Universidade Federal do Pampa
Cristhian Willrich Bilhalva

OGD Search - Uma Ferramenta de Busca para Dados Governamentais

Alegrete
2016

Cristhian Willrich Bilhalva

OGD Search - Uma Ferramenta de Busca para Dados Governamentais

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Fabio Natanael Kepler

Alegrete

2016

Cristhian Willrich Bilhalva

OGD Search - Uma Ferramenta de Busca para Dados Governamentais

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Trabalho de Conclusão de Curso defendido e aprovado em 30 de Novembro de 2016.

Banca examinadora:

Fabio N. Kepler

Prof. Dr. Fabio Natanael Kepler
Orientador

Diego L. Kreutz

Prof. Me. Diego Luís Kreutz
UNIPAMPA

Andréa S. Bordin

Prof.ª Dr.ª Andréa Sabedra Bordin
UNIPAMPA

Dedico este trabalho, aos meus pais, amigos, colegas e
a todos os que me incentivaram e me ajudaram nesta caminhada.

Resumo

Com a criação da lei de acesso à informação em 2011, tornou-se obrigação dos governos a disponibilização de dados governamentais. São os chamados Dados Governamentais Abertos (DGA). Os dados governamentais abertos vêm mostrando sua importância através de várias iniciativas públicas com a criação de aplicativos, sites e ferramentas que usam os DGA como fonte de informação. Neste trabalho, vamos criar uma ferramenta de busca para DGA. Para isso iremos desenvolver um Crawler focado que irá extrair páginas e arquivos apenas de sites do governo. Os dados extraídos pelo Crawler serão indexados no Solr, uma ferramenta OpenSource para criação de motores de busca. Os dados indexados no Solr serão utilizados para a realização de consultas a partir de um Site nomeado OGD Search.

Palavras-chave: Dados; Governo; Análise Estatística; Crawler; Scrapping; Ferramenta de Busca; Motor de Busca; Solr.

Abstract

With the creation of the Brazillian law of information access (Lei de Acesso à informação) in 2011, became responsibility of governments to release government data. This is called Open Government Data (OGD). The OGD have shown its importance through various public initiatives to create applications, websites and tools that use the OGD as a source of information. In this work we plan to create a Search Tool for OGD. To do so, we develop a Focused Crawler that extract pages and files only from governament sites. This data extracted from the crawler is indexed in Solr, a OpenSource tool for creating Search Engines. The data indexed in Solr, will be used to search in a WebSite named OGDSearch.

Key-words: Open Government Data; Statistical Analysis; Crawling;Crawler; Search Engine.

Lista de ilustrações

Lista de siglas

API Application User Interface

CAPTCHA Completely Automated Public Turing test to tell Computers and Humans Apart

CGI Common Gateway Interface

CSV Comma Separated Values

DGA Dados Governamentais Abertos

DOM Document Object Model

e-SIC Sistema Eletrônico do Serviço de Informação ao Cidadão

HTML HyperText Markup Language

JSON JavaScript Object Notation

ODS Open Document Format

OGD Open Government Data

OWL Ontology Web Language

PAC Programa de Aceleração do Crescimento

PDF Portable Document Format

RDF Resource Description Framework

SAX Simple Api for XML

SOAP Simple Object Acess Protocol

URL Uniform Resource Locator

XHTML eXtensible Hypertext Markup Language

XML eXtensible Markup Language

Sumário

1 Introdução

Em 2011, o governo federal Brasileiro criou a lei de acesso à informação (BRASIL. Congresso. Senado, 2011). A lei define que qualquer cidadão pode fazer requisições de dados governamentais sem necessidade de justificativa, em todas as esferas públicas (PRIMEIRA..., 2014). As requisições devem ser atendidas em vinte dias e só podem ser negadas se forem amparadas por alguma lei.

A lei de acesso à informação é basicamente uma regulamentação de um conceito recente, chamado de dados governamentais abertos. Um dado aberto é qualquer informação que pode ser visualizada e compartilhada sem qualquer restrição de licença ou acesso (OPEN..., 2014). No site OpenDefinition.org é possível obter mais informações sobre o conceito de dado aberto ou informação aberta.

Os dados governamentais abertos começaram a aparecer de forma significativa quando o governo dos Estados Unidos motivado pelas iniciativas de transparência de seu presidente Barack Obama, criou o portal da internet Data.gov (ERICKSON et al., 2013). Muitos países, desde então, começaram a abrir seus dados para que pudessem começar a ser usados por cidadãos comuns. O Brasil disponibiliza na internet um portal de dados abertos (PORTAL..., 2014a), assim como diversos outros sites de cidades e estados que disponibilizam dados de seus governos públicos, e portais da transparência de recursos públicos.

Informações e conhecimentos extraídos de bases abertas podem ser usados para a tomada de decisões e o estabelecimento de políticas governamentais. Por exemplo, em (SILVA, 2011), dados de licitações reais extraídas do sistema ComprasNet foram combinados com algoritmos de mineração de dados para encontrar irregularidades como cartéis, simulação de concorrência e direcionamento de editais.

Informações importantes podem ser obtidas por análises estatísticas amplas e confiáveis. Uma ou poucas bases podem ser usadas para obtenção de algum tipo de informação ou conhecimento, mas as informações obtidas são limitadas, por considerarem apenas dados da própria base, e parciais, porque os dados da base são geralmente incompletos. Entretanto, o uso de várias bases para a realização de análises e obtenção de informações úteis é dificultado pelo fato das bases de dados serem segmentadas: os dados estão em diversos sites diferentes, formatos diferentes, padrões diferentes, são repetidos, possuem lacunas, etc. Isso dificulta uma análise integrada, mais extensa e mais profunda do que é possível em apenas uma base de dados. O principal problema, entretanto, é que não há uma base integrada e não há uma forma padronizada ou pré-definida de como integrar diferentes bases.

1.1 Motivação

Vivemos em um governo onde as pessoas devem participar de forma pró-ativa, fiscalizando os serviços públicos, dando ideias e participando das decisões do estado. Para tanto é necessário ter conhecimento sobre os problemas enfrentados e buscar suas soluções com base nos dados.

Os dados governamentais abertos são fontes valiosas de informação, que devem ser exploradas por cidadãos dispostos a participar das políticas do governo. Para isso não basta a existência dos dados, é necessário também que pessoas explorem estes dados com uso de boas ferramentas a fim de encontrar respostas.

Uma das maiores preocupações da sociedade atualmente é quanto aos investimentos públicos. Muito se fala sobre a precariedade dos sistemas de educação, saúde e segurança. Este problema é decorrente do mal uso das verbas públicas bem como falta de organização e de competência das entidades governamentais.

Os dados governamentais abertos podem ser explorados para descobrir, por exemplo, onde vale mais a pena investir. Pode-se também descobrir causas de problemas atuais como epidemias, apenas analisando dados produzidos por secretarias como a da saúde. Diversos problemas relativos ao governo e à população podem ser solucionados com a ajuda dos dados governamentais abertos.

1.2 Objetivos

Nosso objetivo é integrar dados de arquivos e páginas provenientes de diversos sites do governo em um motor de busca exclusivo para DGA. Estes dados poderão então ser utilizados por outras pessoas para realizar os diversos tipos de análises a fim de extrair informações úteis na resolução de problemas governamentais.

No Capítulo 2, iremos explicar de maneira mais profunda a questão dos DGA, conceitos de motores de busca e conceitos sobre o Solr, um software que permite a criação de forma fácil de uma ferramenta de busca. No Capítulo 3, iremos abordar trabalhos já realizados no contexto dos Dados Abertos.

Na seção 4.1 iremos explicar como foi feita a implementação de um Web Crawler utilizado para indexar páginas referentes a Dados Governamentais Abertos. Na seção 4.2 é apresentada uma interface de consulta ao Solr utilizando o servidor Web Apache e Python.

No Capítulo 5 iremos mostrar alguns dos resultados obtidos através de consultas ao Solr e da execução do *crawler*. E por fim, no Capítulo 6, serão apresentadas as concluões sobre os resultados obtidos e possíveis trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo serão abordados conceitos sobre os Dados Governamentais Abertos, *Web Crawling* e sobre a ferramenta de busca Solr. Estes conceitos são necessários para o entendimento de como foi realizado este trabalho.

2.1 Dados Governamentais Abertos

Segundo a definição da OpenDefinition, "dado aberto é um dado que pode ser livremente utilizado, reutilizado e redistribuído por qualquer um" (OPEN..., 2014). Qualquer dado fruto do trabalho administrativo de entidades governamentais que não seja considerado pelo governo como sigiloso, pode ser disponibilizado para que outras pessoas possam usar da forma que quiserem.

