



**INSTITUTO FEDERAL DE EDUCAÇÃO,
CIÊNCIA E TECNOLOGIA DE SÃO PAULO**
Câmpus Campinas

Curso de Tecnologia em Análise e Desenvolvimento de Sistemas

Método de catalogação de URLs com base no Common Crawl

Autor: Alison Lúcio dos Santos
Orientador: Prof. Dr. Andreiuid Sheffer Corrêa

**CAMPINAS
2018**

Autor: Alison Lúcio dos Santos
Orientador: Prof. Dr. Andreiwid Sheffer Corrêa

Método de catalogação de portais de dados abertos a partir de URLs indexadas no Common Crawl

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.

CAMPINAS
2018

Autor: Alison Lúcio dos Santos
Orientador: Prof. Dr. Andreiuid Sheffer Corrêa

**Método de catalogação de portais de dados abertos a partir de URLs indexadas no
Common Crawl**

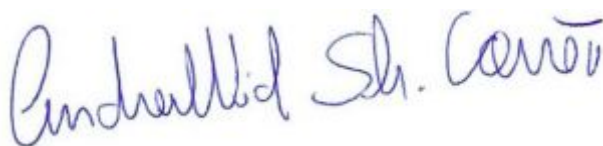
Trabalho de Conclusão de Curso
apresentado como exigência parcial
para obtenção do diploma do curso de
Tecnologia em Análise e
Desenvolvimento de Sistemas do
Instituto Federal de Educação, Ciência e
Tecnologia Câmpus Campinas.

Aprovado em 07 de Novembro de 2018.

BANCA EXAMINADORA

Alcino Vilela Ramos Junior

Prof. Me. Everton Josue da Silva
IFSP Câmpus Campinas



Prof. Dr. Andreiuid Sheffer Corrêa (orientador)

IFSP Câmpus Campinas

*Dedico este trabalho a Deus,
aos meus pais, colegas de classe,
colegas de trabalho, a todos os meus
professores desde de minha infância e a todos
os servidores do Instituto que colaboraram
em minha jornada acadêmica.*

Ficha catalográfica
Instituto Federal de São Paulo – Câmpus Campinas
Biblioteca
Rosangela Gomes – CRB 8/8461

Santos, Alison Lúcio dos

S237m Método de catalogação de URLs com base no Common Crawl / Alison Lúcio dos Santos. – Campinas, SP: [s.n.], 2018.

49f. : il.

Orientador: Andreiwid Sheffer Corrêa Trabalho de Conclusão de Curso (graduação) – Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas, 2018.

1. Portal de dados abertos governamentais. 2. Common Crawl. 3. Data science. 4. Web scraping. 5. Lei de acesso à informação. I. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas. II. Título.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e da capacidade de aprender.

Agradeço a todos os professores que tive durante toda minha vida, pois, desde dos primeiros, eles me prepararam para os desafios de hoje. Aos servidores do IFSP Câmpus Campinas e Câmpus Campos do Jordão onde comecei minha jornada e que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço aos meus pais que desde de sempre foram meus heróis, me apoiando, me incentivando e me dando bons exemplos.

Agradeço aos membros de minha banca por disponibilizarem seu tempo para este trabalho.

Agradeço ao meu orientador pela paciência, pela confiança e pelas muitas sugestões de melhorias e ajustes no trabalho.

Resumo

Este trabalho objetivou desenvolver um método de catalogação de URLs de portais de dados abertos governamentais a partir da base de dados do Common Crawl, que é um projeto livre e aberto para indexação de dados de toda a internet. Foram empregadas técnicas de data science e web scraping, considerando os conceitos e princípios de dados abertos governamentais, Lei de Acesso à Informação brasileira, plataformas de dados abertos e um trabalho seminal onde se desenvolveu uma forma de identificar algumas destas plataformas. Como resultado obtido, foi possível observar a distribuição geográfica dos portais de dados identificados e a especificação da plataforma utilizada. A contribuição deste trabalho se deu com um método que poderá ser reproduzido em grande escala e executado repetidamente, de modo a servir de base para construção de um repositório central e atualizável com endereços virtuais (URLs) dos portais de dados abertos em utilização.

Palavras-chave: Portal de dados abertos governamentais. Common Crawl. Data science. Web scraping. Lei de acesso à informação.

Abstract

This work aimed to develop a method of cataloging open data URLs from the Common Crawl database, which is a free and open project for indexing data from all over the internet. Data science and web scraping techniques were employed, considering open data standards and systems, Brazilian Access to Information Law, open data platforms and seminal work. As a result it was possible to observe a geographical distribution of the data data and a database used. The basis of this work is a method that can be reproduced on a large scale and is executed repeatedly to serve as the basis for building a central and updatable repository with the virtual machines (URLs) of open access data in use.

Keywords: Portal of open government data. Common Crawl. Data science. Web scraping. Law on access to information.

LISTA DE FIGURAS

Figura 1 - Trabalhos envolvidos	16
Figura 2 - Arquitetura de arquivos do Common Crawl	31
Figura 3 - Tabela no banco de dados	33
Figura 4 - Fluxograma das etapas para se obter as URLs governamentais	34
Figura 5 - Gráfico de URLs .gov com palavras-chave por país	41
Figura 6 - Imagem geográfica das URLs encontradas com palavra-chave	44
Figura 7 - Gráfico de portais de dados abertos e suas plataformas por país	44
Figura 8 - Gráfico da quantidade de portais de dados abertos por plataforma	46
Figura 9 - Imagem geográfica das URLs de portais de dados abertos encontrados	47

LISTA DE TABELAS

Tabela 1 - Plataformas e suas respectivas assinaturas	28
Tabela 2 - Tabela criada com os dados do banco de dados	35
Tabela 3 - Tabela criada com dados retirados do sítio	35
Tabela 4 - Tabela com dados mesclados	36
Tabela 5 - Tabela de sinônimos ou palavras referentes a sítios governamentais	37
Tabela 6 - Tabela de URLs .gov com palavras-chave selecionadas	41
Tabela 7 - Quantidade de portais de dados abertos encontradas por país	45
Tabela 8 - Quantidade de portais de dados abertos encontrados por plataformas	46
Tabela 9 - Lista de URL de portais de dados abertos encontrados	47

LISTA DE SIGLAS

API	Aplication Programming Interface - Interface de Programação de Aplicações
CEAP	Cota para Exercício da Atividade Parlamentar
CSV	Comma Separated Values - Valores Separados por Vírgula
GB	Gigabyte
HTML	Hypertext Markup Language - Linguagem de Marcação de Hipertexto
IDs	Identificadores
JSON	JavaScript Object Notation - Notação de Objeto JavaScript
LAI	Lei de Acesso à Informação
MB	Megabyte
OGD	Open Government Data - Dados Abertos Governamentais
PDF	Portable Document Format - Formato Portátil de Documento
TXT	Text
URL	Uniform Resource Locator - Localizador Padrão de Recursos ou endereço
virtual	
XML	eXtensible Markup Language - Linguagem de Marcação Extensível

SUMÁRIO

1 Introdução	13
2 Justificativa	14
3 Objetivos	15
3.1 Objetivo Geral	15
3.2 Objetivos Específicos	15
4 Fundamentação Teórica	16
4.1 Trabalhos relacionados	16
4.1.1 De onde vieram as assinaturas	17
4.2 Web Scraping	17
4.3 Data Science	17
4.4 Direito de acesso à informação	18
4.5 Dados Abertos Governamentais (OGDs)	19
4.6 Common Crawl	21
4.6.1 Formatos de dados	22
4.6.1.1 Formato WARC	22
4.6.1.2 Formato de arquivo WAT	23
4.6.1.3 Formato de arquivo WET	24
4.7 Padrão do Common Crawl	26
4.8 Portais de dados abertos	26
4.9 Plataformas de dados abertos utilizadas neste trabalho	26
4.9.1 CKAN	27
4.9.2 Socrata	27
4.9.3 OpenDataSoft	27
4.9.4 ArcGIS OpenData	28
4.10 Assinaturas	28
5 Métodos	30
5.1 Arquitetura de arquivos do projeto Common Crawl	30
5.2 Acessando os dados do Common Crawl	31
5.2.1 Obtendo as URLs de indexações	31
5.2.2 Escolhendo o tipo de arquivo	32
5.2.3 Baixando o arquivo .WAT	32
5.3 Algoritmo para buscar URLs que tenham .gov	32
5.4 Banco de dados	33
5.5 Determinando de qual país a URL pertence	34
5.6 Sinônimos de .gov	36
5.7 Dificuldades encontradas	39

5.8 Disponibilização do código	39
6 Resultados	40
6.1 Quantidade de URLs .gov com as palavras-chave selecionadas	40
6.2 Quantidade de URLs de portais de dados abertos encontradas	44
6.3 Quantidade de URLs de portais de dados abertos encontrados por plataforma	45
6.4 As URLs de portais de dados abertos encontradas	47
7 Considerações finais	50
8 Trabalhos futuros	51
9 Referências	52

1 Introdução

A Lei de Acesso à Informação (LAI) está criando uma oportunidade única e muito valiosa de acesso aos dados públicos, o que é considerado um caminho para a disponibilização dos denominados Dados Abertos Governamentais (OGD). A LAI obriga as instituições públicas a disponibilizarem para a população seus gastos e quaisquer outros dados de interesse da sociedade, como por exemplo: quantidade de habitantes, escolas, impostos coletados e finanças públicas.

Os portais de dados abertos existentes, que são os mecanismos que viabilizam OGD, estão espalhados no mundo em milhares de sítios de instituições públicas o que dificulta a coleta e análises dos mesmos, por isso acredita-se que a criação de um repositório único de endereços virtuais (URLs) desses portais seria uma ferramenta de enorme importância para a análise e utilização pela população, criando assim uma base onde possa facilmente ter acesso aos portais existentes para fiscalizar seus representantes e avaliar a eficácia e eficiência na gestão dos recursos públicos. No entanto, um dos grandes obstáculos para tal projeto é justamente encontrar e catalogar as URLs corretas para adicionar ao repositório, uma vez que podem ser alteradas e muitas são criadas todos os dias.

O grande problema hoje é que existem diversas formas e plataformas para se criar um portal de dados abertos de OGD, sendo assim, existem diversos padrões a serem identificados, analisados e catalogados ou pior, muitos podem ser criados sem padrão algum (CORRÊA, 2017).

Este trabalho tem como objetivo desenvolver um método de catalogação de portais de dados abertos que utilizem uma das quatro plataformas consideradas e que estejam nos bancos de dados de Common Crawl, que é um projeto de indexação livre e aberto de toda a internet.

