extração de dados presentes em documentos PDF.
Muito embora possua uma versão web, esta versão será descontinuada, por esta razão o trabalho buscará utilizar a API Java da aplicação para integrá-la ao webservice para que desta forma ainda seja possível utilizar esta útil ferramenta, uma vez que a API Java deve receber suporte da comunidade.
A biblioteca Jsoup é uma biblioteca open source capaz de acessar a página HTML e extrair informações destas páginas, enquanto a biblioteca PDFBox é uma biblioteca criada pela apache e é uma solução muito boa para a manipulação de documentos PDF (PARK et al, 2010).

2 METODOLOGIA

O desenvolvimento da aplicação se dá utilizando a linguagem de programação Java, e utilização de bibliotecas externas PDFBox e Jsoup e é baseado nos seguintes processos: coleta de

informações de documentos PDF, coleta de informações de páginas HTML e desenvolvimento de interface para usuário, a inserção do applet Java em uma página HTML.
Basicamente o processo de extração do PDF consiste no download do documento localizado portal de transparência do munícipio usando como referência a url do documento PDF.
Em posse do documento e utilizando a biblioteca PDFBox o documento é convertido em um formato de imagem onde o usuário poderá selecionar uma área específica, armazenado estas coordenadas será possível o envio destas informações para o webservice que realizará a extração da informação e devolverá um documento em formato aberto.
A outra funcionalidade da aplicação que é extração de dados HTML em formato tabular, nesta funcionalidade será utilizada a biblioteca Jsoup, ferramenta capaz de manipular os elementos presentes na estrutura da página do portal de transparência do órgão.
A extração se dará tendo como base a url da página em que se deseja extrair a informação buscada, a referência para extração serão as tags que compõem as tabelas de informações das páginas, para que desta forma seja possível montar um documento em formato aberto e que seja útil para o usuário interessado.

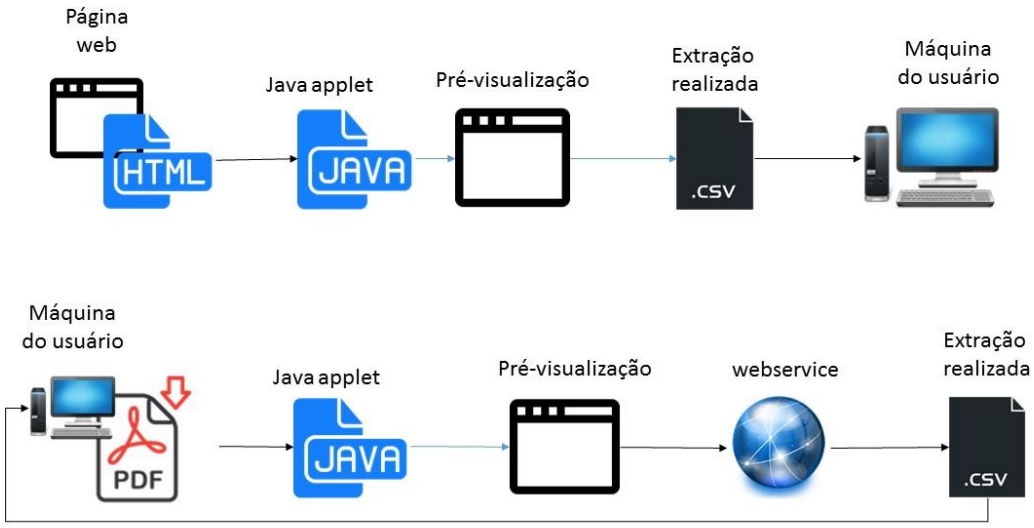

Figura 1 - Esquema de extração

A outro aspecto importante é o desenvolvimento de uma interface agradável e que facilite a utilização da aplicação pelo usuário, para isto serão utilizados os elementos do java swing, para que se desenvolva uma interface já conhecida e agradável para o usuário.
Uma interface agradável é necessária para que a utilização da aplicação seja fácil e o usuário se sinta interessado em utilizar novamente a aplicação.
Integrado a estes processos está a inclusão deste applet Java à uma página HTML, para que desta forma o usuário possa manipular apenas a página, tornando toda a manipulação da ferramenta integrado ao ambiente web.

3 RESULTADOS

Até o presente momento a manipulação do PDF tem sido bastante trabalhosa, para a uma possível utilização deste formato para a ferramenta iniciou-se o trabalho baseando-se em

tecnologias web, porém inúmeros empecilhos e limitações técnicas obrigaram a alteração da tecnologia aplicada na ferramenta, resultando em uma mudança nos cronogramas iniciais.
A tecnologia Java trouxe uma nova perspectiva no desenvolvimento da aplicação, de imediato o tratamento do documento PDF se tornou mais plausível, e ainda passa por trabalhos para futuros testes de integração da ferramenta com webservice, e testes efetivos de extração.
No âmbito da tecnologia HTML a utilização da biblioteca JSOUP, possibilitou a extração de algumas páginas teste, porém apenas é possível extrair páginas que utilizam a protocolo GET em suas requisições, páginas que utilizam requisições POST para obter informações não retornam resultados, outro caso específico é a não padronização da estrutura dos documentos HTML, algumas páginas se estruturam em formas de tabela e a extração destas páginas é comprometida.
A interface não tem sido o foco da pesquisa até o presente momento, a estrutura da ferramenta é funcional, porém é possível um maior aperfeiçoamento.

4 CONCLUSÃO

O acesso à informação é importante para a população e quando estas se encontram em formato não aberto a utilização destas informações se torna muito complicada, neste caso as ferramentas são úteis para auxiliar aqueles que possuem interesse em obter as informações necessárias.
A utilização das ferramentas de desenvolvimento corretas, neste caso o Java, pode auxiliar a alcançar resultados satisfatórios no desenvolvimento de novas soluções para os mais diversos problemas. As bibliotecas externas também auxiliam no desenvolvimento quando oferecem opções para que o desenvolvimento seja facilitado acrescentando funcionalidades que antes deveriam ser tratadas pelo desenvolvedor.

REFERÊNCIAS

CORRÊA, Andreiwid Sheffer; CORRÊA, Pedro Luiz Pizzigatti; DA SILVA, Flávio Soares Corrêa. Transparency portals versus open government data: an assessment of openness in Brazilian municipalities. In: Proceedings of the 15th Annual International Conference on Digital Government Research. ACM, 2014. p. 178-185.

PARK, Sung Hee et al. HTML5 ETDs. In:Proceedigs of International Symposium on Electronic Thesis and Dissertations. Austin, TX, USA. 2010.

W3C. As três leis e os oito princípios de dados abertos, São Paulo, 2010. Disponível em<http://www.w3c.br/pub/Materiais/PublicacoesW3C/dados-abertos-governamentais.pdf>. Acesso em 20 jul 2015.

W3C. Manual dos dados abertos: desenvolvedores. São Paulo, 2011. Disponível em < http://www.w3c.br/pub/Materiais/PublicacoesW3C/manual_dados_abertos_desenvolvedores_web.pdf >. Acesso em 20 jul 2015.