ESTUDO DE MÉTODOS DE ACESSO AUTOMATIZADO PARA CLASSIFICAÇÃO DE CONTEÚDO EM WEBSITES DE ACORDO COM DADOS ABERTOS GOVERNAMENTAIS

ARTHUR P. ROZADO[1], RAUL M. D. SOUZA[2], ANDREIWID S. CORRÊA[3]
[1]Graduando em Tecnologia em Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP , IFSP, Câmpus Campinas, arthurprozado@gmail.com.
[2]Graduando em Tecnologia em Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP , IFSP, Câmpus Campinas, raulmendes932@gmail.com.
[3]Docente, IFSP, Câmpus Campinas, andrewid@ifsp.edu.br.
Área de conhecimento (Tabela CNPq): 1.03.03.04-9 - Sistemas de Informação

Apresentado no 8° Congresso de Inovação, Ciência e Tecnologia do IFSP - 2017
06 a 09 de novembro de 2017 - Cubatão - SP, Brasil

RESUMO: De acordo com a legislação brasileira, os registros públicos governamentais devem ser disponibilizados de forma aberta à sociedade, porém a ineficiência na disponibilização destes dados em formato aberto e legíveis por máquina gera desperdício de recursos e compromete o acesso às informações, atualmente veiculadas por websites não-padronizados e muitas vezes incompatíveis com os princípios de dados abertos. Visando garantir a qualidade das formas de disponibilização de dados atualmente adotadas, este projeto objetiva estudar, comparar e aplicar métodos automatizados de acesso a esses websites para permitir posterior classificação de seus conteúdos. Os resultados serão utilizados para extrair métricas de avaliação que serão expostas à sociedade. Como resultado preliminar, tem-se uma nota resultante da análise de três critérios distintos: acessibilidade, complexidade e qualidade. A acessibilidade consiste na compatibilidade do website em prover conteúdo às pessoas com deficiência, sendo dividida em três níveis de prioridade. A complexidade se refere ao grau de dificuldade imposto aos usuários ao obter as informações e a qualidade é relacionada diretamente ao conteúdo semântico da página HTML, com a checagem de links e a disponibilidade do que é fornecido pelas páginas web.

PALAVRAS-CHAVE: dados abertos governamentais; web scraping; portal de transparência.

STUDY OF AUTOMATED ACCESS METHODS FOR WEBSITES CONTENT CLASSIFICATION ACCORDING WITH OPEN GOVERNAMENT DATA

ABSTRACT: According to Brazilian legislation, government public records must be made available in open form to society, but the inefficiency in the availability of this data in an open and machine-readable format generates waste of resources and compromises access to information, currently transmitted by non-standardized websites and often incompatible with open data principles. In order to guarantee the quality of the forms of data availability currently adopted, this project aims to study, compare and apply automated methods of access to these websites to allow further classification of their contents. The results will be used to extract evaluation metrics that will be exposed to society. As a preliminary result, we have a note resulting from the analysis of three distinct criteria: accessibility, complexity and quality. Accessibility consists of the compatibility of the website in providing content to people with disabilities, being divided into three levels of priority. The complexity refers to the degree of difficulty imposed to the users in obtaining the information and the quality is directly related to the semantic content of the HTML page, with the checking of links and the availability of what is provided by the web pages.

KEYWORDS: open government data; web scraping; transparency portal.

INTRODUÇÃO

Na atual na era da informação, espera-se que dados sejam compartilhados de forma simples, rápida e irrestrita, na maioria dos casos. Através do livre acesso aos dados e a composição de serviços governamentais digitais por meio deles, tem-se um caminho para a excelência dos serviços públicos digitais, essencial para o funcionamento das atividades com a participação da sociedade (CGU, 2011).

O conceito de dados abertos governamentais baseia-se na disponibilização destas informações em formato aberto, permitindo, irrestritamente, o uso, reuso e redistribuição, além de outros requisitos (OPEN KNOWLEDGE FOUNDATION, 2012; TAUBERER, 2014). No Brasil, a legislação para regulamentar a disponibilização de dados é amparada pela Lei de Acesso à Informação (BRASIL, 2011). Mesmo havendo diretrizes legais desde 2011, órgãos e entidades governamentais brasileiros ainda têm mostrado dificuldades em atender a legislação, disponibilizando dados incompatíveis com dados abertos, principalmente devido à heterogeneidade de seus websites, na maioria das vezes sem padronização técnica (CORRÊA et al., 2017; CORRÊA; CORRÊA; SILVA, 2014).

Devido à grande quantidade e volume de dados desses websites, tem-se a necessidade de aplicar métodos automatizados para avaliá-los, obtendo informações sobre desvios dos requisitos de dados abertos governamentais frente à legislação nacional. Isso permitirá evidenciar os problemas com maior eficiência e possibilitar à sociedade cobrança de providências.

MATERIAL E MÉTODOS

O projeto conta com um servidor Apache e um sistema desenvolvido em Java, utilizando métodos de *web scraping* (MITCHELL, 2015) para realizar análises e comparações do conteúdo e funcionamento de websites governamentais.

Para o desenvolvimento do sistema de testes, foi utilizado o IDE (*Integrated Development Environment*) Eclipse Java EE, um ambiente de desenvolvimento gratuito para programação em Java para web. Utilizou-se também o software ATOM, um editor de textos de código aberto com ampla compatibilidade com linguagens de programação.