Dados governamentais são considerados abertos de acordo com oito principios definidos pela Open Government Data (HACKER, 2011). Estes princípios são:

1. Completos: Todos os dados públicos devem ser disponibilizados. Dado público é aquele que não está sujeito a restrições de privacidade, segurança ou outros privilégios;

2. Primários: São apresentados tal como colhidos da fonte, com o maior nível possível de granularidade, sem agregação ou modificação (por exemplo, um gráfico não é fornecido aberto, mas os dados utilizados para construir a planilha que deu origem a ele devem ser abertos);

3. Atuais: Devem ser publicados o mais rápido possível para preservar seu valor. Em geral, têm periodicidade: quanto mais recentes e atuais, mais úteis para seus usuários;

4. Acessíveis: São disponibilizados para a maior quantidade possível de pessoas, atendendo, assim, aos mais diferentes propósitos;

5. Compreensíveis por máquina: Devem estar estruturados de modo razoável, possibilitando que sejam processados automaticamente (por exemplo, uma tabela em PDF é muito bem compreendida por pessoas, mas para um computador é apenas uma imagem; uma tabela em formato estruturado, como CSV ou XML, é processada mais facilmente por softwares e sistemas);

6. Não discriminatórios: Devem estar disponíveis para qualquer pessoa, sem necessidade de cadastro ou qualquer outro procedimento que impeça o acesso;

7. Não proprietários: Nenhuma entidade ou organização deve ter controle exclusivo sobre os dados disponibilizados;

8. Livres de licenças: Não devem estar submetidos a copyrights, patentes, marcas registradas ou regulações de segredo industrial.

Os dados abertos não necessariamente precisam ser governamentais, empresas e outras entidades também podem disponibilizar dados de forma aberta, para que qualquer um possa usar para seu propósito específico. Porém, neste trabalho nos limitamos a utilizar os dados disponibilizados pelos governos, para que também sejam utilizados como uma forma de avaliação do estado atual dos DGA no Brasil.

2.1.1 Portal de Dados Abertos x Portal da Transparência

Em um Portal de Dados Abertos, qualquer tipo de dados pode ser disponibilizado, sejam eles estatísticos, de orçamento e recursos públicos ou informativos. Neste tipo de site são disponibilizados arquivos de vários formatos. Deseja-se que estes arquivos estejam disponíveis para processamento por computadores, o que nem sempre acontece.

Já em um Portal da Transparência, é possível apenas visualisar informações sobre despesas, receitas e demais operações envolvendo recursos públicos. Apenas alguns destes sites disponibilizam a informação através arquivos que possam ser processados por máquina. Na maioria dos casos a informação está disponível apenas para a visualização em um navegador da internet.

2.1.2 Fontes de dados

No Brasil, existem várias fontes de DGA. No portal de dados abertos do Governo Federal dados.gov.br são disponibilizadas várias séries de dados referentes a diferentes setores públicos. É possível encontrar dados sobre saúde, transporte, segurança, educação, gastos públicos, processos eleitorais, entre outros (SOBRE..., 2014). Por exemplo, em uma análise no portal, nós encontramos uma série de dados referentes às obras do Programa de Aceleração do Crescimento (PAC), indicadores de saúde (taxas de incidência de AIDS e Dengue), educação (listas de instituições de educação, taxa de analfabetismo, etc.), segurança, trabalho, entre outros.

Além do dados.gov.br, o governo federal possui outros websites de onde é possível extrair mais dados:

- Portal da transparência dos recursos públicos federais, onde é possível conseguir dados em tempo real de gastos governamentais (PORTAL..., 2014b);

- Sistema Eletrônico do Serviço de Informação ao Cidadão (e-SIC), onde através de um cadastro é possível fazer solicitação de informações do governo federal e acompanhar o pedido no sistema (SISTEMA..., 2014);
- SICONV, sistema que possibilita consultar informações sobre convênios do governo federal, que conta com uma Application User Interface (API) para consulta dos dados de forma automatizada (DADOS..., 2014a).

Não só o governo federal aderiu aos dados abertos. Estados e municípios de todo o país disponibilizam serviços de obtenção de DGA. Abaixo alguns exemplos de estados que exercem esta prática:

- Rio Grande Do Sul: Possui portal chamado Dados RS, nos mesmos moldes do portal do governo federal dados.gov.br (DADOS..., 2014b). Portal da Transparência (TRANSPARÊNCIA..., 2014). Mapa da transparência, com um infográfico que mostra de forma geral, os gastos do estado do Rio Grande do Sul (MAPA..., 2014). E uma página de solicitação de informações onde é necessário preenchar uma série de campos com informações pessoais (REQUERIMENTO..., 2014).
- São Paulo: Possui o Governo Aberto SP, um portal de dados abertos com dados do estado de São Paulo. Um Portal da transparência que conta inclusive com um WebService em SOAP, tornado possível a consulta aos dados de forma automatizada. E assim como o governo federal e do estado do RS, um Serviço de Informação ao Cidadão, no entanto este não é eletrônico.
- Santa Catarina: Santa Catarina não conta com um portal de dados abertos, porém possui um portal da transparência onde é possível realizar consultas às movimentações financeiras do estado. O acesso por máquinas às vezes é dificultado pelo uso de CAPTCHA em algumas páginas de consulta.

2.1.3 Tipos de dados

Um grande problema enfrentado ao tentar-se agregar valor aos DGA é a diversidade de padrões e formatos. Para diferentes conjuntos serem combinados, é necessário um esforço para padronizar os dados de maneira que no final sejam processados por uma única ferramenta. Em algumas vezes formatos de arquivos proprietários são disponibilizados, necessitando então de uma licença para usar o software capaz de interpretar tais dados.

Um formato simples de dados é o CSV, onde os campos de cada registro são separados por vírgula e os registros são separados por linha. Tal formato é muito fácil de se interpretar por máquina devido a sua simplicidade. O XML e o HyperText Markup Language (HTML) usam marcadores para delimitar os campos e dar nomes e atributos.

Formatos como Open Document Format (ODS) são utilizados em programas de planílha eletrônica, necessitando assim o uso do software para a conversão ou de API's específicas. O PDF também é encontrado em conjuntos de dados abertos. O problema com este formato é que é muito dificil de extrair sua informação automaticamente, pois este formato, assim como imagens por exemplo, foi projetado apenas para a visualização da informação.

Muitos outros formatos existem, no site dados.gov.br encontramos arquivos em formato XML, HTML, JSON, CSV, ODS, RDF, OWL, entre outros.

2.2 Motores de Busca

Um motor de busca é um conjunto de softwares utilizado para pesquisar documentos contendo um termo específico (WHAT..., 2016). Um motor de busca é geralmente composto por um crawler, um *Indexer* e uma interface de consulta - esta ultima responsável por entregar a informação que o usuário do motor de busca precisa.

2.2.1 Web Crawler

Um *Web Crawler* (também conhecido como robô ou spider) é um sistema para download em massa de páginas da internet (OLSTON; NAJORK, 2010). Um *Web Crawler* utiliza os *links* (relações) existentes entre uma página e vai seguindo-os em busca de páginas que possam ter uma informação específica. *Web Crawlers* são utilizados principalmente em motores de Busca, para coletar o máximo de páginas existentes na internet e indexá-las em um banco de dados para posterior consulta de usuários (OLSTON; NAJORK, 2010).

2.2.2 Indexer

Um *Indexer* ou indexador, é responsável por armazenar os termos das páginas, ou as próprias páginas capturadas pelo crawler em um *Index* ou Índice (WHAT..., 2016). Essas páginas podem ser então encontradas através de consultas ao *index*. Os algoritmos utilizados para as consultas também podem fazer parte do software responsável pelo *Index*, auxiliando na organização dos resultados. Por Exemplo, em (BAEZA-YATES; SAINT-JEAN, 2003), é mencionado que o *Index* pode ser organizado de acordo com a distribuição das consultas, de maneira que as consultas mais relevantes sejam mantindas na memória principal e as de menos importância no disco rígido ou memória secundária.

2.2.3 Interface de Consulta

A interface de consulta é a parte do motor de busca responsável por fazer a ligação entre o usuário e o *Index*. Através da interface de consulta o usuário entrará com os termos

de consulta necessários para que ele encontre o que precisa, então a interface retornará com os resultados mais relevantes, o que é relevante ou não é definido através de um algoritmo de relevância.

Do manual referência do Solr (APACHE. . . , 2016b), "relevância é o grau no qual o resultado da consulta satisfaz um usuário que está procurando por informação". A relevância pode ser subjetiva - o que é relevante para uma determinada pessoa pode não ser relevante para outra. Isto insere um certo grau dificuldade na implementação de ferramentas de busca: Como definir o que é relevante?

O principal papel das ferramentas de busca é prover algoritmos que possibilitem determinar para um usuário ou um grupo de usuários o que é mais relevante. De acordo com (AHONEN-MYKA, 2007), os critérios mais comuns de avaliação de relevância são:

- *precision* (precisão): A quantidade de páginas relevantes em relação ao resultado de uma consulta;
- *recall*: A quantidade de páginas relevantes em relação à todas as páginas relevantes possíveis.

Para que estes critérios possam ser aplicados é necessário possuir um conjunto de regras ou julgamentos que definam quando um resultado é relevante ou não. Em (INGERSOLL, 2009) são listados vários métodos para avaliar a relevância de uma consulta:

- *Ad-hoc*: Inventar alguma consulta e verificar se os resultados parecem bons ou ruins;
- Grupos de foco: Escolher um grupo de usuários e deixar que eles interajam com a ferramenta por um algum período de tempo. Despois pede-se um *feedback* sobre a experiência destes usuários com a ferramenta;
- TREC ou outras avaliações públicas: Executar um estudo de relevância utilizando um conjunto de consultas, documentos e julgamentos de relevância criados por um grupo terceirizado. A instância mais conhecida desta técnica é a TREC (*Text Retrieval Evaluation Conference*), mantida pelo Instituto Nacional de Padrões e Tecnologias dos Estados Unidos;
- Classificação Online: Deixar que os usuários avaliem as consultas utilizando um sistema de classificação numérico ou com estrelas por exemplo;

Estes são apenas alguns dos vários métodos existentes de avaliação das consultas, além disso nada impede que novos métodos sejam criados.

