# MÉTODOS DE EXTRAÇÃO DE DADOS TABULARES A PARTIR DE HTML E PDF

Felipe Gustavo Lopes[1], Evandro Henrique Couto de Paula[2], Fábio Reis Santos[3] e Andreiwid Sheffer Corrêa[4]

## Introdução

As prefeituras brasileiras ainda estão se adaptando a publicação de dados no formato aberto, que foi introduzido pela LAI (Lei de Acesso à Informação), Lei nº 12.527, de 18 de novembro de 2011 [1]. Por esta ser uma lei recente, os tipos de arquivos mais utilizados e preferidos pelas administrações municipais na divulgação dos dados públicos ainda são PDF e HTML.

O processo de adaptação é longo, por isso este trabalho propõe métodos para extrair conteúdo tabular, originalmente publicado em formato PDF e HTML, e disponibilizá-lo como dados no formato universal CSV (*comma-separated values*) para possibilitar o processamento automatizado por máquina e o acesso a partir de qualquer ferramenta. Os métodos foram implementados em C# e testados a partir de documentos reais publicados em portais de transparência municipais.

## Fundamentação Teórica

O CKAN (*Comprehensive Knowledge Archive Network*) [2], uma das principais plataformas de dados abertos da atualidade, oferece ferramentas que facilitam a publicação de dados, assim como sua manipulação, disponibilizando-os facilmente em um único portal, desde que o arquivo tratado esteja em formato aberto. Para oferecer esse tipo de acesso à sociedade, é extraído o conteúdo tabular de arquivos PDF e de páginas HTML existentes nos portais de transparência municipais e então a tabela extraída é salva em um arquivo CSV, um tipo de arquivo universal, definido como “colunas separadas por ponto e vírgula”.

Páginas HTML possuem uma definição explícita de tabela, pois é através da *tag* <table> que uma tabela é gerada e organizada. Além disso, seletores como “ID” e “class”, quando são únicos em um arquivo, permitem que seja buscado um único elemento em especial dentro de toda a página [3]. Assim, identifica-se a tabela desejada para então extrair o texto.

O tipo de arquivo PDF não possui uma definição interna de tabela. Ao gerar um arquivo deste tipo, internamente são definidos parâmetros sobre a posição de cada cadeia de texto (*string*) na página, com um valor numérico para a vertical e outro para a horizontal. Buscando trechos no arquivo com um mesmo valor para a sua posição vertical, por exemplo, torna-se possível extrair uma única coluna da página. Estes parâmetros estão ocultos na visualização padrão, mas compõem a organização estrutural deste arquivo. Onde há algum texto, este trecho e a sua posição na página vêm entre os parâmetros BT para início e ET para fim [4].

## Materiais e Métodos

Os métodos propostos por este trabalho foram implementados utilizando a plataforma de desenvolvimento Microsoft Visual Studio 2012, com o *framework* ASP.NET e código escrito em C#. Para extração de páginas HTML, foi utilizada como apoio a biblioteca Html Agility Pack [5]. Para os documentos PDF, foi utilizada a biblioteca iTextSharp [6].

Na extração de conteúdo a partir de páginas HTML, conforme indicado nos passos 1A e 2A da Fig. 1, busca-se dentro do código-fonte da página (1A) o seletor “ID”, o seletor “class” ou a própria *tag* <table>, para a tabela, e posteriormente <tr> e <td> ou <th>, para linhas e colunas, respectivamente. Assim, torna-se possível a extração de todo o texto que está dentro da tabela (2A), pois é permitida a identificação de tabelas específicas, que é onde estão os dados de interesse, pois muitas vezes há mais de uma tabela em uma única página.

Para PDF, passos 1B e 2B da Fig. 1, o processo é realizado com todo o conteúdo do arquivo, incluindo os dados sobre o posicionamento de cada cadeia de texto, que são seus delimitadores (1B). Busca-se, então, somente o texto desejado, que são os trechos entre os delimitadores “BT” e “ET“ (2B). Desta forma, é

[1] Estudante do curso de Tecnologia em Análise e Desenvolvimento de Sistemas, IFSP, Campinas/SP. E-mail: felipe.lopes@outlook.com
[2] Estudante do curso de Tecnologia em Análise e Desenvolvimento de Sistemas, IFSP, Campinas/SP. E-mail: evandro.coutodepaula@gmail.com
[3] Estudante do curso de Tecnologia em Análise e Desenvolvimento de Sistemas, IFSP, Campinas/SP. E-mail: fabitous@gmail.com
[4] Professor do curso de Tecnologia em Análise e Desenvolvimento de Sistemas, IFSP, Campinas/SP. E-mail: andreiwid@ifsp.edu.br

possível compor as células da tabela em uma matriz, organizando-as de acordo com seus índices de linha e coluna.

Para os dois tipos de arquivo, após ser realizada a extração e organização dos dados, são adicionados símbolos ";" (ponto e vírgula) como delimitadores de coluna, observando-se os espaços vazios e a mescla de células. Por fim, todo o texto é salvo em um novo arquivo do tipo CSV (passo 3 da Fig. 1).

