Universidade de São Paulo

Escola de Artes, Ciências e Humanidades

Graduação em Sistemas de Informação

Participantes

Clayton dos Santos Lima NºUSP: 8517140

Gabriel Bonassi Zulpo NºUSP: 8921951

Gabriel Henrique de Oliveira NºUSP: 8517202

Gleidson Fernandes de Souza NºUSP: 8657647

Giovane Gabrielli NºUSP: 8516420

Equipe 13 - Projeto 16

São Paulo

Junho de 2017

1 Introdução

A administração estatal no período anterior a 2012 poderia ser considerado uma “caixa preta”, sem informações claras e acessíveis. Com a Lei de Acesso à Informação entrando em vigor na data de 16 maio de 2012 a situação mudou. Através de tecnologias de informação como o E-Sic, plataforma que recebe solicitações para o cumprimento da LAI, dados em formato aberto começaram a fluir. Isso clareou o cenário da administração estatal, melhorou a transparência.

O Governo Aberto tem como objetivo tornar o processo da administração pública mais transparente, eficaz e participativo. Nesses princípios pensamos numa forma de colaborar.

Coletando os dados da plataforma Lattes com o Diário Livre e a relação partidária conseguimos montar um banco de dados no formato aberto que facilita posterior análises.

2 Objetivos

2.1 Objetivo Geral

Com este projeto, o grupo teve que cruzar as informações do Diário Livre com a plataforma Lattes e publicá-las de forma a serem reutilizadas no futuro.

2.2 Objetivos Específicos

- Estudar a base de dados da plataforma Lattes e do Diário Livre;
- Adaptar ferramentas de extração já existentes para a solução do problema;
- Extrair informações sobre os cargos de confiança de alto escalão na cidade de São Paulo, cargos com código DAS-16;
- Publicar informações resultantes seguindo o sistema de estrelas da W3C;

3 Metodologia

Para o projeto se desenvolver adequadamente, ele foi dividido em três etapas: Extração dos dados, Cruzamento de informações e Republicação.

Com isso em mente, o grupo trabalhou com três bases de dados: o Diário Livre, a plataforma Lattes e a base dados referentes à filiação partidária dos indicados à cargos comissionados.

Após uma breve explicação sobre cada base, será feita uma discussão mais detalhada a respeito das três etapas realizadas.

O Diário livre nasceu de um trabalho de conclusão de curso de um aluno da EACH-USP, cujo objetivo era republicar as informações do diário oficial de São Paulo de maneira aberta. Desde então, o Diário Livre tem fornecido as informações oficiais da cidade de São Paulo em formatos mais fáceis de serem lidos e processados por máquina, o que mostrou ser de extrema ajuda na extração e cruzamento de informações sobre os cargos de confiança de alto escalão no estado com outras bases de dados.

Uma das bases utilizadas é referente ao currículo acadêmico dos pesquisadores cadastrados na plataforma Lattes. Desenvolvida e mantida pelo Centro Nacional de Pesquisa e Qualidade (CNPQ), a ferramenta conta com o currículo de diversos pesquisadores e abrange os aspectos da vida acadêmica, como graduação, pós-graduação, especializações e áreas de atuação.

Sobre a Filiação partidária, uma base com as informações sobre pessoas que já se filiaram a algum partido brasileiro foi encontrada no endereço eletrônico http://www.tse.jus.br/partidos/filiacao-partidaria/relacao-de-filiados . Essa base de dados também nos foi fornecida em formato aberto, no caso específico, arquivos com extensão (.csv).

3.1 Extração dos Dados

Apresentaremos aqui como foram extraídos os dados utilizados para gerar a base de dados disponibilizada neste trabalho.

Foram eles:

- Nomes dos funcionários nomeados ao cargo comissionado de maior escalão da prefeitura da cidade de São Paulo (DAS-16).
- Informações sobre pessoas filiadas aos maiores partidos políticos no estado de São Paulo.
- Extração de informação de currículos acadêmicos da plataforma Lattes.

3.1.1 Extração Diário Livre

O Diário Livre (Martano, 2015), foi criado com as informações do Diário Oficial Municipal(DOM) da cidade de São Paulo, com o objetivo de republicar estas informações em formato aberto, para que fosse possível sua leitura e utilização por ferramentas computacionais.

Ele possui uma ferramenta de busca que ajuda o usuário, podendo ser um computador, a encontrar as informações relevantes e baixá-las em formato texto, JSON e XML, além de disponibilizar as informações completas dos anos de 2003 a 2013 em formato texto.

Para este trabalho, extraímos as informações em formato texto com a utilização de um algoritmo. Tal algoritmo faz uso de expressões regulares para correta extração dos campos contendo os nomes dos 85 comissionados. Após extrair estes nomes, consolidamos em um arquivo tabulado em formato aberto (.csv) para efetuarmos os cruzamentos futuros.