2.3 Conceitos da Ferramenta Solr

O Solr é uma plataforma de busca baseada no motor de busca Apache Lucene (APACHE..., 2016a), que permite a criação de um *index* bem como sua consulta, através de uma série de recursos voltados para recuperação de informações. Esta seção tem como objetivo introduzir alguns conceitos particulares desta ferramenta. A Figura 1 mostra uma visão geral dos recursos da ferramenta Solr que serão abordados nas próximas seções.

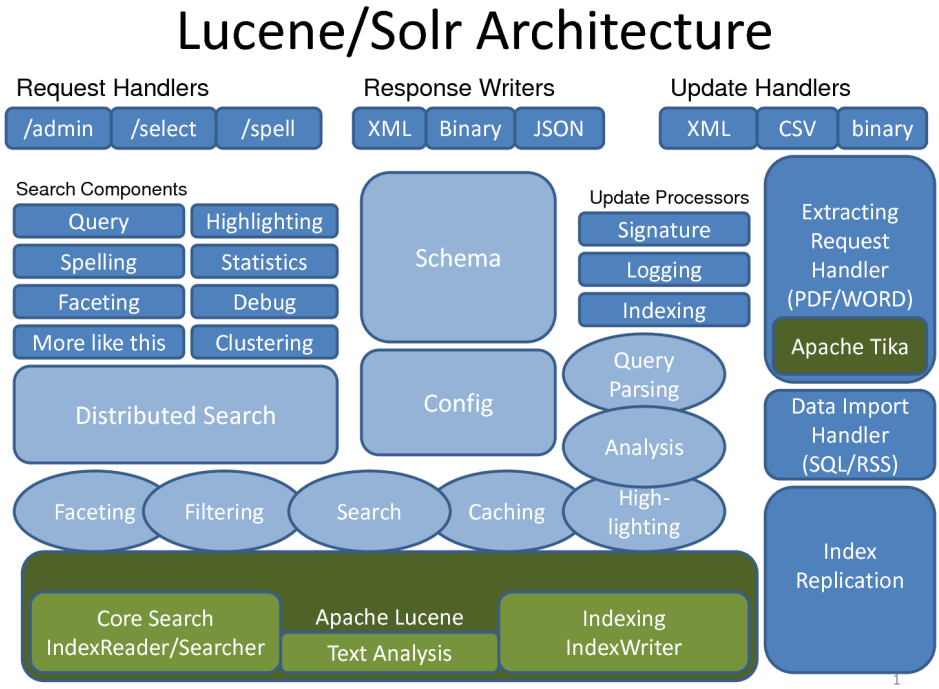


Figura 1 – Arquitetura do Solr.

Fonte: (EIPE, 2016)

2.3.1 Schema

Como em qualquer outro sistema de armazenamento de dados, O Solr necessita de uma modelagem da informação. A Modelagem dos dados no Solr é feita pelo *Schema*. O *Schema* é configurado a partir de um arquivo XML com o nome padrão managed_schema.xml. Neste arquivo são especificados os campos de cada registro, analizadores, tokenizadores, filtros, campos dinamicos, entre outros recursos que facilitam na hora da indexação ou da consulta de registros.

2.3.1.1 Fields

Os *fields* (campos), são propriedades de um documento indexado no Solr. Antes de começar a indexar, é necessário ter um modelo definido com os campos especificando

quais os tipos de informação que eles irão receber. O Solr vem com vários tipos definidos, que servem para representar informações de texto, valores, de data e hora, entre outros tipos.

2.3.1.2 CopyFields e Dynamic Fields

Em alguns momentos pode ser necessário copiar a informação de um ou mais campos para algum campo especifico. Neste caso pode-se utilizar um *CopyField*, que faz justamente esta função.

Já o *DynamicField* permite que o Solr possa indexar o valor de um campo que não foi explicitamente declarado, através do uso de um coringa definido pelo caractere *.

2.3.2 RequestHandlers

Os *Request Handlers* são responsáveis pelas principais funções do Solr, são os RequestHandlers que tem a função de indexar e de realizar as consultas. Neste trabalho utilizamos dois RequestHandlers:

- ExtractingRequestHandler

  O *ExtractingRequestHandler* utiliza o Apache Tika (APACHE..., 2016c) para extrair informação e meta-dados de arquivos dos mais variados formatos e transformar em uma *stream* (fluxo) XHTML que alimenta um objeto SAX. O objeto SAX por sua vez envia eventos ao *RequestHandler* com as informações a serem extraídas do documento XHTML (EXTRACTINGREQUESTHANDLER..., 2016). Através de parâmetros enviados ao *ExtractingRequestHandler* é possível especificar quais informações devem ser mapeadas aos campos definidos no *Schema*.

- SearchHandler

  O *SearchHandler* é responsável por realizar a consulta ao *index* do Solr. O *SearchHandler* ao ser acionado recebe os parametros da consulta através da URL passada ao servidor Solr e retorna ao cliente o resultado da consulta com o formato especificado. O Solr pode gerar a consulta em diversos formatos, incluindo Python, JSON e XML.

3 Trabalhos Relacionados

Na dissertação de mestrado de Carlos Vinícios Sarmento Silva entitulada “Agentes de mineração e sua aplicação no domínio de auditoria governamental”, dados sobre licitações são obtidos através do sistema ComprasNet, e auxiliam na descoberta de fraudes pela Controladoria Geral Da União (SILVA, 2011).

No artigo Open Government Data in Brazil (BREITMAN et al., 2012), são discutidos o estado dos DGA no Brasil e as lições aprendidas com o uso desta prática.

Na dissertação de mestrado da Patrícia Azevedo (AZEVEDO, 2013), ferramentas são utilizadas para integrar diferentes conjuntos de dados sobre enchentes na Bacia do Rio Doce.

Em (GERMANO, 2013), são estudados modelos de negócios de sete ferramentas que utilizam os DGA. O estudo mostra casos de sucesso do uso de aplicativos de cunho social que usam os DGA. O estudo também aponta problemas enfrentados no desenvolvimento dos aplicativos como qualidade dos dados abertos e disponibilidade.

Em (ANGÉLICO, 2012) é criado um catálogo de Dados Governamentais Abertos através do acesso de sites de diversos órgãos e entidades do Executivo Federal entre março e setembro de 2012.

O Manual de Dados Abertos (HACKER, 2011), aborda os conceitos e a importância da abertura de dados governamentais. Neste trabalho é dado foco para a maneira como os dados devem ser disponibilizados por parte do governo e são apontadas iniciativas públicas que devem ser tomadas para a sua popularização.

Em (CRAVEIRO; SANTANA; ALBUQUERQUE, 2013), é feito um estudo sobre os dados orçamentais abertos disponibilizados em sites da internet. Um framework é proposto para a avaliação dos dados obtidos de 54 sites do governo e de sites de 34 sites Brasileiros de auditoria.

Em (NAZÁRIO; SILVA; JOSÉ, 2012), é feita uma avaliação no Portal Da Transparência do Governo Federal de acordo com vários quesitos definidos em um framework. Alguns quesitos são consistência, corretude, conveniência e segurança.

4 OGD Search: Uma ferramenta para DGA

Neste capítulo vamos apresentar o *OGD Search*, uma ferramenta de busca que permitirá procurar páginas e arquivos de diversos formatos de páginas de entidades governamentais brasileiras. Na seção 4.1, é apresentado um *crawler* capaz de indexar páginas e documentos do governo. Na subseção 4.1.4 é explicado como é processo de indexação das páginas e dos arquivos extraídos pelo *crawler*. Na seção 4.2 é apresentada a interface de consulta do *OGD Search*.

4.1 O Crawler

Para a obtenção das páginas e arquivos de Sites do governo, foi desenvolvido um *Web Crawler*. O *Web Crawler* proposto, foi implementado em Python, utilizando a bilioteca urllib2 para o download das páginas e arquivos e a bilioteca lxml para obtenção do DOM das páginas baixadas. A Figura 2 mostra uma visão simplificada do algoritmo de crawling implementado.

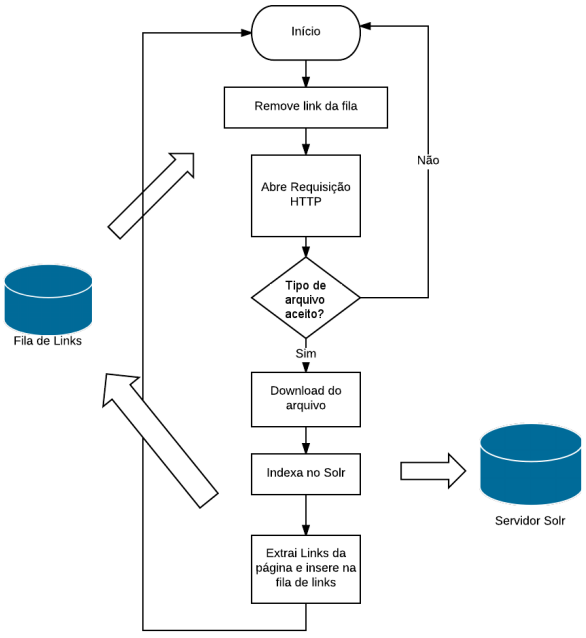


Figura 2 – Visão simplificada do algoritmo do Crawling.