2 Justificativa

A motivação deste trabalho está pautada na afirmativa de que os maiores agentes contra a corrupção são os cidadãos que investem seu tempo e por vezes até seus recursos para fiscalizar o uso de recursos públicos (LAMBRANHO, 2016). Com as LAIs adotadas por diversos países eles ficaram empoderados pelo acesso aos dados, mas ainda há muito o que melhorar, como por exemplo a qualidade dos portais de dados abertos e o melhor uso dos OGDs na gestão pública para impacto social positivo.

O acesso de qualquer pessoa a OGD de seu interesse pode resultar no início de mudanças na gestão de políticas públicas de governo uma vez que toda pessoa poderá fiscalizar os gastos públicos geridos pelos seus representantes.

O problema é que as URLs nem sempre seguem um padrão de nomenclatura para dizer se pertencem a órgãos governamentais, como por exemplo a sigla .gov, muito menos para dizer se elas referenciam portais de dados abertos e cada instituição pode criar o seu próprio portal de dados, faz-se necessário filtrar as URLs referentes a órgãos governamentais e depois descobrir se elas levam ou possuem um link que leve a algum portal de dados abertos compatível com OGD.

Um repositório central de URLs de portais de dados abertos governamentais possibilitará a análise de dados e extração de métricas de gestão dos recursos públicos, desta forma viabilizando também a avaliação da qualidade dos sistemas empregados como portal e implementação de OGD.

3 Objetivos

3.1 Objetivo Geral

O objetivo deste trabalho foi desenvolver um método de catalogação de portais de dados abertos a partir de URLs indexadas no Common Crawl.

3.2 Objetivos Específicos

- I. Desenvolver um método de acesso aos dados do Common Crawl;
- II. Desenvolver um método que implemente e use as assinaturas de identificação das quatro plataformas de dados abertos consideradas;
- III. Modelar um banco de dados para a catalogação;
- IV. Definir um modelo para automatizar o processo para funcionar em larga escala e repetidamente usando como parâmetros as URLs que tenham o sufixo .gov ou sinônimos do mesmo utilizados em outros idiomas/culturas.

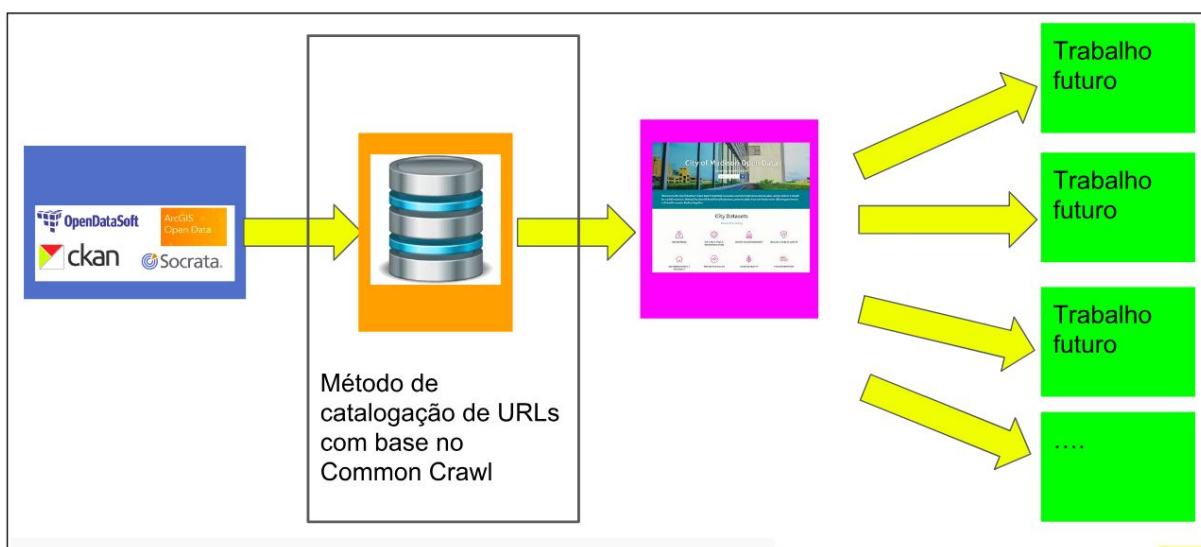
4 Fundamentação Teórica

Para a realização deste trabalho foi necessário o estudo de tecnologias, técnicas, leis e diretrizes relevantes a este trabalho. Estes assuntos serão tratados nos sub-capítulos posteriores.

4.1 Trabalhos relacionados

Este trabalho considerou as assinaturas desenvolvidas por Corrêa, Zander e da Silva (2018) para criar uma base de dados que deverá ser utilizada em trabalhos futuros para o desenvolvimento de um repositório central e atualizável de URLs de portais de dados abertos em um sítio onde ficarão expostas para consulta pública. A partir do repositório central, poderão ser desenvolvidas inúmeras aplicações utilizando os portais de dados catalogados de diversos países. A Figura 1 mostra no retângulo azul o trabalho de Corrêa, Zander e da Silva (2018); o retângulo laranja é onde se situa a proposta deste trabalho; a partir deste trabalho, no retângulo roxo, é visualizado o trabalho futuro que desenvolverá um repositório central de URLs de portais de dados abertos governamentais que servirão de base para inúmeros projetos que poderão ser idealizados, que são representados na Figura 1 pelos retângulos verdes.

Figura 1 . Trabalhos envolvidos



Fonte: Elaborado pelo autor (2018)

4.1.1 De onde vieram as assinaturas

O artigo *Investigating Open Data Portals Automatically: A Methodology and Some Illustrations* realizado por Côrrea, Zander e da Silva (2018), está representado na Figura 1 pelo retângulo azul, teve como objetivos desenvolver uma metodologia para investigar a adoção e uso de plataformas de disponibilização de dados abertos em todo o mundo e dizer qual seria a metodologia mais adequada para investigar a adoção de portais de dados abertos.

Foram duas as contribuições deste trabalho, foram elas: O desenvolvimento de métodos para pesquisar automaticamente portais de dados abertos, identificar plataformas específicas de disponibilização de dados abertos (que identificou e catalogou as assinaturas que foram utilizadas no trabalho corrente), coletar metadados sobre seus conjuntos de dados subjacentes. Após a coleta dos dados foram feitas análises para mostrar que os dados podem produzir insights no que diz respeito à compreensão, adoção e uso de cada portal de dados abertos de acordo com cada plataforma de disponibilização de dados utilizada.

4.2 Web Scraping

Web Scraping é um conjunto de técnicas utilizadas para extrair dados diretamente de páginas web, ou seja tudo o que estiver no formato HTML, arquivos XML, CSV, ou de qualquer outro formato no sítio podem ser extraídos (MCKINNEY, 2018).

Foram usadas técnicas de Web Scraping para obter dados de páginas do sítio do Common Crawl e de outros sítios necessários para a realização deste trabalho, para isso foram utilizadas as bibliotecas requests e BeautifulSoup do Python.

4.3 Data Science

Data Science é uma ciência que estuda e visa extrair conhecimento de dados estruturados e desestruturados já disponíveis, sobretudo na internet. É uma ciência multidisciplinar que pode lidar com todo o ciclo de vida do dado desde sua geração, captura, transformação e análise por fim gerando informações relevantes (GRUS, 2016).

Este trabalho utilizou técnicas de Data Science como coleta de dados de fontes diferentes, transformação dos dados em informações relevantes e análise dos mesmos utilizando a linguagem Python juntamente com as bibliotecas Plotly e Pandas.

4.4 Direito de acesso à informação

A transparência pública refere-se ao procedimento de deixar acessível à população informações referentes a serviços e gastos pagos com recursos públicos, como salários, benefícios, cotas, contratos, resultados técnicos de quaisquer tipos de fiscalização e quaisquer ações realizadas por órgãos públicos municipais, estaduais e federais (BRASIL, 2016?).

O direito de acesso à informação visa atender ao inciso XXXIII do Art. 5º da constituição, onde diz que:

Todos têm direito a receber dos órgãos públicos informações de seu interesse particular, ou de interesse coletivo ou geral, que serão prestados no prazo da lei, sob pena de responsabilidade ressalvadas aquelas cujo o sigilo seja imprescindível à segurança da sociedade e do estado (BRASIL, 1988).

Para complementar o artigo constitucional foi criada a lei nº 12.527, de novembro de 2011, que impõe diretrizes para a liberação de dados, dos quais alguns incisos são mencionados abaixo. Elas pertencem ao Art 3º (BRASIL, 2011):

O inciso I diz que a negação do direito aos dados deve existir somente em caso de exceção pois a regra geral deve ser a disponibilização dos mesmos. O inciso II diz que os dados devem ser liberados mesmo que ninguém os peça. Já o inciso III diz que os dados devem ser disponibilizados na internet.

O fato de existir LAIs não significa que o acesso a informações seja fácil ou plenamente incentivado. É difícil saber quanto dos nossos impostos são gastos na iluminação das ruas, pesquisas médicas, saneamento básico, segurança, saúde, educação, salários e benefícios para políticos e servidores públicos. Estas informações estão nos dados fornecidos pelas instituições públicas normalmente viabilizados como portais de dados abertos. Aqui no Brasil temos os portais de transparência que disponibilizam dados a população, mas nem sempre seus dados são considerados OGDs, no capítulo “Dados abertos governamentais” serão explicados os oito princípios básicos para que um dado seja considerado um OGD.

Existem pessoas que exercem o seu direito de fiscalizar e denunciam os políticos e servidores públicos usando OGDs, como exemplo temos o projeto Operação Serenata de Amor, que usa a tecnologia com OGDs para fiscalizar, cobrar e corrigir nossos representantes públicos.

O projeto Operação Serenata de Amor tem como objetivo acabar com a corrupção no Brasil e está obtendo resultados, ele consiste em apontar e disponibilizar em seu site oficial chamado de Jarbas¹ discrepâncias em gastos de deputados para com a verba que cada um recebe mensalmente destinada para uso exclusivo e em exercício da função chamada de Cota para o Exercício da Atividade Parlamentar - CEAP (DEVEGILI, 2017). Foram encontradas discrepâncias como o deputado que pediu o reembolso de 6.205,00 reais (seis mil duzentos e cinco reais) por uma refeição, o deputado que pediu reembolso de 6.000,00 reais (seis mil reais) de gasolina, o que daria quase 30 tanques de combustível, e também foi verificado que 219 deputados costumam usar o valor máximo permitido por mês (APOIA.SE, 2018). Existem órgãos públicos Brasileiros que respeitam a lei da transparência mas não respeitam os princípios de OGDs que serão explicados no próximo capítulo.

4.5 Dados Abertos Governamentais (OGDs)

Dados Abertos Governamentais são dados do governo que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa, inúmeras vezes (OPEN KNOWLEDGE INTERNACIONAL, 2012?).