3.1.2 Extração Filiação partidária

O site do Tribunal Superior Estadual disponibiliza as informações de filiações partidárias, separadas por partidos e Estado, em formato aberto (.csv). A interface é apresentada na imagem abaixo:

Fonte:< http://www.tse.jus.br/partidos/filiacao-partidaria/relacao-de-filiados>

Baixamos manualmente as informações dos partidos do estado de São Paulo, importamos para uma ferramenta de banco de dados e consolidamos em uma única base de dados utilizando SQL. Esta base de dados foi utilizada para fazer os cruzamentos com os nomes das pessoas nomeadas para cargos comissionados DAS-16.

3.1.3 Extração Plataforma Lattes

Inicialmente o grupo utilizou um algoritmo desenvolvido em linguagem python desenvolvido por alunos de graduação da EACH-USP e disponível no github (https://github.com/foo0x29a/sadbois), e que será chamado de A1, que acessava a plataforma Lattes e fazia o download automaticamente dos currículos segundo uma área de atuação. Com esse algoritmo, iniciou-se um trabalho de adaptação para que ao invés de buscar por área de atuação, a busca fosse feita por nome, já que para o cruzamento das informações seria necessário o nome dos indivíduos.

A plataforma estudada até então, possuía um sistema de captcha que impedia o acesso computadorizado aos dados dos currículos, ao mesmo tempo que o grupo enfrentava dificuldades na adaptação do algoritmo A1. Nesse momento encontramos um novo algoritmo de extração, também disponível no github (https://github.com/lucachaves/lattes-crawler) e que não utiliza nenhuma ferramenta

para quebrar o captcha. A esse algoritmo chamaremos A2. Ao estudar o A2 o grupo descobriu que ele acessava uma nova plataforma Lattes, que não possuía captcha.

Apesar desse último avanço, os problemas com relação ao Lattes não acabaram. Analisando melhor a plataforma, observou-se que os currículos eram indexados na página por meio de dois números de identificação distintos, um de dez dígitos (id10) e outro de dezesseis (id16), de maneira que, para ter acesso à versão do currículo disponível no site seria necessário ter o id16 e para fazer o download dessa mesma versão em xml seria necessário ter o id10.

Como o grupo precisava dos dois identificadores, já que o A2 se utilizava tanto do nome da pessoa quanto dos dois ids para conseguir baixar a informação em xml, baseado na lei de acesso à informação foi feito um pedido de informação para a CNPQ, solicitando um documento com três colunas, o nome do pesquisador(a), e os índices de 10 e 16 dígitos. No endereço eletrônico da CNPQ é disponibilizado um arquivo com o id16 de todos os currículos do sistema, porém com apenas essa informação, sem identificador (nome) e os id's de 16 e 10 dígitos, não foi possível utilizá-lo nesse trabalho.

Com uma resposta muito insatisfatória da parte do CNPQ, que relutava à cooperar, e com os dados de nome completo já extraídos do Diário Livre, o grupo decidiu verificar manualmente cada nome dos comissionados DAS-16 na plataforma nova do Lattes e caso encontrado extrair seus dois ids, já que se tratavam de 85 nomes no total.

O procedimento adotado pelo grupo para a extração manual dos índices no Lattes se deu da seguinte maneira: um nome era procurado na ferramenta de busca da plataforma, se encontrado um e apenas um registro com o nome exatamente igual ao procurado, eram obtidos os dois índices e associados ao nome da pessoa.

Portanto, casos com mais de um registro com o mesmo nome procurado (homônimos), registros com um nome similar ou faltando algum nome foram ignorados para esse trabalho.

3.2 Cruzamento de informações

3.2.1 Diário Livre - Filiação partidária

Com base nestas informações extraídas, pode-se cruzar, utilizando o nome do funcionário comissionado como chave de cruzamento, as informações do diário livre com as de filiação partidária.

Subimos ambos os conjuntos em um banco de dados e efetuamos o cruzamento. Foram encontrados 15 pessoas das 85 com informação nas duas bases, removendo homônimos.

Foram encontrados 5 pessoas com homônimos, podendo conter alguns significados. Dentre eles temos duas hipóteses: a primeira, que o nome da pessoa é comum o bastante para ter mais de um filiado com mesmo nome na história dos partidos. A segunda, mais interessante para os propósitos do trabalho, que a pessoa entrou e saiu de um ou mais partidos por diversas vezes.

Verificando está segunda hipótese, podemos perceber que existem pessoas, além do nosso conjunto de dados, que possuem um número grande de entradas e saídas de partidos diversos durante o tempo. O que seria normal se não houvessem muitos casos de entradas e saída no mesmo dia e com poucos dias de diferença.