Fonte: Autor.

O processo de *crawling* começa com a adição das páginas sementes na fila de *links* a serem extraídos. Logo após o *crawler* entra em um *loop* que extrai cada página referente a uma URL removida da fila. O DOM desta página contendo a estrutura da mesma é

então extraído, com o intuito de localizar todos os *links* contidos na página. Estes *links* serão então inseridos na fila de extração para que o processo continue.

4.1.1 Mantendo o Escopo de Extração

Nem todos os *links* podem ser inseridos na fila de extração, caso contrário, o *crawler* irá fugir do nosso escopo. Para manter o *crawler* focado, foi desenvolvido um simples algoritmo que elimina as *urls* que não forem do governo. A maneira utilizada para identificar *urls* do governo é simples - basta detectar o sufixo .gov.br.

4.1.2 Árvore de busca e Algoritmo de Busca

A maneira como o *crawling* é realizado, depende diretamente da estrutura utilizada para armazenar os *links* a serem visitados. Se for uma estrutura de dados do tipo fila (o ultimo *link* a entrar na fila é o último a ser retirado), a busca será em largura e se for uma estrutura de dados do tipo pilha (o ultimo *link* colocado na estrutura é o primeiro a ser retirado) a busca será em profundidade.

Buscar em largura significa que as páginas a serem processadas primeiro, serão todas as páginas vizinhas da página que está sendo processada. A Figura 3 mostra este comportamento.

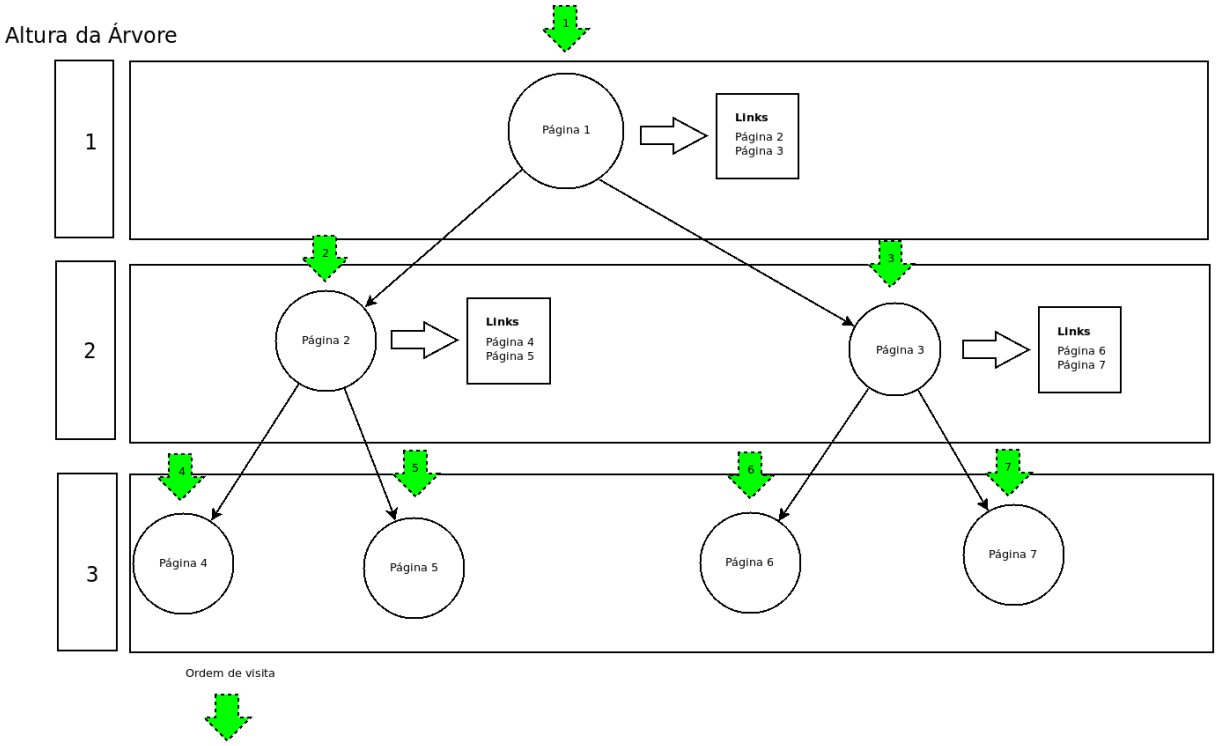


Figura 3 – Exemplo da busca em largura.

Já na busca em profundidade, será processado primeiro uma página desta página vizinha ignorando as demais páginas e deixando-as por último. Assim o *crawler* irá procurar por páginas até o fundo da árvore, depois retornando à visitar as páginas anteriores.

O principal problema neste método é que é necessário que exista algo que limite até que página o *crawling* deve ir, caso contrário as páginas anteriores nunca serão processadas.

A Figura 4 exemplifica o funcionamento da busca em profundidade. Para compreender os exemplos, repare na seta que indica a ordem de visita das páginas.

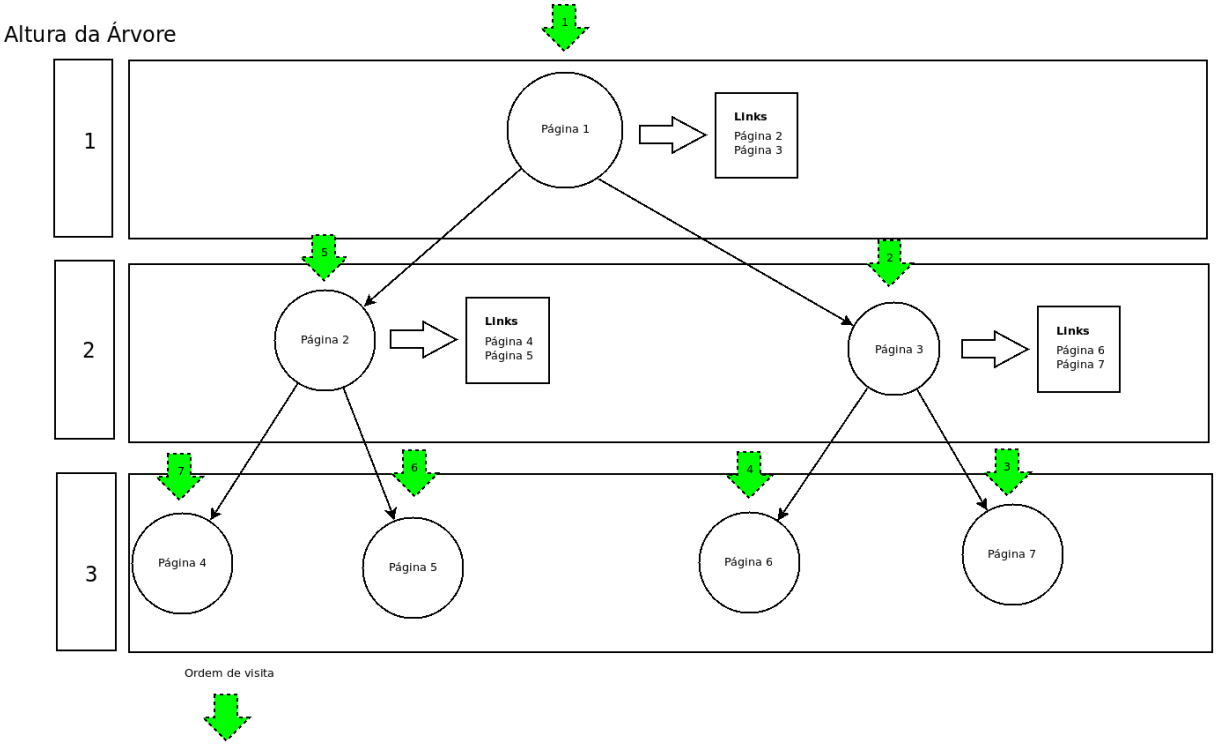


Figura 4 – Exemplo da busca em profundidade.

Em nosso caso adotamos uma estratégia mesclada. Como a idéia é priorizar a indexação de arquivos, quando uma página com um limiar alto de arquivos é encontrada, o *crawler* muda o algoritmo de busca em largura para busca em profundidade. Entretanto, quando a página voltar a ter poucos *links* para arquivos, o *crawler* irá voltar a utilizar a busca em largura.

4.1.3 Facetas de domínio

Todos os *sites* do governo seguem um padrão em seu domínio. Percebe-se que a maioria ou todos os *sites* do governo tem seu nome de domínio terminado em .gov.br. Além disso a estrutura do domínio parece constituir-se de uma organização lógica. Por

exemplo, o *site* da receita federal - organização pertencente ao ministério da fazenda - tem seu domínio escrito como idg.receita.fazenda.gov.br.

Outro exemplo é o *site* do governo do estado do Rio Grande do Sul, o nome de domínio deste *site* é www.estado.rs.gov.br.