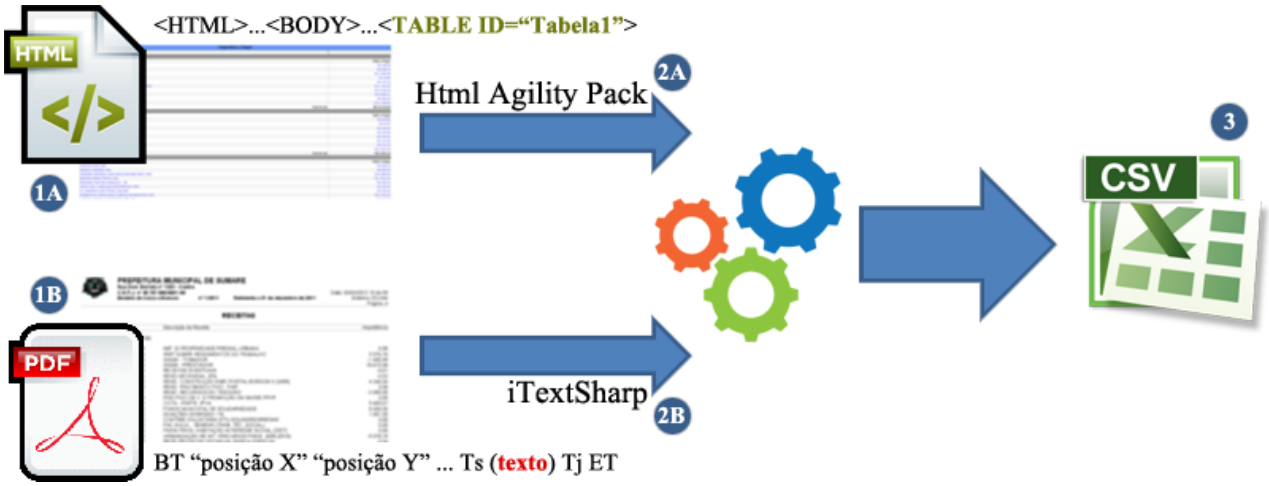


**Figura 1. Métodos de extração de informações tabulares a partir de páginas HTML e documentos PDF.**

## Apresentação e Discussão dos dados

Foram utilizados casos reais da Prefeitura de Americana [7] para a extração de arquivos PDF. As informações dispostas nas tabelas em alguns dos documentos utilizados estavam com uma visualização ruim, com colunas coladas em outras. Isso foi importante para determinar se o método de extração de PDF baseado em seus delimitadores iria, de fato, funcionar corretamente. Tabelas HTML foram utilizadas também de casos reais, mas da Prefeitura de Pedreira [8]. A organização do código-fonte das tabelas analisadas estava fora de padrão, com muitas *tags* de tabela dentro de outras, se opondo a forma de desenvolvimento *tableless* [9]. Para extrair corretamente todo o conteúdo, foi necessário identificar de maneira única a tabela em questão através de algum elemento dentro da *tag* <table>.

Os resultados obtidos corresponderam com precisão aos documentos originais. Os dados tabulares extraídos, tanto de páginas HTML, quanto de arquivos PDF, foram estruturados em arquivos do tipo CSV de maneira fiel a como foram dispostos originalmente.

A partir dos métodos aqui implementados, pretende-se como trabalho futuro desenvolver um sistema para buscar informações tabulares em HTML e PDF disponibilizadas nos portais de transparência municipais e assim realizar a extração dos dados e salvá-los em arquivos CSV. Para isto, estão sendo estudadas ferramentas que auxiliem na automação de navegadores web, para que se possa acessar páginas e arquivos dinâmicos.

Com esse sistema, pretende-se levar informações publicadas pelas prefeituras para a sociedade no formato aberto, possibilitando aos cidadãos o acesso irrestrito, universal e independente de tecnologia a estes dados. Esta iniciativa impulsionará o Brasil para o cumprimento da LAI e servirá como passo importante na adaptação da administração pública rumo aos conceitos de dados abertos.

## Referências

[1] CGU - Acesso à Informação. Disponível em: <http://www.acessoainformacao.gov.br/acessoainformacaogov/acesso-informacao-brasil/>. Acesso em: 7 mai. 2014.
[2] ckan - The open source data portal software. Disponível em: <http://ckan.org/>. Acesso em: 7 mai. 2014.
[3] HTML elements - HTML5. Disponível em: <http://www.w3.org/TR/html-markup/elements.html>. Acesso em: 7 mai. 2014.
[4] LOWAGIE, Bruno. PDF Document Structure. ***The ABC of PDF with iText*: PDF Syntax essentials**. Leanpub, 2014. p. 27-74.
[5] Html Agility Pack. CodePlex. Disponível em: <http://htmlagilitypack.codeplex.com/>. Acesso: em 7 mai. 2014.
[6] iText, Programmable PDF software | iText Software. Disponível em: <http://itextpdf.com/>. Acesso em: 7 mai. 2014.
[7] Americana - Site Oficial. Disponível em: <http://www.americana.sp.gov.br/>. Acesso em: 13 mai. 2014.
[8] Prefeitura Municipal de Pedreira. Disponível em: <http://www.pedreira.sp.gov.br/>. Acesso em: 13 mai. 2014.
[9] Tableless layout HOWTO. Disponível em: <http://www.w3.org/2002/03/csslayout-howto>. Acesso em: 8 mai. 2014.