3.2.2 Diário Livre - Plataforma Lattes

Uma vez obtido os nomes dos indicados aos cargos comissionados e seus respectivos índices extraídos da plataforma Lattes, o grupo utilizou o algoritmo A2 modificado (https://github.com/XhaMbuwandong/lattes-sucks_crawler) para fazer o download automaticamente dos currículos em formato xml.

Juntamente com os arquivos xml, juntou-se um arquivo em (.csv) com três colunas: uma com os nomes encontrados no diário livre e no Lattes seguindo o quesito de extração detalhado anteriormente e as outras duas colunas com os índices de dezesseis e dez dígitos respectivamente.

O arquivo “NomesDAS16_temLattes_extraido” citado acima foi disponibilizado no Github da disciplina:

https://github.com/EACH-Lab2017/ACH3778_Governo_Aberto/tree/master/Equipes/Equipe%2013 .

3.3 Republicação

Todos os materiais utilizados pelo grupo se encontram na plataforma Github no link citado acima[1]. Nos próximos parágrafos será feita uma breve descrição de cada arquivo a fim de facilitar a leitura.

Sobre os algoritmos utilizados, A1 e A2 se encontram dentro da pasta “Projeto Extrator Lattes” sendo que o arquivo chamado “sadbois-master” representa o A1 enquanto o “crawler_lattes.py” representa o A2 está na pasta “lattes-bothers_crawler”.

Para os arquivos utilizados nas extrações, dentro da pasta “Projeto Extrator Lattes/lattes-bothers_crawler” pode-se encontrar o arquivo (.csv) com os nomes dos nomeados aos cargos de confiança de alto escalão (DAS-16) juntamente com seus respectivos índices de 10 e 16 dígitos (NomesDAS16_temLattes_extraido.csv). Também é possível visualizar na mesma pasta uma pasta própria chamada “data/cv_xml” contendo os currículos baixados em xml.

4 Resultados

Partindo do Diário Livre, foram extraídos 85 nomes (com duplicações) das pessoas nomeadas aos cargos comissionados de alto escalão (DAS-16), referentes aos anos de 2003 a 2013. Desses nomes, após o cruzamento dos dados com a plataforma Lattes, restaram apenas 12 (sem duplicações), enquanto foram encontrados 23 nomes sem homônimos ligados à algum partido por meio dos dados da base de filiação, e 440 nomes, levando em consideração os homônimos.

Uma observação importante realizada pelo grupo foi que, com relação ao currículo Lattes, a maior parte das pessoas encontradas pertenciam a área de saúde,

tendo graduações e especializações na área médica e trabalhando em hospitais pela capital.

A partir do cruzamento das diversas informações obtidas neste trabalho, o grupo conseguiu associar as informações acadêmicas às pessoas que possuem altos cargos de confiança na cidade de São Paulo e se elas já se associaram alguma vez a algum partido político. Não cabe ao grupo julgar se ocorreu responsabilidade no momento da nomeação do cargo por meio da informação partidária e acadêmica associada a cada indivíduo, porém os dados estão abertos e publicados de maneira livre para que qualquer pessoa que possua algum interesse em estudá-los de maneira mais aprofundada consiga efetivar seu objetivo, e essa era a proposta que permeou e incentivou o grupo a desenvolver o projeto.

5 Discussão

Com o desenvolvimento do projeto vivenciamos toda a experiência de solicitação, consulta e análise de dados abertos. Observamos as dificuldades e barreiras existentes para a aquisição de dados simples, como por exemplo, os dados dos pesquisadores cadastrados na plataforma Lattes, e a necessidade de conhecimento técnico para realizar tal análise, visto que os dados quando disponíveis ainda necessitam de certo tratamento devido ao grande volume. Com isso torna-se inviável a elaboração de informação a partir desses dados por algum cidadão que não possua tais habilidades técnicas. Portanto, iniciativas como a nossa de republicação de dados já tratados, organizados e unificados torna o processo de gerar informação muito mais acessível a todo cidadão que deseja contribuir com o desenvolvimento de um governo mais aberto.

6 Referências Bibliográficas

MARTANO, A. M. R. Diário Livre: co-criação de uma ferramenta para publicação de um diário oficial em formato aberto. Dissertação (Mestrado) | Universidade de São Paulo, 2015.

7 Bibliografia/Webgrafia

Disponível em: <http://www.tse.jus.br/partidos/filiacao-partidaria/relacao-de-filiados>. Acessado em 05 junho de 2017.
Disponível em: <http://ferramentas.artigo19.org/historico>. Acessado em 19 junho de 2017.
Disponível em: <http://www.acessoainformacao.gov.br/central-de-conteudo/publicacoes/arquivos/cartilhaacessoainformacao.pdf>. Acessado em 19 junho de 2017.