Como forma de aproveitar essa organização dos domínios, resolvemos extrair estes níveis de domínio e armazená-los no processo de indexação das páginas, de forma a usá-las como facetas durante o processo de consulta realizado pelo usuário. A figura Figura 5 mostra como o domínio de uma página é processado pelo *crawler*. Do domínio são extraídas palavras chave e determinado se uma página deve ou não ser processada conforme explicado na subseção 4.1.1.

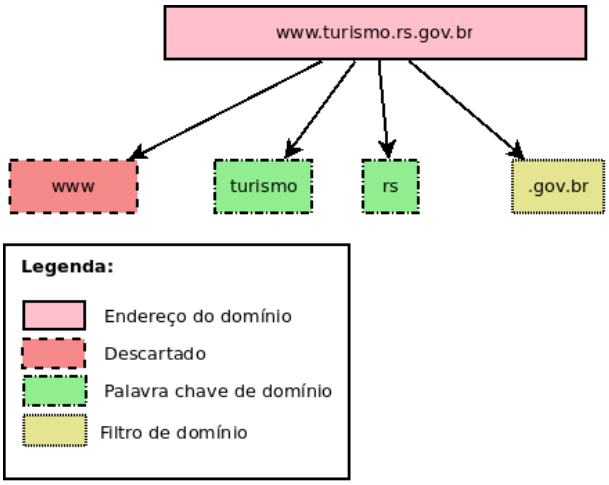


Figura 5 – Separação do domínio em níveis.

A Figura 6 mostra a possibilidade do uso de facetas por domínio na página do *OGD Search*.

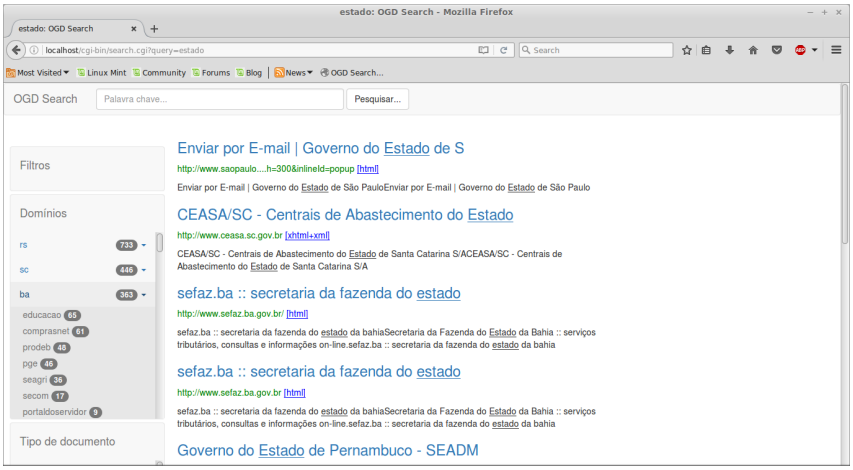


Figura 6 – Visualização das facetas de domínio.

4.1.4 Indexando páginas no Solr

Durante o processo de *crawling*, as páginas baixadas são enviadas ao servidor Solr através de uma requisição HTTP para que as mesmas sejam indexadas. Para indexar as páginas, é utilizado o *handler* do Solr *ExtractingRequestHandler*. Além disso informações adicionais são enviadas como parâmetros da requisição HTTP, como os domínios e sub-domínios extraídos da URL da página conforme explicado na subseção 4.1.3. A Figura 7 mostra como funciona o processo de indexação das páginas.

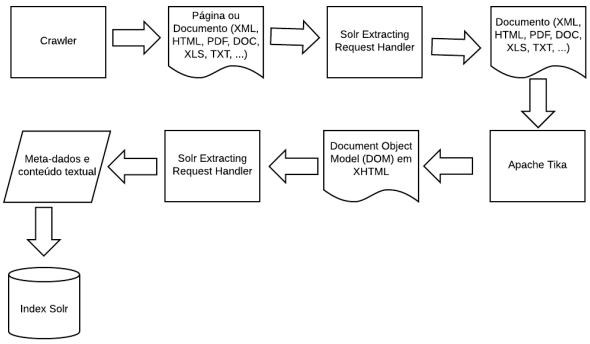


Figura 7 – Processo geral de indexação dos documentos/páginas.

Fonte: Autor.

Como explicado também na subseção 2.3.2, o *ExtractingRequestHandler* é capaz extrair a informação de uma infinidade de formatos de arquivo utilizando o Apache Tika. Para isso, é necessário especificar ao *ExtractingRequestHandler* quais os atributos do DOM XHTML serão extraídos e para qual campo do solr será mapeado. Esta configuração é feita no arquivo de configuração do Solr como mostra a Figura 8.

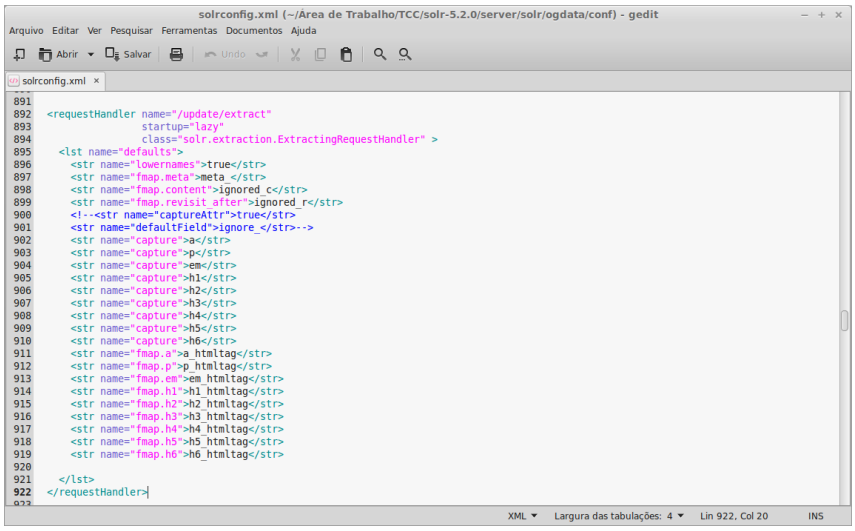


Figura 8 – Configuração do Extracting Request Handler.

Fonte: Autor.

O parâmetro *capture* indica quais os elementos do DOM XHTML devem ter seus valores extraídos, e o parâmetro fmap indica para quais campos do Solr esses valores devem ser indexados. Além disso o *ExtractingRequestHandler* gera automaticamente uma série de campos com meta-dados como data de criação e modificação, título, codificação, autor, entre outras informações contidas nos documentos.

4.2 Interface de Consulta

Para que o usuário possa realizar uma pesquisa através do *OGD Search*, foi desenvolvida uma interface *Web* utilizando um servidor Apache e um CGI em Python. A consulta entrada pelo usuário na página do *OGD Search* é transformada em uma consulta na sintaxe do Solr. A Figura 9 mostra o diagrama simplificado do funcionamento da interface de consulta.

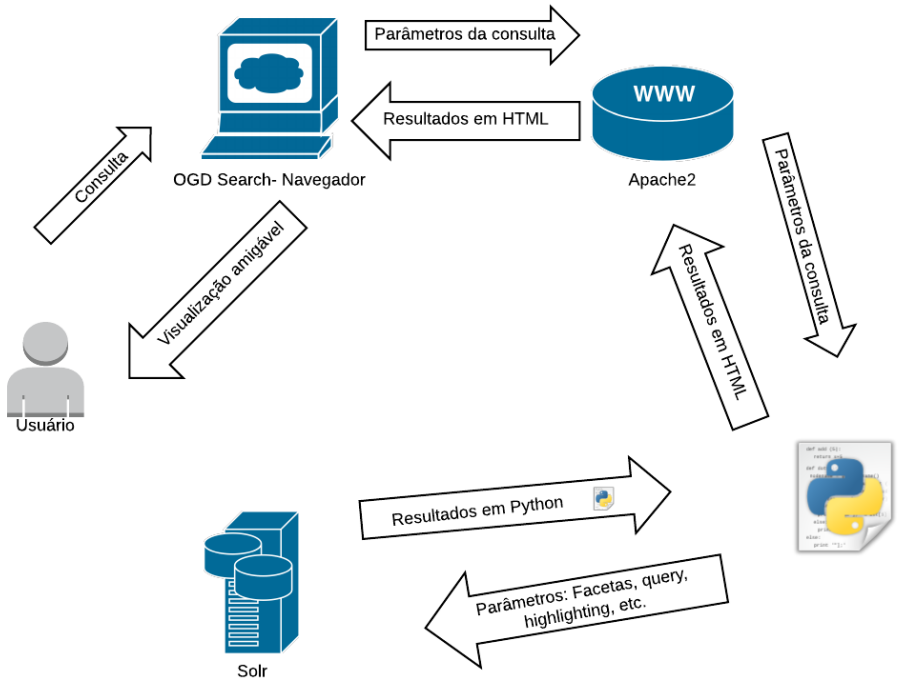


Figura 9 – Diagrama de funcionamento da Interface de Consulta.

Fonte: Autor.

Depois da consulta ser enviada ao *SearchHandler* do servidor Solr, o mesmo retorna os resultados ao CGI em python, que os interpreta e envia ao servidor apache e ao usuário final em forma de uma página *Web*. A Figura 10 mostra a visualização da interface de consulta.