Existem princípios para o que são dados abertos. Abaixo são citados os 8 principais que foram criados na Califórnia em 2007 por 30 pessoas e foram chamados de Oito Princípios de Dados Abertos Governamentais (TAUBERER, 2007):

I. Completos

Todos os dados públicos devem ser disponibilizados, não devem ter limitações de privacidade, segurança ou privilégio, devem ser também gratuitos e localizáveis. Por este motivo a disponibilização dos dados não pode ser parcial, o acesso aos dados deve ser totalmente liberado a todos sem nenhum tipo de discriminação e todos os dados devem ser fornecidos gratuitamente e devem ser rastreáveis para a confirmação de suas fontes.

II. Primário

Os dados devem ser coletados na origem, com o nível mais alto possível de granularidade, sendo assim, dados disponibilizados que já tenham passado por algum tipo de

¹ https://jarbas.serenata.ai/dashboard/chamber_of_deputies/reimbursement/

filtragem ou limpeza não respeitam esse princípio, por exemplo: os dados de uso de recursos devem apresentar se possível cada gasto realizado e com a maior quantidade de informações possíveis e não somente uma soma do montante dos gastos no fim do mês.

III. Atuais

Devem ser liberados ao público o mais cedo possível, já que a finalidade de um OGD é empoderar o cidadão para analisar o uso de recursos pelo governo. Esses dados não serão considerados abertos se forem disponibilizados tardiamente. Um exemplo de descumprimento é: uma licitação para a construção de um túnel, porém esses dados só são disponibilizados após o término da construção.

IV. Acessível

O maior número possível de pessoas deve ter acesso aos dados uma vez que os mesmos devem ser disponibilizados no maior número possível de formatos já reconhecidos pela indústria, como CSV, TXT, XML entre outros.

V. Compreensível por máquina

Os dados devem ser estruturados em formatos que a máquina possa entender sem que um desenvolvedor tenha que transformá-los, isso deixa o processo automatizado e muito mais rápido. Para serem considerados abertos os dados devem ser de fácil manipulação e processamento automático. Ou seja, quando os dados vêm em um arquivo TXT ou CSV eles podem ser manipulados facilmente e se esses dados respeitam os outros princípios eles são considerados como dados abertos, mas se esses mesmos dados vêm em um arquivo PDF ou um arquivo escaneado eles não são considerados dados abertos pois esses formatos não são compreensíveis por máquinas, a qual somente consegue exibi-los, mas não os compreende para manipulá-los impedindo o processamento automatizado.

VI. Não discriminatório

Para acessar esses dados não deve se obrigar os usuários a fazerem cadastro ou algo do tipo que possa possibilitar sua identificação. Para que os dados sejam considerados dados

abertos o acesso a eles não deve ter nenhum mecanismo de reconhecimento ou cadastro de usuários.

VII. Não proprietário

O controle dos dados disponibilizados não devem ser exclusivos de organizações e entidades, sua manipulação não deve depender da licença ou compra de uma ferramenta ou serviço.

VIII. Livre de licença

Os dados não podem pertencer a patentes, marcas registradas ou regulações de segredo industrial. Restrições razoáveis quanto à privacidade, segurança e outros privilégios são aceitas, desde que transparentes e bem justificadas.

É possível que um portal de dados abertos siga a LAI mas não siga os princípios de dados abertos uma vez que não seguir a LAI significa o descumprimento da lei que pode acarretar em punições e não seguir os princípios de dados abertos significa o descumprimento de diretrizes o que não acarreta em punição.

4.6 Common Crawl

Common Crawl é uma organização sem fins lucrativos que tem um projeto com o mesmo nome que rastreia a Web e fornece o conteúdo ao público gratuitamente e sem restrições. A organização começou a rastrear a web em 2008 e seu conteúdo consiste em bilhões de páginas da web rastreadas várias vezes por ano (MORGAN, 2016). O Google também faz o rastreamento de sítios da web, mas a principal diferença entre o Common Crawl e o Google é que o segundo não disponibiliza sua base de dados ao público e atua com interesses prioritariamente comerciais.

O Common Crawl é um repositório aberto de dados de rastreamento da web e que podem ser acessados e analisados por qualquer pessoa, nele contém petabytes de dados coletados nos últimos 7 anos. Esse rastreamento ocorre uma vez por mês, e normalmente é chamado de indexação mensal. Ele contém dados brutos da página da web, metadados extraídos e extrações de texto. Esses dados são guardados na Amazon como parte do programa Public Dataset Program, e são de acesso gratuito (COMMON CRAWL, 2015).

Na base de dados do Common existem centenas de milhares de URLs de diferentes lugares do mundo, organizações, empresas e de outros meios, este trabalho catalogou URLs de portais de dados abertos que utilizaram para sua criação uma das plataformas de dados abertos, que serão explicadas no capítulo “Plataformas de dados abertos utilizadas”. Estas plataformas não tem ligação alguma com o Common Crawl, como o Common Crawl armazena tudo que encontra na internet consequentemente ele armazenou dados de portais de dados abertos que foram criados com essas plataformas, nos possibilitando usar sua base para encontrar suas URLs.

4.6.1 Formatos de dados

Nos repositórios de Common Crawl podem ser encontrados os seguintes tipos de dados (COMMON CRAWL, 2015):

- Arquivos WARC: armazenam os dados brutos de rastreamento, abaixo tem um exemplo de um sítio que seus dados foram indexados pelo Common Crawl e está em um arquivo .WAT;
- Arquivos WAT: armazenam metadados calculados para os dados armazenados no WARC;
- Arquivos WET: armazenam texto puro extraído dos dados armazenados no WARC;
- Arquivos ISO: armazenam todos os detalhes.

4.6.1.1 Formato WARC

O formato WARC são os dados brutos do rastreamento, fornecendo um mapeamento direto para o processo de rastreamento. O formato não apenas armazena a resposta HTTP dos sítios com os quais ele contata (WARC-Type: response), mas também armazena informações sobre como essas informações foram solicitadas (WARC-Type: request) e metadados no próprio processo de rastreamento (WARC-Type : metadados).

Para as próprias respostas HTTP, a resposta bruta é armazenada. Isso não inclui apenas a resposta em si, o que seria obtido se fosse baixado o arquivo, mas também as informações do cabeçalho HTTP, que podem ser usadas para obter uma série de informações interessantes. Abaixo temos um trecho de uma arquivo .WARC para exemplo.

WARC/1.0

WARC-Type: metadata

WARC-Target-URI:

http://collections-test.carli.illinois.edu/cdm/ref/collection/iwu_argus/id/1874/

WARC-Date: 2018-09-26T18:00:14Z

WARC-Record-ID: <urn:uuid:23ee8598-0240-44ef-9933-7ef85943aa03>

WARC-Refers-To: <urn:uuid:a326e476-b0de-44e2-8dab-171208cd7db2>

Content-Type: application/json

Content-Length: 1303

```
{
  "Container": {
    "Filename": "CC-MAIN-20180918130631-20180918150631-00000.warc.gz",
    "Compressed": true,
    "Offset": "43381111",
    "Gzip-Metadata": {
      "Inflated-Length": "631",
      "Footer-Length": "8",
      "Inflated-CRC": "681524181",
      "Deflate-Length": "446",
      "Header-Length": "10"
    },
    "Envelope": {
      "Format": "WARC",
      "WARC-Header-Length": "426",
      "Actual-Content-Length": "201",
      "WARC-Header-Metadata": {
        "WARC-Target-URI": "http://collections-test.carli.illinois.edu/cdm/ref/collection/iwu_argus/id/1874/",
        "WARC-Warcinfo-ID": "<urn:uuid:8d75a1d7-86c5-4788-9e02-7e28190391ba>",
        "WARC-Current-To": "<urn:uuid:a59294a3-05c7-4913-9e19-583d46e7088e>",
        "WARC-Date": "2018-09-18T14:51:19Z",
        "Content-Length": "201",
        "WARC-Record-ID": "<urn:uuid:a326e476-b0de-44e2-8dab-171208cd7db2>",
        "WARC-Type": "metadata",
        "Content-Type": "application/warc-fields",
        "Block-Digest": "sha1:HVJ74RTFOTIQAWBDJ4NDIZ6QFVQ77C3I",
        "Payload-Metadata": {
          "Actual-Content-Type": "application/metadata-fields",
          "WARC-Metadata-Metadata": {
            "Metadata-Records": [
              {
                "Value": "733",
                "Name": "fetchTimeMs"
              },
              {
                "Value": "UTF-8",
                "Name": "charset-detected"
              },
              {
                "Value": "{\\\"reliable\\\":true,\\\"text-bytes\\\":35067,\\\"languages\\\":[\\\"code\\\":\\\"en\\\",\\\"code-iso-639-3\\\":\\\"eng\\\",\\\"text-covered\\\":0.99,\\\"score\\\":921.0,\\\"name\\\":\\\"ENGLISH\\\"]}",
                "Name": "languages-cld2"
              }
            ],
            "Actual-Content-Length": "201",
            "Trailing-Slop-Length": "0"
          }
        }
      }
    }
  }
}
```

4.6.1.2 Formato de arquivo WAT

Os arquivos WAT contêm metadados importantes sobre os registros armazenados no formato WARC mencionado acima. Esses metadados são calculados para cada um dos três tipos de registros (metadados, solicitação e resposta). Se as informações rastreadas forem HTML, os metadados calculados incluirão os cabeçalhos HTTP retornados e os links (incluindo o tipo de link) listados na página.

Esta informação é armazenada como JSON. Para manter os tamanhos de arquivo o menor possível, o JSON é relativamente ilegível para humanos. Para inspecionar um arquivo JSON, pode-se usar uma das muitas ferramentas de conversão JSON disponíveis on-line. Abaixo temos um trecho de um arquivo .WAT para exemplo.