Para análise dos websites, verificaram-se padrões de acessibilidade com base no Modelo de Acessibilidade em Governo Eletrônico (GOVERNO ELETRÔNICO, 2016), adaptação brasileira do padrão W3C denominado *Web Content Acessibility Guidelines* - WCAG (CALDWELL et al., 2008).

A complexidade do website é testada com um método desenvolvido especialmente para o projeto. Inicialmente, gera-se uma árvore da estrutura do website a partir da página inicial. Dessa forma, mapeiam-se as possibilidades de páginas que o usuário poderia ser redirecionado. Após obter a estrutura do website, as páginas são analisadas uma a uma em busca por código HTML específico, como as *tags* do tipo <FORM> que compõem a especificação de formulários web. Caso algum formulário seja encontrado, é realizada uma classificação de acordo com seu conteúdo (campos do formulário), podendo existir três casos: cadastro, busca ou requisição para dados.

Assim, quando um campo senha é encontrado, o formulário é classificado como cadastro. Quando um único campo de texto é encontrado, é classificado como busca. Os casos que não atendem a nenhuma das duas primeiras é classificado como requisição de dados: campos que devem ser preenchidos para ter acesso às informações (página dinâmica para a *deep web*). A nota de complexidade é composta pela média entre a distância e a dificuldade. A distância é a quantidade mínima de páginas entre a página inicial e os dados, enquanto a dificuldade é a existência de qualquer um dos formulários.

A qualidade é medida por uma média extraída de dados de formatação da página HTML e uma checagem de conteúdo, como, por exemplo, se todos os links disponíveis na página funcionam e se nenhum conteúdo desta está indisponível ou incompleto.

Foram utilizadas as bibliotecas JSoup do Java para adquirir conteúdo HTML e uma API (*Application Programming Interface*) que fornece um serviço automático de checagem de páginas web sobre questões de acessibilidade que são englobadas pelos padrões de dados abertos.

Foi utilizado um banco de dados MariaDB para armazenamento de dados extraídos dessas análises, assim como valores relevantes para o projeto. Estes são constantemente atualizados, garantindo que os dados coletados correspondam ao estágio atual do website.

RESULTADOS E DISCUSSÃO

A Tabela 1 a seguir apresenta um resumo dos dados gerados a partir das análises deste projeto. Os campos P1, P2 e P3 mostram o quantitativo de problemas de acessibilidade encontrados por prioridade. Os campos Complexidade e Qualidade são obtidos pela análise das páginas web e a Nota é a métrica utilizada neste projeto para quantificar a adequação do website a partir dos critérios definidos no método desta pesquisa.

Tabela 1. Amostra de resultados obtidos pelo acesso automatizado e classificação dos websites.

Website	P1	P2	P3	Complexidade	Qualidade	Nota
http://campinas.sp.gov.br/servico-ao-cidadao/portal-da-transparencia/	2	5	5	7	8	5,4
http://transparencia.prefeitura.sp.gov.br/Paginas/home.aspx	4	8	8	8	8	7,2

Fonte: elaborado pelo autor.

CONCLUSÕES

A classificação por método automatizado dos websites governamentais mostra-se possível. Com os resultados preliminares obtidos com esta pesquisa, já é possível identificar desvios entre o que é encontrado nos websites governamentais e o que é esperado para a disponibilizados de dados a partir das diretrizes legais existentes.

Com a checagem das métricas definidas neste projeto sendo executada constantemente e repetidas vezes, será possível evidenciar os problemas com maior eficiência. Assim, cobrar melhorias necessárias das equipes técnicas responsáveis pelos websites governamentais.

AGRADECIMENTOS

Agradecemos ao Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (PIBIFSP) – Edital 023/2016 – pelas bolsas de iniciação científica e suporte a este projeto de pesquisa.

REFERÊNCIAS

BRASIL. Lei nº 12.527 de 18 de novembro de 1996. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. Brasília. 18 nov. 2011.

CALDWELL, B. et al. Web Content Accessibility Guidelines (WCAG) 2.0. Disponível em: <http://www.w3.org/TR/WCAG20/>. Acesso em: 20 abr. 2017.

CGU. Acesso à Informação Pública: Uma introdução à Lei nº 12.527, de 18 de novembro de 2011. Brasília: CGU, 2011.

CORRÊA, A. S. et al. Transparency and open government data: a wide national assessment of data openness in Brazilian local governments. Transforming Government: People, Process and Policy, v. 11, n. 1, 8 mar. 2017.

CORRÊA, A. S.; CORRÊA, P. L. P.; SILVA, F. S. C. DA. 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, 2014.

GOVERNO ELETRÔNICO. eMAG - Modelo de Acessibilidade em Governo Eletrônico. Disponível em: <http://emag.governoeletronico.gov.br/>. Acesso em: 10 mai. 2017.

MITCHELL, R. Web Scraping with Python. Collecting Data from the Modern Web. O'Reilly, 2015.

OPEN KNOWLEDGE FOUNDATION. Open Data Handbook Documentation, 14 nov. 2012. Disponível em: <http://opendatahandbook.org/>. Acesso em: 18 abr. 2017

TAUBERER, J. Open Government Data: The Book - Second Edition, 2014. Disponível em: <https://opengovdata.io/>. Acesso em: 18 abr. 2017.