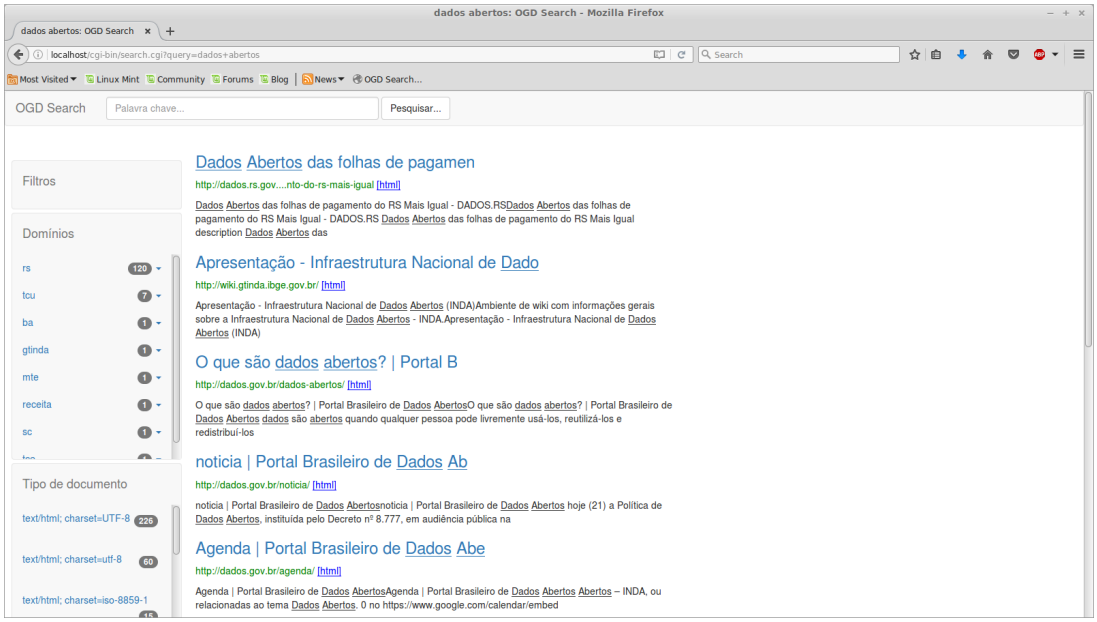


Figura 10 – Página de consulta do OGD Search.

Fonte: Autor.

5 Resultados

Este capítulo tem por objetivo mostrar os resultados obtidos através deste trabalho. Separamos os resultados basicamente em duas esferas:

- *Crawling*: Avaliação do *crawler* desenvolvido em Python. São avaliados os aspectos de eficiência quanto a quantidade de páginas processadas e quanto a relevância das mesmas em relação à página semente.
- Consultas: São avaliadas consultas fictícias quanto à relevância utilizando o método ad-hoc descrito na subseção 2.2.3.

Então, realizamos testes em ambas esferas e após, realizamos uma análise quanto aos resultados obtidos em ambos testes.

5.1 Testes

Para realizar as avaliações, foram executados dois testes. Cada teste consiste em duas etapas, um teste de *crawling* com duração de trinta minutos e três testes com consultas distintas através da interface *web* implementada.

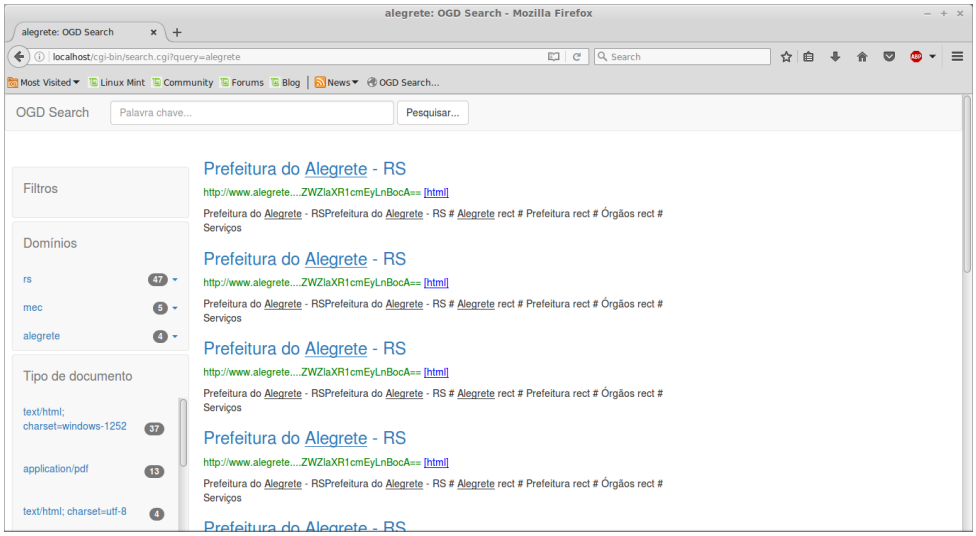
5.1.1 Primeiro Teste

No primeiro teste executamos o crawler durante aproximadamente trinta minutos utilizando como semente a página da Prefeitura Municipal de Alegrete. A Tabela 1 mostra os dados da execução do *crawling* no primeiro teste.

Tabela 1 – Dados de execução do crawling do teste 1.

Semente	http://www.alegrete.rs.gov.br/site/
Páginas baixadas	1549
Páginas restantes na fila	7514
Páginas indexadas no Solr	1397
Início da execução	17:00:28
Final da execução	17:30:02
Tempo total de execução	00:29:34

Durante a executação do *crawler*, foram indexadas 1397 páginas no servidor Solr. Na primeira consulta à página do *OGD Search* foi utilizado o termo de consulta “alegrete”. A página de pesquisa retornou 56 resultados - 13 em formato PDF e 43 em formato HTML conforme mostra a Figura 11.



Prefeitura do Alegrete - RS

http://www.alegrete....ZWZlaXR1cmEyLnBocA== [html]

Prefeitura do Alegrete - RSPrefeitura do Alegrete - RS # Alegrete rect # Prefeitura rect # Órgãos rect # Serviços

Prefeitura do Alegrete - RS

http://www.alegrete....ZWZlaXR1cmEyLnBocA== [html]

Prefeitura do Alegrete - RSPrefeitura do Alegrete - RS # Alegrete rect # Prefeitura rect # Órgãos rect # Serviços

Prefeitura do Alegrete - RS

http://www.alegrete....ZWZlaXR1cmEyLnBocA== [html]

Prefeitura do Alegrete - RSPrefeitura do Alegrete - RS # Alegrete rect # Prefeitura rect # Órgãos rect # Serviços

Prefeitura do Alegrete - RS

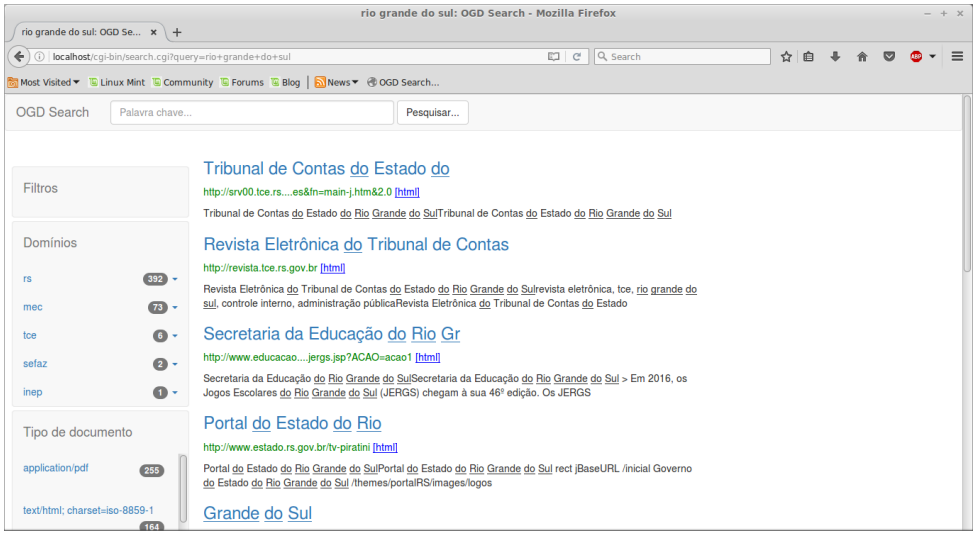
http://www.alegrete....ZWZlaXR1cmEyLnBocA== [html]

Prefeitura do Alegrete - RSPrefeitura do Alegrete - RS # Alegrete rect # Prefeitura rect # Órgãos rect # Serviços

Prefeitura do Alegrete - RS

Figura 11 – Visualização da consulta na página do OGD Search.

Na segunda consulta foram utilizados os termos "rio grande do sul". Na figura Figura 12 é possível visualizar os resultados do *OGD Search*.