WARC/1.0

WARC-Type: metadata

WARC-Target-URI: <http://archeopasja.pl/kategoria/archeoprzewodnik/>

WARC-Date: 2018-10-01T08:19:16Z

WARC-Record-ID: <urn:uuid:92d59494-0c5f-4b04-b88d-63b4fd805a7e>

WARC-Refers-To: <urn:uuid:4bc2bb99-1a5d-49cc-b95d-184c5b9de811>

Content-Type: application/json

Content-Length: 1272

```
{"Container":{"Filename":"CC-MAIN-20180926140948-20180926161348-00061.warc.gz","Compressed":true,"Offset":"18375838","Gzip-Metadata":{"Inflated-Length":"600","Footer-Length":"8","Inflated-CRC":"433779632","Deflate-Length":"432","Header-Length":"10"}}, "Envelope":{"Format":"WARC","WARC-Header-Length":"395","Actual-Content-Length":"201","WARC-Header-Metadata":{"WARC-Target-URI":"http://archeopasja.pl/kategoria/archeoprzewodnik/","WARC-Warcinfo-ID":"<urn:uuid:8d974b65-0b36-4013-906c-d3451b09e6d2>","WARC-Concurrent-To":"<urn:uuid:49acf4f7-015d-41fb-bcc7-774db07763f0>","WARC-Date":"2018-09-26T15:19:01Z","Content-Length":"201","WARC-Record-ID":"<urn:uuid:4bc2bb99-1a5d-49cc-b95d-184c5b9de811>","WARC-Type":"metadata","Content-Type":"application/warc-fields"},"Block-Digest":"sha1:Y3HYKUHLV3F44JCG3DOBYTD6RA7D7X AQ","Payload-Metadata":{"Actual-Content-Type":"application/metadata-fields","WARC-Metadata-Metadata":{"Metadata-Records":[{"Value":"1768","Name":"fetchTimeMs"}, {"Value":"UTF-8","Name":"charset-detected"}, {"Value":{"reliable":true,"text-bytes":5684,"languages":[{"code":"pl","code-iso-639-3":"pol","text-covered":0.99,"score":1174.0,"name":"POLISH"}]},"Name":"languages-cld2"}]},"Actual-Content-Length":"201","Trailing-Slop-Length":"0"}}}
```

4.6.1.3 Formato de arquivo WET

Como muitas tarefas exigem apenas informações textuais, o conjunto de dados do Common Crawl fornece arquivos WET que contêm apenas texto simples extraído. A maneira

como esses dados textuais são armazenados no formato WET é bem simples. Os metadados do WARC contêm vários detalhes, incluindo a URL e o comprimento dos dados em texto plano. Abaixo temos um trecho de uma arquivo .WAT para exemplo.

```
WARC/1.0
WARC-Type: conversion
WARC-Target-URI:
http://caro.tv/products/super-bright-car-headlights-h7-led-h8h11-hb39005-hb49006-70w-7000
lm-auto-front-bulb-automobiles-headlamp-6000k-car-lighting/
WARC-Date: 2018-09-26T14:44:47Z
WARC-Record-ID: <urn:uuid:9fc41209-e535-4a1f-abcf-4284d77d0921>
WARC-Refers-To: <urn:uuid:8e633dbf-ee54-4b25-afa1-5f30c72c83c0>
WARC-Block-Digest: sha1:QZEBSTSLJLVNXUH6UYYBOF4BS5GLKJZM
Content-Type: text/plain
Content-Length: 9631
Super Bright Car Headlights H7 LED H8/H11 HB3/9005 HB4/9006 70W 7000lm
Auto Front Bulb Automobiles Headlamp 6000K Car Lighting > Caro.tv Caro.tv
Products
Buyer Protection
Shipping & Delivery
Free Worldwide Shipping
Categories
Dvd Players
Cheap DVD Players
DVD Players with GPS
OEM Players
All categories
Home > Products > Car LED lights > Super Bright Car Headlights H7 LED H8/H11
HB3/9005 HB4/9006 70W 7000lm Auto Front Bulb Automobiles Headlamp 6000K Car
Lighting
Super Bright Car Headlights H7 LED H8/H11 HB3/9005 HB4/9006 70W 7000lm
Auto Front Bulb Automobiles Headlamp 6000K Car Lighting
```

96 % of buyers enjoyed this product!

In stock

...

4.7 Padrão do Common Crawl

O Common Crawl segue a ISO 28500 da qual já está na versão ISO 28500:2017, mas a documentação desta versão é paga, para este trabalho foi estudada a versão ISO 28500:2009. Este documento é um conjunto de normas para padronização de processos que visam dar diretrizes básicas para o processo de indexação de informações da internet, sendo assim o Common Crawl segue um padrão para indexar conteúdo da internet, é importante frisar que estas normativas da ISO independente de quais sejam não obrigam ou flexibilizam os processos que se destinam a padronizá-los porém dão as características básicas para se ter um processo padronizado (ISO, 2014?).

O principal desafio nesta parte do trabalho foi a documentação não mencionar muitas informações encontradas nos arquivos disponibilizados do Common Crawl.

4.8 Portais de dados abertos

Um portal de dados abertos é o ponto central para a busca e o acesso a dados públicos, onde qualquer pessoa com acesso a web pode acessar e encontrar conjuntos de dados (GOVERNO DIGITAL, 2017). Um conjunto de dados também pode ser chamado de dataset e é o principal insumo dos processos de análise de dados, são normalmente representados por dados tabulares em formato de planilhas onde as linhas são os registros dos acontecimentos e as colunas são as características desses acontecimentos (AQUARELA, 2018).

4.9 Plataformas de dados abertos utilizadas neste trabalho

As plataformas de dados abertos em resumo são um poderoso sistema de gerenciamento de dados que torna os dados acessíveis, fornecendo ferramentas para simplificar a publicação, compartilhamento, localização e uso de dados (CKAN, 2018), elas normalmente são usadas para a criação de portais de dados abertos. Neste trabalho foram utilizados os resultados obtidos de “Investigating open data portals automatically: a methodology and some illustrations” (CORRÊA; ZANDER; DA SILVA, 2018) como ponto

de partida foram utilizadas as mesmas plataformas de dados abertos, foram elas, CKAN, Socrata, openDataSoft e ArcGIS Open Data.

4.9.1 CKAN

A plataforma de dados abertos CKAN leva o mesmo nome da associação responsável por gerenciar e supervisionar ela, esta fundação foi fundada em 2014.

A plataforma CKAN é uma solução de gerenciamento de dados e portal de dados completo e open source, ela oferece uma maneira simplificada de tornar seus dados detectáveis e apresentáveis. Cada conjunto de dados recebe sua própria página para listar os recursos de dados e uma rica coleção de metadados, tornando-o um catálogo de dados valioso e de fácil pesquisa (CKAN, 2018).

4.9.2 Socrata

A empresa Socrata foi fundada em 2007 com o objetivo de criar uma plataforma em nuvem que permitisse que organizações do setor público pudessem gerenciar e compartilhar facilmente seus dados. A plataforma tem o mesmo nome da empresa promete tornar os dados acessíveis para qualquer pessoa com uma conexão a internet, ao mesmo tempo em que também lhes dava o poder de visualizar e analisar os dados com facilidade.

A empresa Socrata se diz líder de mercado em tornar os dados existentes do governo detectáveis, utilizáveis e acionáveis para os funcionários do governo e as pessoas que eles servem. O Socrata fornece uma plataforma de dados como serviço e aplicativos de nuvem exclusivamente para organizações governamentais municipais, estaduais e federais (TYLER TECHNOLOGIES, 2018). A plataforma Socrata não é de código aberto e para sua plena utilização deve-se pagar assim como os serviços que a empresa Socrata também oferece.

4.9.3 OpenDataSoft

Fundada em 2011 a empresa OpenDataSoft é responsável pela plataforma de dados abertos que tem o mesmo nome da empresa que oferece soluções de compartilhamento de dados que ela diz ser líderes de mercado que redefinem o gerenciamento de dados corporativos. O OpenDataSoft oferece um repositório único de dados organizados e ferramentas para acessá-los, visualizá-los, analisá-los e compartilhá-los. O OpenDataSoft é uma plataforma paga e de código fechado (OPENDATASOFT, 2018).

4.9.4 ArcGIS OpenData

A plataforma de dados abertos ArcGIS OpenData pertence a empresa Esri fundada em 1969, esta empresa se especializou em dados geográficos, o ArcGIS segundo a Esri é o mais poderoso software de mapeamento e análise espacial do mundo. Ele aplica um conceito chamado Science of Where para conectar todos, em qualquer lugar através de uma linguagem visual comum. Ele combina mapeamento e análise para revelar uma visão mais profunda dos dados, ideal para ser utilizado por construtoras e consultoras de engenharia civil, mas também é utilizados para mostrar dados como vendas em mapas (ESRI, 2018).

4.10 Assinaturas

Todas as quatro plataformas retornam um arquivo JSON com informações específicas de cada uma quando invocadas. Três das quatro plataformas de dados abertos catalogadas no trabalho de Correa, Zander e da Silva (2018) têm chamadas para a assinatura diferentes, com exceção da ArcGIS Open Data que é a mesma do OpenDataSoft, porém com retorno específico para esta plataforma, conforme mostrado na Tabela 1.

Tabela 1. Plataformas e suas respectivas assinaturas

Plataforma	Assinatura	URL sítio de dataset
CKAN	/api/action/site_read	http://opendata.hccg.gov.tw
Socrata	/api/catalog/v1	https://www.data.act.gov.au
OpenDataSoft	/api/v2	http://data.outdoornebraska.gov
ArcGIS Open Data	/api/v2	http://data.outdoornebraska.gov

Fonte: Elaborado pelo autor (2018)

Quando essas assinaturas são adicionadas após a URL de domínio de um sítio onde se encontra um portal de dados abertos (denominado Endpoint), o retorno deve ser um arquivo em formato JSON, por exemplo, quando colocamos no browser a URL www.data.rio (sítio oficial de datasets do Rio de Janeiro) e no seu fim adicionamos a assinatura do ArcGIS (Plataforma usada neste caso) <https://www.data.rio/api/v2> nos obtemos um retorno em JSON

com dados deste portal de dados abertos, mas se colocarmos outra assinatura isso não acontece. Abaixo está um trecho exemplo do JSON retornado.

```
{ ...  
  
  { "collection": "pages": {  
    "collection": "https://www.data.rio/api/v2/pages",  
    "object": "https://www.data.rio/api/v2/pages/{:id}"  
  },  
  "params": {  
    "q": "search term",  
    "page": {  
      "size": "page size for each results. defaults to 10",  
      "number": "page number for results. defaults to 1"  
    },  
    "include": "related resources to include in search results",  
    "fields": "subset of fields for resources queried or included",  
    "filter": "a filter applied to search results"  
  }  
}
```