Tribunal de Contas do Estado do

http://srv00.tce.rs....es&fn=main-j.htm&2.0 [html]

Tribunal de Contas do Estado do Rio Grande do SulTribunal de Contas do Estado do Rio Grande do Sul

Revista Eletrônica do Tribunal de Contas

http://revista.tce.rs.gov.br [html]

Revista Eletrônica do Tribunal de Contas do Estado do Rio Grande do Sulrevista eletrônica, tce, rio grande do sul, controle interno, administração públicaRevista Eletrônica do Tribunal de Contas do Estado

Secretaria da Educação do Rio Gr

http://www.educacao....jergs.jsp?ACAO=acao1 [html]

Secretaria da Educação do Rio Grande do SulSecretaria da Educação do Rio Grande do Sul > Em 2016, os Jogos Escolares do Rio Grande do Sul (JERGS) chegam à sua 46ª edição. Os JERGS

Portal do Estado do Rio

http://www.estado.rs.gov.br/tv-piratini [html]

Portal do Estado do Rio Grande do SulPortal do Estado do Rio Grande do Sul rect jBaseURL /inicial Governo do Estado do Rio Grande do Sul /themes/portalRS/images/logos

Grande do Sul

Figura 12 – Visualização da consulta na página do OGD Search.

A página de consulta retornou 499 resultados. Os formatos retornados nas consultas estão especificados na Tabela 2.

Tabela 2 – Formatos das páginas retornadas na consulta 2.

Formato da página	Quantidade de páginas
PDF	255
HTML	227
Excel	12
Word	5

Na terceira consulta foram pesquisados os termos “alegrete educação”. Esta consulta retornou apenas 2 resultados, cujos formatos eram ambos PDF conforme mostra a Figura 13.

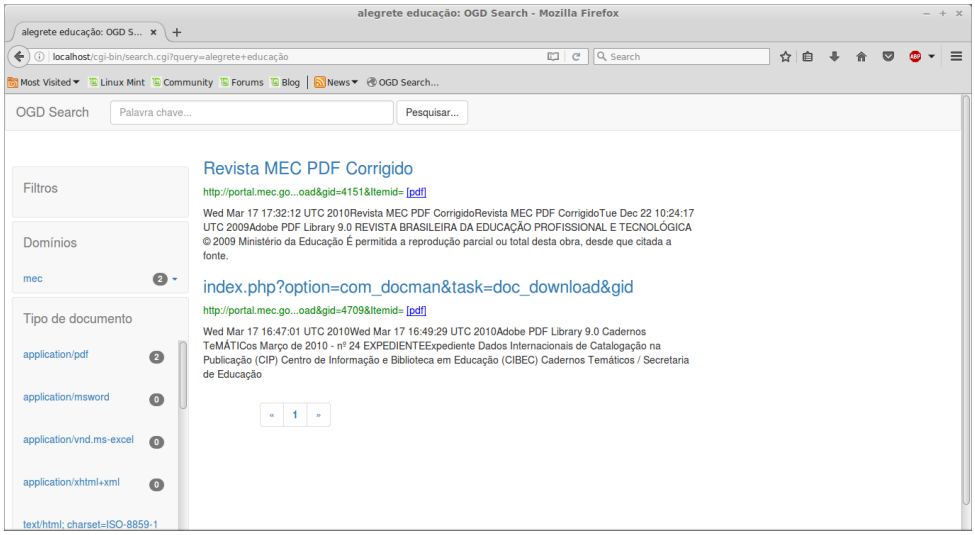


Figura 13 – Visualização da consulta na página do OGD Search.

5.1.2 Segundo Teste

No segundo teste executamos o *crawler* também durante trinta minutos, utilizando desta vez como semente a página do Portal da Transparência do Governo Federal. Os dados da execução do *crawling* são mostrados na Tabela 3.

Tabela 3 – Dados de execução do crawling do segundo teste..

Semente	http://www.portaldatransparencia.gov.br/
Páginas baixadas	1298
Páginas restantes na fila	7625
Páginas indexadas no Solr	1038
Início da execução	19:00:49
Final da execução	19:31:30
Tempo total de execução	00:30:41

Depois de executado o *crawler* foram indexadas 1038 páginas no servidor Solr. Na primeira consulta utilizamos o termo “transparência” na página de consulta do *OGD Search*. A consulta retornou 451 resultados, onde 337 eram páginas do formato HTML, 113 do formato PDF e apenas um documento do formato Microsoft Word. A Figura 14 mostra a visualização dos resultados desta primeira consulta.

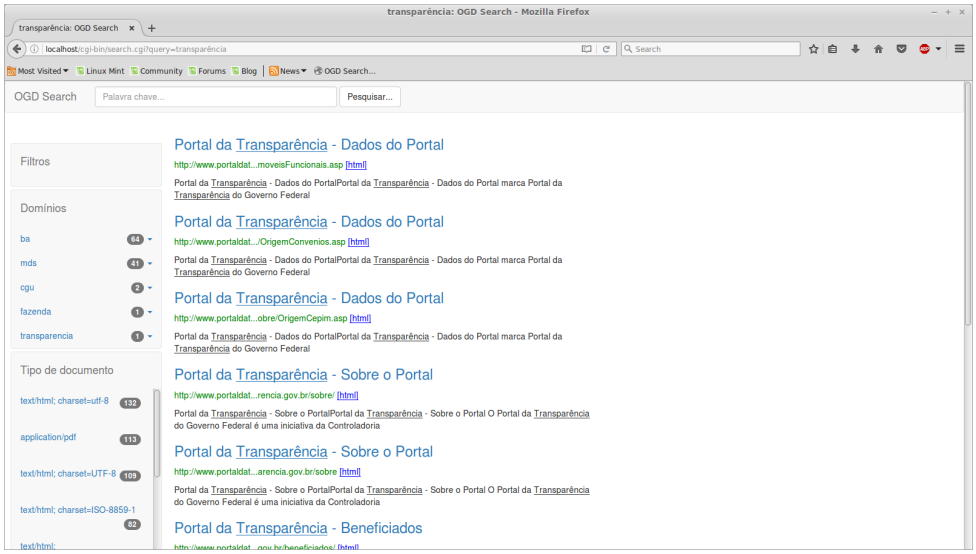


Figura 14 – Visualização da consulta na página do OGD Search.

Na segunda consulta foram utilizados os termos "recursos saúde". A consulta retornou 41 resultados dos quais 12 eram de formato HTML e 29 de formato PDF. A figura Figura 15 mostra a visualização da consulta.

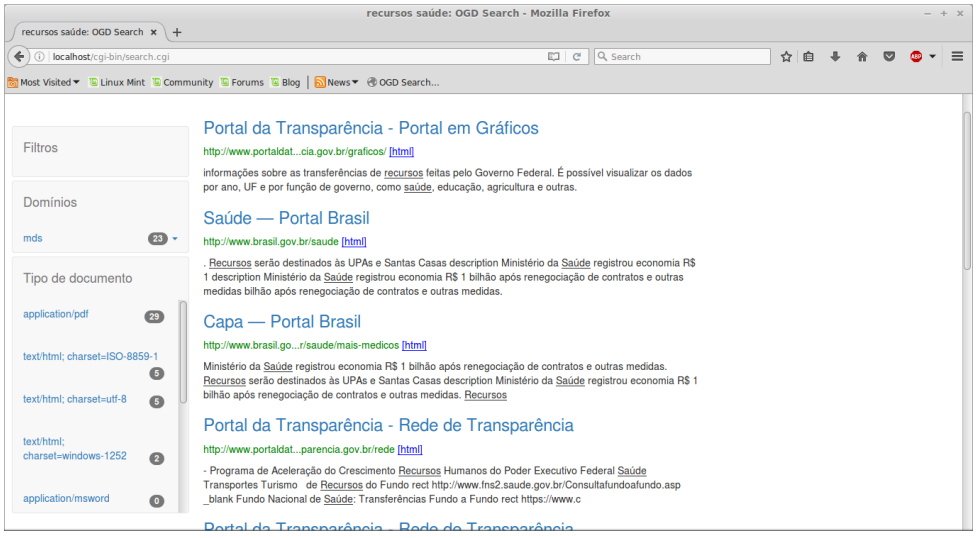


Figura 15 – Visualização da consulta na página do OGD Search.

Na terceira consulta, foram utilizados os termos "PEC 241" na página de busca do *OGD Search*. Esta última consulta retornou apenas 3 resultados, todos de formato HTML. A Figura 16 mostra a visualização desta última consulta.

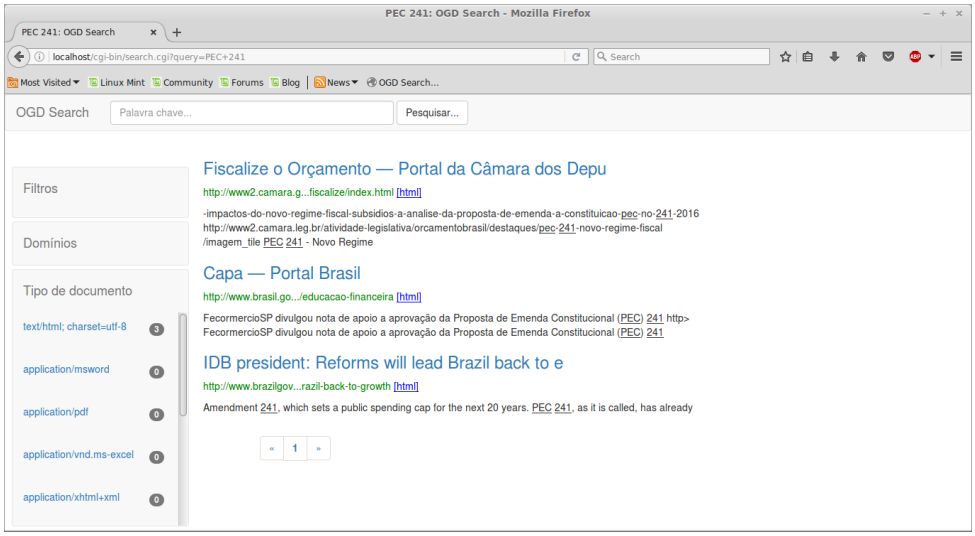


Figura 16 – Visualização da consulta na página do OGD Search.