5 Métodos

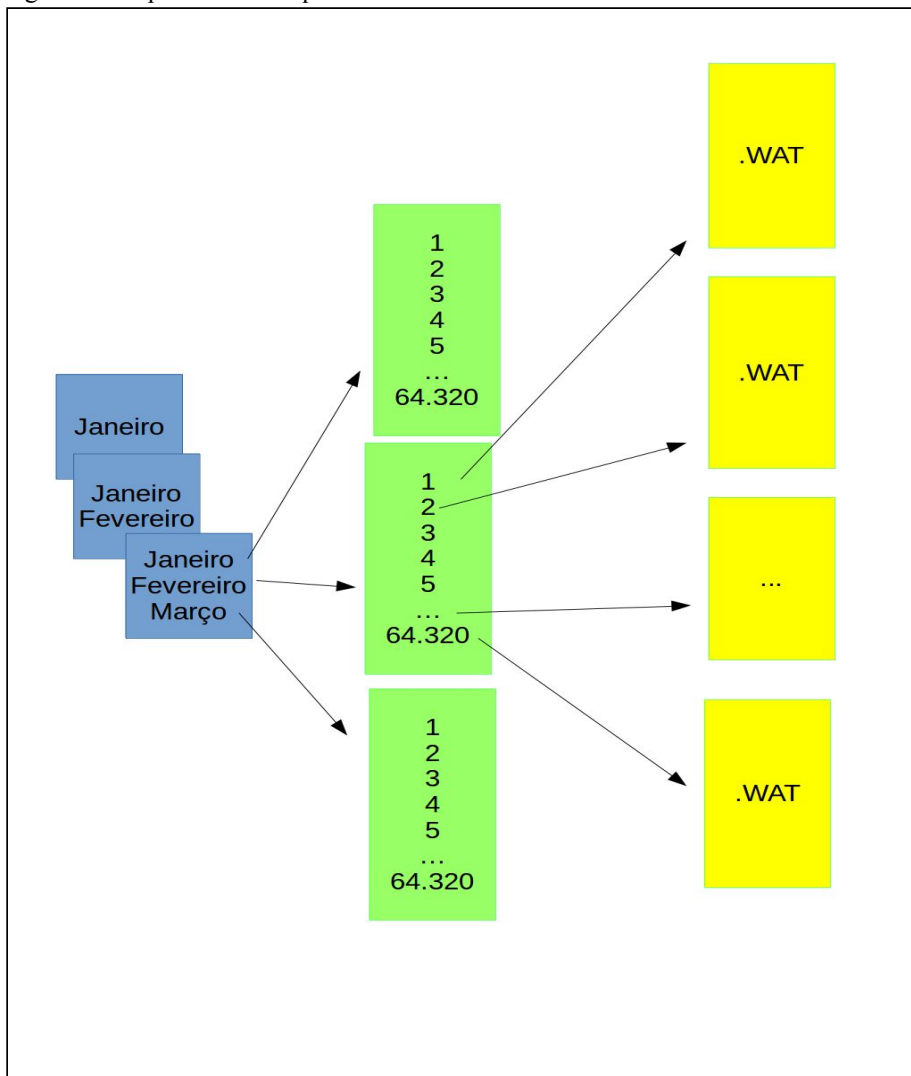
Neste capítulo será demonstrado as etapas que levaram a obtenção dos resultados demonstrados no capítulo “Resultados”.

5.1 Arquitetura de arquivos do projeto Common Crawl

Entender esta arquitetura foi muito importante para a criação da lógica para automatização da coleta de dados uma vez que é inviável fazer o download de toda uma indexação mensal já que ela passa de 220 Terabytes de dados.

A Figura 2 representa a arquitetura de arquivos do Common Crawl, em azul temos os retângulos que representam as indexações que podem ser encontradas no site do Common Crawl. A cada mês é feita uma nova indexação com o nome do mês que ele representa. Para cada indexação é gerado um arquivo que em média tem 7,3 MB representado pelo retângulo verde, contendo em média 64.320 (sesenta e quatro mil trezentos e vinte) linhas, onde cada linha representa uma URL para uma fração da indexação, que está contido em um arquivo .WAT, representado pelo retângulo amarelo, em média cada arquivo .WAT quando descompactado contém aproximadamente 6,5 GB de conteúdo.

Figura 2 . Arquitetura de arquivos do Common Crawl



Fonte: Elaborado pelo autor(2018)

5.2 Acessando os dados do Common Crawl

Para cumprir com os objetivos deste trabalho, foi necessário automatizar o acesso e a obtenção de dados do projeto do Common Crawl, por este motivo foram realizadas todas as etapas a seguir, para tal foram utilizadas as bibliotecas Requests e BeautifulSoup do Python.

5.2.1 Obtendo as URLs de indexações

Acessando a página² que contém as URLs para todas as indexações realizadas pelo Common Crawl, foi possível obter todo conteúdo HTML da página, em seguida foi criada

² <https://commoncrawl.s3.amazonaws.com/cc-index/collections/index.html>

uma lista de IDs de URLs de indexações e foi escolhida a última por acreditar que está sempre vai estar mais atualizada do que as anteriores.

5.2.2 Escolhendo o tipo de arquivo

O próximo passo foi usar o ID correspondente a última indexação para acessar a página³ onde podemos escolher o tipo de arquivo para baixar. Cada indexação tem um conjunto de arquivos de diferentes extensões e para diferentes finalidades e por este motivo foi preciso especificar qual tipo de arquivo baixar. No caso deste trabalho foram utilizados os arquivos .WAT representado pelo pós fixo wat.paths.gz. Este arquivo após descompactado tem o tamanho médio de 7 MB. É neste arquivo que tem aproximadamente 64 mil URLs para frações da indexação da qual escolhemos usando seu ID no início deste passo.

5.2.3 Baixando o arquivo .WAT

Para baixar cada arquivo referente a uma fração da indexação foi adicionado a URL padrão⁴ a URL que está no arquivo.WAT, deixando a URL completa⁵.

5.3 Algoritmo para buscar URLs que tenham .gov

O arquivo .WAT tem diversos dados de inúmeros sítios sendo que a maioria não pertencem a sítios governamentais por este motivo foi preciso criar estratégias para melhorar a performance do algoritmo para analisar somente o necessário.

Analisando os dados foi descoberto que a URL fica na mesma linha que a tag WARC-Target-URI, então o algoritmo procura esta linha e por consequência encontra a URL. Assim que é encontrado salva esta URL no banco de dados. Foi selecionada como importante a característica da URL ter .gov por ela ser comum em URLs governamentais.

³ <https://commoncrawl.s3.amazonaws.com/crawl-data/CC-MAIN-2018-39/index.html>

⁴ <https://commoncrawl.s3.amazonaws.com/>

⁵

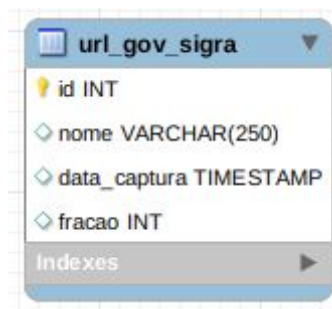
<https://commoncrawl.s3.amazonaws.com/crawl-data/CC-MAIN-2018-39/segments/1537267155413.17/wat/CC-MAIN-20180918130631-20180918150631-00000.warc.wat.gz>

5.4 Banco de dados

Para a finalidade deste trabalho houve a necessidade da criação de somente uma tabela de banco de dados, ela serve para guardar as URLs que tenham .gov, esta tabela contém os seguintes campos:

- id - Para identificação única de cada URL;
- nome - Guarda a URL;
- data_captura - Guarda a data e hora que a URL foi encontrada;
- fracao - Guarda o número da fração da indexação da qual foi encontrada a URL.

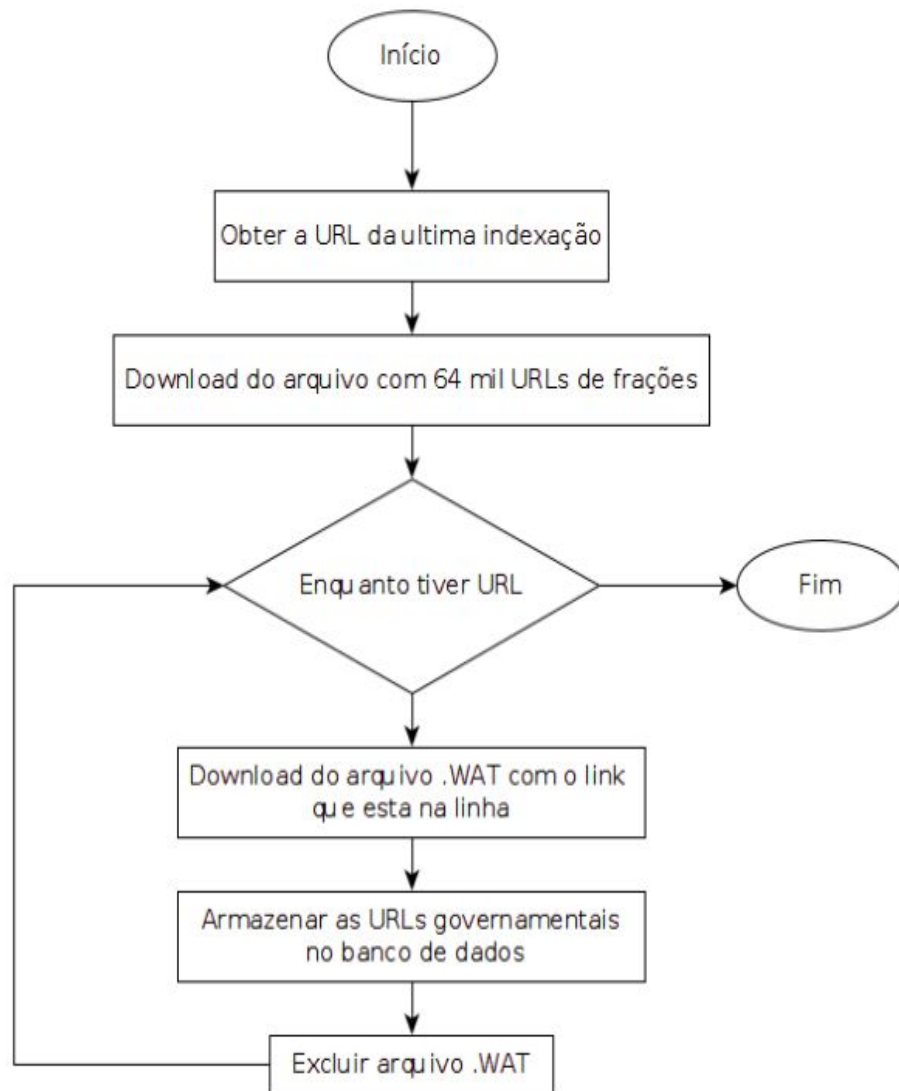
Figura 3 . Tabela no banco de dados



Fonte: Elaborado pelo autor (2018)

A Figura 4 representa o fluxograma dos passos realizados até aqui.

Figura 4. Fluxograma das etapas para se obter as URLs governamentais



Fonte: Elaborado pelo autor (2018)

5.5 Determinando de qual país a URL pertence

Foi importante saber a localização que cada URL pertence para analisar quais países têm mais portais de dados abertos. Para isso, foram utilizadas as bibliotecas Pandas, Requests e BeautifulSoup.

Observou-se que a maioria das URLs que contêm .gov tem uma sigla depois do .gov, que normalmente identifica seu país de origem, com exceção dos Estados Unidos da América. Por este motivo foi extraído o HTML do sítio Dataset Publishing Language⁶ onde tem a sigla, latitude, longitude e nome de cada país. Uma característica importante nos dados de latitude e

⁶ https://developers.google.com/public-data/docs/canonical/countries_csv

longitude fornecidos neste sítio é que eles usam as coordenadas grau decimal e não as coordenadas grau minuto segundo que é a que o Google Maps usa, as coordenadas grau decimal são as utilizadas na biblioteca Plotly e esta biblioteca é a que foi utilizada neste trabalho para criar as imagens geográficas.

1. Foi gerado através do MySQL um arquivo CSV com todas URLs que contenham trechos com alguma dessas palavras-chave: free ou trans ou open ou data;
2. Utilizando pandas foi criada a Tabela 2 com os dados deste arquivo CSV.

Tabela 2 - Tabela criada com os dados do banco de dados.

id	url	data_captura	indice_fracao	sigla
10392	https://freeflow.mdba.gov.au	2018-08-25 22:09:40	0	au
63610	http://ftp4.de.freessbie.org/pub/misc/conferenc...	2018-08-30 06:59:55	1910	us
100473	https://handicap-free.nat.gov.tw	2018-09-01 07:12:49	2946	tw
118952	https://www.handicap-free.nat.gov.tw	2018-09-04 02:43:21	3435	tw
6176	http://data.environment.nsw.gov.au	2018-09-11 07:03:23	3793	au

Fonte: Elaborado pelo autor (2018)

3. Foi preciso limpar os dados do HTML capturado anteriormente para capturar somente as siglas, latitude, longitude e nome do país, conforme especificado na Tabela 3.

Tabela 3 - Tabela criada com dados retirados do sítio

sigla	latitude	longitude	pais
ad	42.546245	1.60155	Andorra
ae	23.424076	53.84781	United Arab Emirates
af	33.93911	67.70995	Afghanistan
ag	17.060816	-61.79642	Antigua and Barbuda
ai	18.220554	-63.06861	Anguilla
al	41.153332	20.16833	Albania
am	40.069099	45.03818	Armenia
an	12.226079	-69.06008	Netherlands Antilles
ao	-11.202692	17.87388	Angola
aq	-75.250973	-0.07138	Antarctica
ar	-38.416097	-63.61667	Argentina

Fonte: Elaborado pelo autor (2018)

4. Foi mesclado as duas tabelas usando como coluna de referência a coluna sigla que é comum nas duas tabelas, usando como método o inner para que a mesclagem fosse realizada usando os dados da coluna sigla comum nas duas tabelas. Na Tabela 4 é possível ver a junção dos registros das duas tabelas contendo os registros de nome id, url, data_captura, indice_fraao, sigla, plataforma_opendata, latitude, longitude e pais.

Tabela 4 - Tabela com dados mesclados

id	url	data_captura	indice_fraao	sigla	plataforma_opendata	latitude	longitude	pais
10392	https://freeflow.mdba.gov.au	2018-08-25 22:09:40	0	au	NaN	-25.274398	133.77513	Australia
63610	http://ftp4.de.freessbie.org/pub/misc/conferenc...	2018-08-30 06:59:55	1910	us	NaN	37.090240	-95.71289	United States
100473	https://handicap-free.nat.gov.tw	2018-09-01 07:12:49	2946	tw	NaN	23.697810	120.96051	Taiwan
118952	https://www.handicap-free.nat.gov.tw	2018-09-04 02:43:21	3435	tw	NaN	23.697810	120.96051	Taiwan
6176	http://data.environment.nsw.gov.au	2018-09-11 07:03:23	3793	au	NaN	-25.274398	133.77513	Australia
17602	https://www.transportation.gov/pr	2018-09-11 14:45:47	4096	us	NaN	37.090240	-95.71289	United States
21059	https://data.nj.gov/	2018-09-11 17:15:29	4191	us	NaN	37.090240	-95.71289	United States

Fonte: Elaborado pelo autor (2018)

5.6 Sinônimos de .gov

O .gov é usado como padrão em muitos sítios governamentais, como é o caso do Brasil e Estados Unidos da América, no início deste trabalho acreditou-se que somente utilizando este padrão seria possível catalogar muitos portais de dados abertos governamentais de diferentes países, mas assim que começamos a obter resultados foi verificado que tínhamos um problema, países como Japão, e muitos países europeus que são referência em portais de dados abertos não apareceram, foi quando começou uma investigação onde se descobriu que muitos países não usam o .gov como indicativo de sítio governamental e por isso foi criada a Tabela 5 onde é mostrado o sinônimo de .gov ou palavra característica de 57 países. Estas informações poderão ser utilizadas em futuros trabalhos.

Tabela 5 - Tabela de sinônimos ou palavras referentes a sítios governamentais

Sinônimo de .gov	País
.dk	Dinamarca
.govt.nz	Nova Zelândia
.fi	Finlândia
.se	Suécia
.swiss	Suíça
/no/ ou .no	Noruega
/sg ou .sg ou singapore_opendata	Singapura
.nl	Holanda
.ca	Canadá
/de ou .de	Alemanha
.lu	Luxemburgo
.uk	Reino Unido
.au	Austrália
.is ou /iceland	Islândia
.be	Bélgica
.kh	Hong Kong
gv.at ou .at	Áustria
.gov/	Estados Unidos
.ie/	Irlanda
go.jp	Japão
.eu	Europa
.al	Albania
/andorra ou .ad	Andorra
/belarus ou .by	Bielorrússia

/bosnia-and-herzegovina ou .ba	Bósnia
.bg	Bulgária
open-data-croatia ou data-zagreb-hr ou /cratia ou .hr	Croácia
.sk	Eslováquia
.si	Eslovênia
.es ou /es	Espanha
.ee	Estônia
.fr	França
.gr	Grécia
/hu ou .hu	Hungria
.it	Itália
.lv	Letônia
/lechtenstein	Liechtenstein
.lt	Lituânia
.mt	Malta
.md	Moldávia
.mc	Mônaco
/montenegro	Montenegro
.nl	Amsterdã
.pl	Polônia
/datasets	Também pode ser utilizado como busca generalizada
.pt	Portugal
.cz	República Checa
.mk ou /macedonia-open-data	República de Macedônia
.ro	Romênia
.rs	Sérvia

.am	Armênia
.az	Azerbaijão
.cy	Chipre
.ge	Georgia
.kz	Casaquistão
.ru	Rússia
.tr	Turquia

Fonte: Elaborado pelo autor (2018)

5.7 Dificuldades encontradas

Pelo tamanho da indexação mensal foi necessário deixar o algoritmo rodando em um servidor com internet de boa qualidade o que não impediu que houvesse problemas de indisponibilidade ou lentidão de download, pois o Common Crawl desconecta o usuário ou diminui a velocidade de envio de arquivos depois de alguns dias de conexão o que gerou perda de conexão. O Common Crawl também dá a opção de usar sua API para buscar URLs mas deixa bem claro em seu sítio que ela deve ser utilizada com cautela para não deixar os seus servidores sobrecarregados assim o seu melhor uso é para pequenas buscas o que não foi o caso deste trabalho uma vez que deveria analisar toda uma indexação mensal que tem mais de 220 Terabytes de dados. Uma infraestrutura deve ser montada para que o algoritmo rode constantemente durante meses pois mesmo com uma internet de boa velocidade após 28 dias foi analisado somente 39,8 % de uma indexação mensal. O padrão de pós-fixos .gov não se aplica para a muitos países, a tabela de sinônimos traz uma possível solução para isso aumentando a eficiência nas buscas mas diminuindo a performance do algoritmo quando os sinônimos forem utilizados.

5.8 Disponibilização do código

Todo o código desenvolvido para a realização deste trabalho pode ser encontrado no Github⁷.

⁷ https://github.com/TurcoDick/DataScience_open_data_TCC

6 Resultados

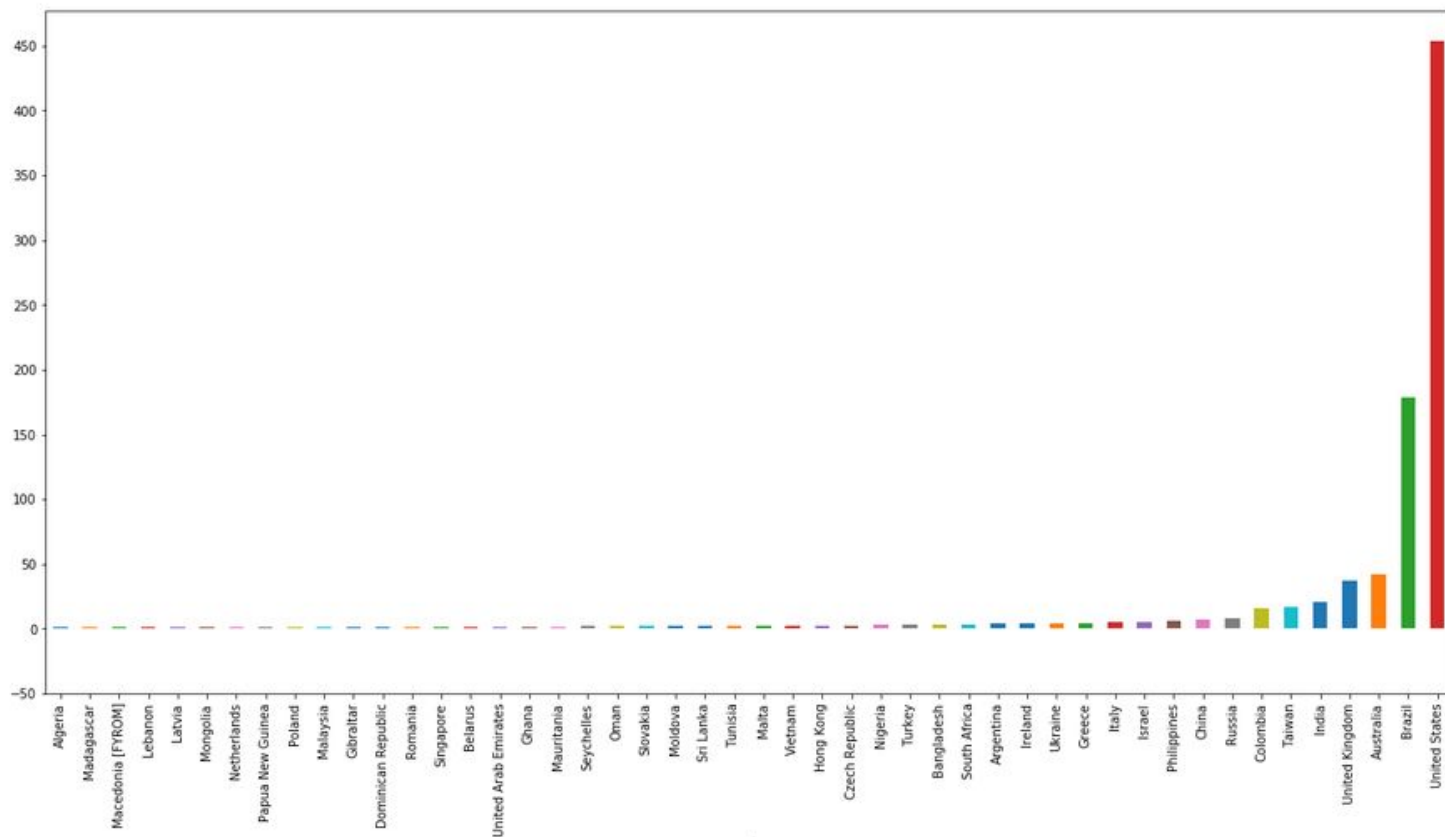
Neste capítulo será apresentado os resultados obtidos neste trabalho e os objetivos alcançados. A data da atualização dos resultados é de 16 de Outubro de 2018. Foram encontradas 897.524 (Oitocentos e noventa e sete mil quinhentos e vinte quatro) URLs governamentais, dessas URLs foram filtradas um grupo de URLs que fazem uso de alguma palavra-chave, com este segundo grupo de URLs foram usadas as assinaturas das plataformas gerando uma base de dados de URLs de portais de dados abertos governamentais.

6.1 Quantidade de URLs .gov com as palavras-chave selecionadas

Para a otimização da performance do método de análise foi selecionado um conjunto de palavras-chave em inglês referentes ao significado de portais de dados abertos, são elas: free que significa livre, open que significa aberto, trans que é abreviação de transparency que significa transparência e data que significa dados.

Foram encontradas 862 URLs com alguma palavras-chave em sua URL, na Figura 5 é possível ver a quantidade de URLs por país, os mesmos dados também podem ser vistos na Tabela 6.

Figura 5. Gráfico de URLs .gov com palavras-chave por país



Fonte: Elaborado pelo autor (2018)

Tabela 6 - Tabela de URLs .gov com palavras-chave selecionadas

País	quantidade de URLs encontradas
United States	454
Brazil	179
Australia	42
United Kingdom	37
India	21
Taiwan	17
Colombia	16
Russia	8
China	7

Philippines	6
Israel	5
Italy	5
Greece	4
Ukraine	4
Ireland	4
Argentina	4
South Africa	3
Bangladesh	3
Turkey	3
Nigeria	3
Czech Republic	2
Hong Kong	2
Vietnam	2
Malta	2
Tunisia	2
Sri Lanka	2
Moldova	2
Slovakia	2
Oman	2
Seychelles	2
Mauritania	1
Ghana	1
United Arab Emirates	1
Belarus	1
Singapore	1
Romania	1

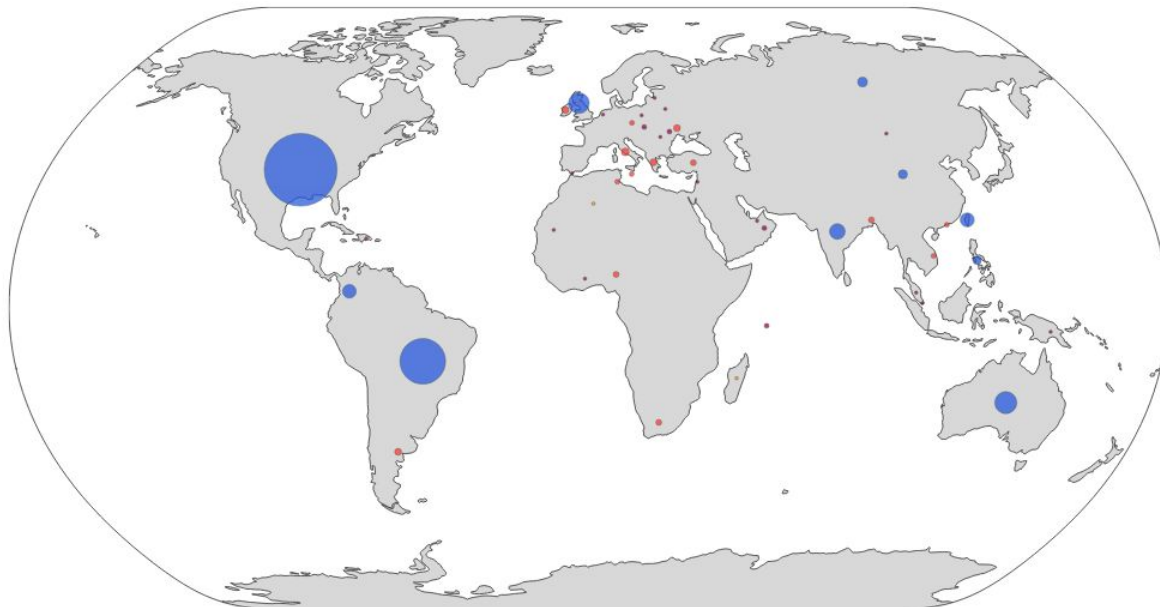
Dominican Republic	1
Gibraltar	1
Malaysia	1
Poland	1
Papua New Guinea	1
Netherlands	1
Mongolia	1
Latvia	1
Lebanon	1
Macedonia [FYROM]	1
Madagascar	1
Algeria	1

Fonte: Elaborado pelo autor (2018)

A Figura 6 mostra geograficamente o que a Tabela 6 quantifica, ela mostra onde estão as maiores concentrações de URLs governamentais com as palavras-chave seleccionadas encontradas por este trabalho, foram utilizadas cores diferentes somente para dar destaque nos pontos pequenos, esta visualização pode ser vista dinamicamente no sítio⁸.

⁸ <https://plot.ly/~AlisonLucio/8>

Figura 6 . Imagem geográfica das URLs encontradas com palavras-chave.

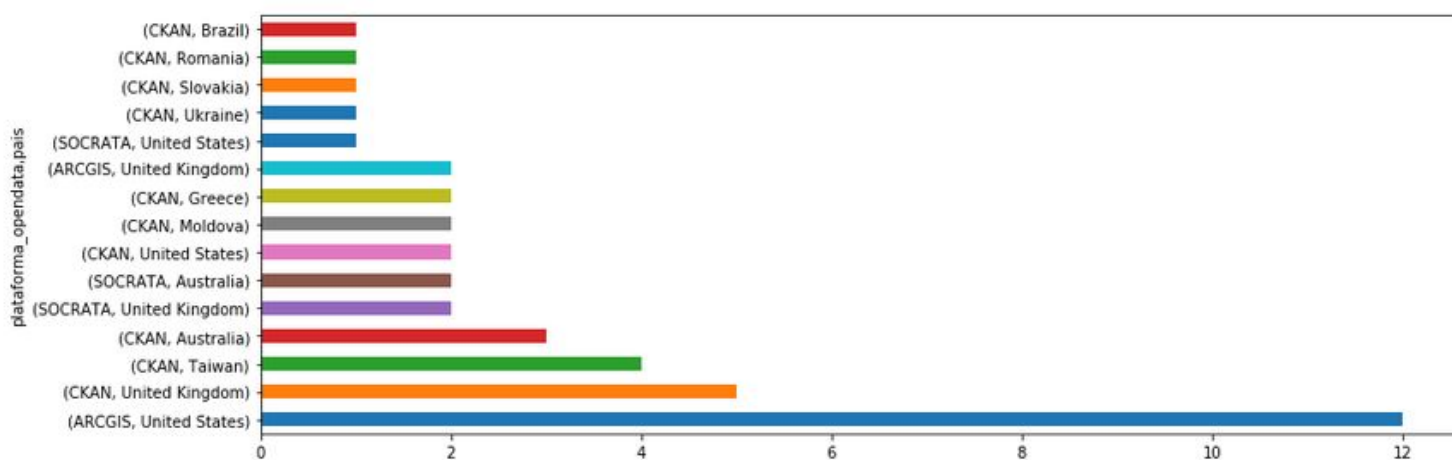


Fonte: Elaborado pelo autor (2018)

6.2 Quantidade de URLs de portais de dados abertos encontradas

Foram encontradas 41 URLs de portais de dados abertos. Na Figura 7 e na Tabela 7 é possível ver a quantidade de portais de dados abertos com sua respectiva plataforma de disponibilização de dados abertos por país.

Figura 7 . Gráfico de portais de dados abertos e suas plataformas por país.



Fonte: Elaborado pelo autor (2018)

Tabela 7 - Quantidade de portais de dados abertos encontradas por país e plataforma

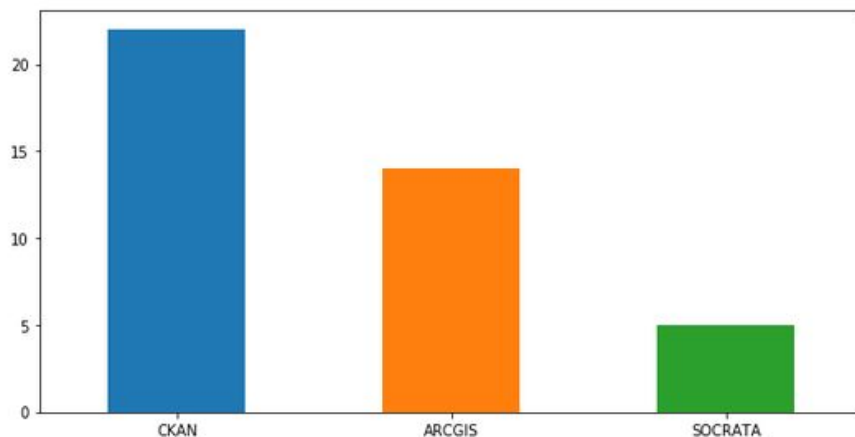
país	plataforma opendata	quantidade
United States	ARCGIS	12
United Kingdom	CKAN	5
Taiwan	CKAN	4
Australia	CKAN	3
United States	CKAN	2
United Kingdom	SOCRATA	2
United Kingdom	ARCGIS	2
Moldova	CKAN	2
Greece	CKAN	2
Australia	SOCRATA	2
United States	SOCRATA	1
Ukraine	CKAN	1
Slovakia	CKAN	1
Romania	CKAN	1
Brazil	CKAN	1

Fonte: Elaborado pelo autor (2018)

6.3 Quantidade de URLs de portais de dados abertos encontrados por plataforma

A Figura 8 e a Tabela 8 mostram a quantidade total de portais de dados abertos encontrados por plataforma.

Figura 8. Gráfico da quantidade de portais de dados abertos por plataforma.



Fonte: Elaborado pelo autor (2018)

Tabela 8 - Quantidade de portais de dados abertos encontrados por plataforma

Plataforma	Quantidade
CKAN	22
ARCGIS	14
SOCRATA	5

Fonte: Elaborado pelo autor (2018)

A Figura 9 mostra o que a Tabela 8 quantifica, a localização dos portais de dados abertos encontrados. Para esta ilustração só foi utilizada a localização por país e não por estado ou cidade sendo assim só existe um ponto em cada país com algum portal de dados abertos independente da quantidade encontrada. Esta imagem pode ser vista dinamicamente no sítio⁹.

⁹ <https://plot.ly/~AlisonLucio/6/quantidade-de-urls-que-respondem-a-alguma-assinatura/#/>

Figura 9 - Imagem geográfica das URLs de portais de dados abertos encontrados.



Fonte: Elaborado pelo autor (2018)

6.4 As URLs de portais de dados abertos encontradas

A Tabela 9 mostra a URL de cada portal de dados abertos encontrado, nela é possível verificar a URL, sigla, nome do país e a plataforma do portal de dados abertos utilizada.

Tabela 9 - Lista de URL de portais de dados abertos encontrados

URL	Sigla	País	Plataforma
http://opendata.suffolkcountyny.gov/	us	United States	ARCGIS
https://data.bloomington.in.gov/	us	United States	CKAN
https://data.cdc.gov.tw	tw	Taiwan	CKAN
http://opendata.wigan.gov.uk	gb	United Kingdom	ARCGIS
http://opendata.hccg.gov.tw	tw	Taiwan	CKAN
https://www.opencrete.gov.gr	gr	Greece	CKAN
http://opendata.slocounty.ca.gov/	us	United States	ARCGIS
http://data.outdoornebraska.gov/	us	United States	ARCGIS
https://www.data.act.gov.au	au	Australia	SOCRATA

http://opendata.firstmap.delaware.gov/	us	United States	ARCGIS
https://opendata.camden.gov.uk	gb	United Kingdom	SOCRATA
http://data.gov.ua	ua	Ukraine	CKAN
http://data.gov.ro	ro	Romania	CKAN
https://data.gov.au	au	Australia	CKAN
http://data.roanokecountyva.gov/	us	United States	ARCGIS
https://data.melbourne.vic.gov.au	au	Australia	SOCRATA
https://www.data.vic.gov.au	au	Australia	CKAN
https://catalogue.data.wa.gov.au	au	Australia	CKAN
https://data.gov.uk	gb	United Kingdom	CKAN
https://data.gov.sk	sk	Slovakia	CKAN
https://www.opendatani.gov.uk	gb	United Kingdom	CKAN
http://data.taichung.gov.tw	tw	Taiwan	CKAN
http://opendata.columbus.gov/	us	United States	ARCGIS
http://datalb.longbeach.gov/	us	United States	ARCGIS
https://opendata.cheshireeast.gov.uk	gb	United Kingdom	SOCRATA
http://openmappingdata.lambeth.gov.uk	gb	United Kingdom	ARCGIS
http://opendata.yakimawa.gov/	us	United States	ARCGIS
http://www.data.gov.md	md	Moldova	CKAN
https://openpaymentsdata.cms.gov/	us	United States	SOCRATA
https://data.birmingham.gov.uk	gb	United Kingdom	CKAN
http://dados.transportes.gov.br	br	Brazil	CKAN
http://www.data.gov.gr	gr	Greece	CKAN

http://opendata.e-land.gov.tw	tw	Taiwan	CKAN
http://data.belmont.gov/	us	United States	ARCGIS
http://data.gov.md	md	Moldova	CKAN
http://geodata.vermont.gov/	us	United States	ARCGIS
http://opendata.dc.gov/	us	United States	ARCGIS
http://anrgeodata.vermont.gov/	us	United States	ARCGIS
https://data.dundee.city.gov.uk	gb	United Kingdom	CKAN
https://data.london.gov.uk	gb	United Kingdom	CKAN
http://opendata.fortsmithar.gov/	us	United States	CKAN

Fonte: Elaborado pelo autor (2018)

7 Considerações finais

Este trabalho mostrou que é possível catalogar URLs de portais de dados abertos governamentais de uma forma automatizada. Foram analisados 30.122 frações da indexação de 6,5 GB em média cada o que corresponde à aproximadamente 67 Terabyte de dados analisados onde identificou-se 897.524 URLs governamentais, sendo que destas 8.024 continham uma das palavras-chave consideradas para identificação de portais de dados. Deste grupo de URLs foram encontradas e catalogadas 41 URLs de portais de dados abertos espalhados pelo mundo. Foi criada uma tabela com 57 sinônimos de .gov utilizados em outros idiomas/culturas que identificam sítios governamentais. Estes sinônimos, quando utilizados, deverão aumentar a quantidade de portais de dados abertos encontrados, uma vez que este trabalho só utilizou o pós-fixado .gov.

O método desenvolvido foi executado durante 28 dias e conseguiu analisar 39,8% de uma indexação mensal de 220 terabytes de dados do Common Crawl. Apesar de não ter sido possível analisar completamente o arquivo de indexação, devido às limitações de infraestrutura computacional, os resultados obtidos mostraram que o método pode ser reproduzido em escala e frequência maiores.

8 Trabalhos futuros

Como trabalhos futuros deverão ser realizados os seguintes trabalhos: Comparação entre os portais de dados abertos encontrados pela metodologia desenvolvida no trabalho “Método de catalogação de URLs com base no Common Crawl” com portais de dados abertos já conhecidos como o dados.gov.br e data.gov para saber o grau de maturidade do método desenvolvido. A adição dos sinônimos de .gov ao método para coletar e catalogar mais URLs de portais de dados abertos de todo o mundo. Desenvolver um método para burlar a diminuição de velocidade do Common Crawl explicado no capítulo “Dificuldades encontradas”. E como último trabalho futuro o desenvolver um método de catalogação de URLs que não use o Common Crawl.

9 Referências

APOIA.SE (Brasil). **Operação Serenata de Amor**. 2018. Disponível em: <<https://apoia.se/serenata>>. Acesso em: 10 abr. 2018.

AQUARELA. **Datasets, o que são e como utilizá-los**. 2018. Disponível em: <<https://www.aquare.la/datasets-o-que-sao-e-como-utiliza-los/>>. Acesso em: 28 out. 2018.

BRASIL. **Constituição**. República Federativa do Brasil de 1988. TÍTULO I DOS PRINCÍPIOS FUNDAMENTAIS. Brasília, DF: Senado Federal, 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 7 abr. 2018.

BRASIL. Senado Federal. Senado. **Transparência: O que é transparência pública?**. 2016?. Disponível em: <<https://www12.senado.leg.br/transparencia/sobre-1>>. Acesso em: 01 jul. 2018.

BRASIL. Casa Civil. Decreto n. 12.527 18 de novembro de 2011. **Diário Oficial da União**. Brasília 18 de novembro de 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112527.htm>. Acesso em: 1 abr. 2018.

CKAN Association (Estados Unidos) (Org.). **CKAN Association**. 2018. Disponível em: <<https://ckan.org/about/association/>>. Acesso em: 13 out. 2018.

COMMON CRAWL. COMMON CRAWL. **COMMONCRAWL.ORG**. 2015. DISPONÍVEL EM: <[HTTPS://COMMONCRAWL.ORG/](https://commoncrawl.org/)>. ACESSO EM: 2 ABR. 2018.

CORRÊA, Andreiuid Sheffer. **opendataportalsurvey**. **GitHub**. Campinas, 2017. Disponível em: <<https://github.com/Andreiuid/opendataportalssurvey>>. Acesso em: 10 mar. 2018.

CORRÊA, Andreiuid Sheffer. **Uma arquitetura de referência colaborativa para estruturação de dados abertos governamentais**. São Paulo, 2017. 222 p Tese (Engenharia de Computação) - Escola Politécnica da Universidade de São Paulo.

CORRÊA, Andreiuid Sheffer; ZANDER, Pär-Ola; DA SILVA, Flavio Soares Correa. **Investigating Open Data Portals Automatically: A Methodology and Some Illustrations**. dg.o '18, 2018, New York, NY, USA. *Anais...* New York, NY, USA: ACM, 2018. p. 82:1–82:10. Disponível em: <<http://doi.acm.org/10.1145/3209281.3209292>>. Acesso em: 19 ago. 2018.

DEVEGILI, Leandro: **Operação Serenata de Amor**. Direção de Leandro Devegili. Produção de Leandro Devegili. Florianópolis : Social Good Brasil, 2017. Entretenimento (17:56). Disponível em: <<https://www.youtube.com/watch?v=0bcQcZSMdoQ>>. Acesso em: 10 mai. 2018.

ESRI (Estado Unidos). **ArcGIS**. 2018. Disponível em: <<https://www.esri.com/en-us/about/about-esri/what-we-do>>. Acesso em: 13 out. 2018.

GOVERNO DIGITAL (Brasil). **MINISTÉRIO DO PLANEJAMENTO, DESENVOLVIMENTO E GESTÃO: Portal Brasileiro de Dados Abertos**. 2017. Disponível em: <<https://www.governodigital.gov.br/transformacao/cidadania/dados-abertos/portal-brasileiro-de-dados-abertos>>. Acesso em: 28 out. 2018.

GRUS, Joel. **Data Science do Zero: Primeiras regras com o Python**. Rio de Janeiro: Alta Books, 2016. 336 p. (O'REILLY), Welington Nascimento.

ISO (Genebra). Organização Internacional (Org.). **International Organization for Standardization**. 2014?. Disponível em: <<https://www.iso.org/standards.html>>. Acesso em: 10 jun. 2018.

LAMBRANHO, Lúcio (Florianópolis). **O voluntário que faz uma cidade economizar milhões por ano**. 2016. Disponível em: <<https://www.bbc.com/portuguese/brasil-37526368>>. Acesso em: 28 out. 2018.

MCKINNEY, Wes. **Python para Análise de Dados: Tratamento de dados com Pandas, Numpy e Ipython**. São Paulo: Novatec, 2018. 632 p. (O'REILLY), Lúcia A. Kinoshita.

MORGAN, Derek. Explorando o Common Crawl com Python. **ORGAN.INFO**. Baltimore, 2016. Disponível em: <<https://dmorgan.info/posts/common-crawl-python/>>. Acesso em: 2 abr. 2018.

OPENDATASOFT (Paris). **OpenDataSoft**. 2018. Disponível em: <<https://www.opendatasoft.com/>>. Acesso em: 13 out. 2018.

OPEN KNOWLEDGE INTERNACIONAL (Estados Unidos). **What is Open? 2012?**. Disponível em: <<https://okfn.org/opendata/>>. Acesso em: 05 maio 2018.

TAUBERER, Joshua. **The annotated 8 principles of Open government Data. opengovdata.org**. Sebastopol, California, 2007. Disponível em: <<https://opengovdata.org/>>. Acesso em: 29 mar. 2018.

TYLER TECHNOLOGIES (Estados Unidos). **Socrata: Mission**. 2018. Disponível em: <<https://socrata.com/company-info>>. Acesso em: 13 out. 2018.