5.2 Considerações Sobre os Resultados do Crawler

Os resultados de ambos os testes mostraram alguns problemas relacionados ao *crawling* das páginas. Percebe-se em ambos os testes, que o *crawler* foge do contexto da página semente de forma rápida. Natural já que não foi implementado um controle quanto a permanência do *crawler* no domínio em que ele está.

Uma das formas de resolver este problema é ordenar as páginas na fila de páginas a ser visitadas, de forma que as páginas do domínio atual tenham a preferência. Uma das vantagens da dispersão quanto ao domínio da semente, é a de que o *crawler* acaba procurando informações que podem ser interessantes em outros domínios.

Outro provável motivo para a dispersão do *crawler* em relação a página da semente, é o algoritmo utilizado para a fila de páginas. Como mencionado na subseção 4.1.2 o algoritmo trabalha em profundidade quando encontra *links* para arquivos. Logo se o *crawler* encontrar uma página de um domínio que não seja o da página semente com vários arquivos, ele continuará a extrair este domínio até que não restem mais páginas com arquivos.

Por último, o tempo de execução do *crawler* parece curto, a idéia de se usar um intervalo de tempo curto é justamente o de testar a eficiência do algoritmo de busca. Um *crawler* ideal, consegue coletar mais páginas relevantes em menos tempo.

5.3 Considerações sobre as Consultas

A primeira consideração a se fazer sobre as consultas, é que as mesmas dependem diretamente da eficiência do *crawler*. Como o tempo de execução do *crawler* foi curto,

naturalmente o número de resultados relevantes será menor. A primeira consulta de ambos os testes mostra o que foi explicado na seção 5.2, como termo de consulta foi o nome relacionado diretamente ao nome das páginas sementes selecionadas, o ideal seria que fossem obtidos mais resultados na consulta, isto representaria um *recall* maior. No entanto a precisão demonstrada na primeira consulta de ambos os testes foi satisfatória, já que retornaram as páginas mais relevantes em relação aos resultados das consultas.

6 Conclusão

Neste trabalho aprendemos vários conceitos referentes a extração de dados, *Web Crawling*, indexação de dados e sobre a ferramenta Solr.

A extração de dados com o uso do crawler, necessitou de um algoritmo para direcionar de maneira mais eficiente a visita às páginas, poupando tempo, processamento e distribuindo melhor a extração das páginas. Este recurso fica pendente para um próximo trabalho.

O Solr auxiliou muito em todo o processo. Com ele a indexação dos arquivos se torna muito fácil, pois ele consegue tratar uma grande variedade de formatos. As consultas também são muito simples de ser implementadas, bastando apenas estudar a sintaxe de consulta do Solr.

A visualização dos resultados da busca ficou bem limpa e de fácil compreensão, porém faltaram recursos mais avançados. Existem idéias de criar uma visualização em grafos da relação entre as páginas extraídas pelo *Web Crawler*. Esta idéia chegou a começar a ser implementada, mas por falta de tempo ficou adiada para um trabalho futuro.

As consultas parecem ter uma relevância boa, o próximo passo é realizar avaliações mais profundas utilizando o feedback de usuários reais. Para isso será necessário hospedar o *OGD Search* em um servidor e contratar um domínio.

Mesmo com vários problemas de implementação, este trabalho pode servir como base para melhorias tornando assim o OGD Search uma ótima ferramenta para auxiliar os cidadãos na busca por informações úteis em bases de dados de governos.

Referências

AHONEN-MYKA, H. *Portal da Transparência - Glossário*. 2007. Citado na página 21.

ANGÉLICO, F. *Catálogo de dados*. [S.l.]: Projeto CGU, MPOG e UNESCO. Relatório Final, 2012. Citado na página 25.

APACHE Solr. 2016. Disponível em: <http://lucene.apache.org/solr/>. Acesso em: 9 de Junho de 2016. Citado na página 22.

APACHE Solr Reference Guide. 2016. Disponível em: <https://archive.apache.org/dist/lucene/solr/ref-guide/apache-solr-ref-guide-5.2.pdf>. Acesso em: 9 de Junho de 2016. Citado na página 21.

APACHE Tika - Apache Tika. 2016. Disponível em: <http://tika.apache.org/>. Acesso em: 10 de Junho de 2016. Citado na página 23.

AZEVEDO, P. C. N. D. Uma proposta para visualização de linked data sobre enchentes na bacia do rio doce. *Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento*, v. 2, n. 1, 2013. Citado na página 25.

BAEZA-YATES, R.; SAINT-JEAN, F. A three level search engine index based in query log distribution. In: SPRINGER. *String Processing and Information Retrieval*. [S.l.], 2003. p. 56–65. Citado na página 20.

BRASIL. Congresso. Senado. Lei nº 12.527, de 18 de novembro de 2011. 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, DF, nov. 2011. Citado na página 15.

BREITMAN, K. et al. Open government data in brazil. *IEEE Intelligent Systems*, Institute of Electrical and Electronics Engineers, Inc., USA United States, v. 27, n. 3, p. 45–49, 2012. Citado na página 25.

CRAVEIRO, G. S.; SANTANA, M. T.; ALBUQUERQUE, J. P. Assessing open government budgetary data in brazil. In: *ICDS 2013, The Seventh International Conference on Digital Society*. [S.l.: s.n.], 2013. p. 20–27. Citado na página 25.

DADOS Abertos SICONV. 2014. Disponível em: <http://api.convenios.gov.br/siconv/doc/>. Acesso em: 6 de Agosto de 2014. Citado na página 19.

DADOS RS. 2014. Disponível em: <http://dados.rs.gov.br/>. Acesso em: 6 de Agosto de 2014. Citado na página 19.

EIPE, J. *CS Repository: Notes on Apache SOLR*. 2016. Citado na página 22.

ERICKSON, J. S. et al. Open government data: A data analytics approach. 2013. Citado na página 15.

EXTRACTINGREQUESTHANDLER - Solr Wiki. 2016. Disponível em: <http://wiki.apache.org/solr/ExtractingRequestHandler>. Acesso em: 10 de Junho de 2016. Citado na página 23.

GERMANO, E. C. *Modelos de negócios adotados para o uso de dados governamentais abertos: estudo exploratório de prestadores de serviços na cadeia de valor dos dados governamentais abertos*. Tese (Doutorado) — Universidade de São Paulo, 2013. Citado na página 25.

INGERSOLL, G. *Debugging Search Application Relevance Issues - Lucidworks.com*. 2009. Disponível em: <https://lucidworks.com/blog/2009/09/02/debugging-search-application-relevance-issues/>. Acesso em: 28 de Junho de 2016. Citado na página 21.

MAPA da Transparência. 2014. Disponível em: <http://www.mapa.rs.gov.br/>. Acesso em: 6 de Agosto de 2014. Citado na página 19.

NAZÁRIO, D. C.; SILVA, P. F. da; JOSÉ, A. Avaliação da qualidade da informação disponibilizada no portal da transparência do governo federal. 2012. Citado na página 25.

OLSTON, C.; NAJORK, M. Web crawling. *Foundations and Trends in Information Retrieval*, Now Publishers Inc., v. 4, n. 3, p. 175–246, 2010. Citado na página 20.

OPEN Knowledge Definition. 2014. Disponível em: <http://opendefinition.org/>. Acesso em: 6 de Agosto de 2014. Citado 2 vezes nas páginas 15 e 17.

PORTAL Brasileiro de Dados Abertos. 2014. Disponível em: <http://dados.gov.br/>. Acesso em: 6 de Agosto de 2014. Citado na página 15.

PORTAL da Transparência nos Recursos Públicos Federais. 2014. Disponível em: <http://www.portaltransparencia.gov.br/>. Acesso em: 6 de Agosto de 2014. Citado na página 18.

PRIMEIRA Lei de Acesso no mundo que prevê dados abertos. 2014. Disponível em: <http://dados.gov.br/noticia/primeira-lei-de-acesso-no-mundo-que-preve-dados-abertos/>. Acesso em: 6 de Agosto de 2014. Citado na página 15.

REQUERIMENTO de Informações - Central do Cidadão. 2014. Disponível em: <http://www.centraldocidadao.rs.gov.br/informacoes>. Acesso em: 6 de Agosto de 2014. Citado na página 19.

SILVA, C. V. S. Agentes de mineração e sua aplicação no domínio de auditoria governamental. 2011. Citado 2 vezes nas páginas 15 e 25.

SISTEMA de Acesso à Informação. 2014. Disponível em: <http://www.acessoainformacao.gov.br/sistema/site/index.html>. Acesso em: 6 de Agosto de 2014. Citado na página 19.

SOBRE o dados.gov.br | Portal Brasileiro de Dados Abertos. 2014. Disponível em: <http://dados.gov.br/sobre/>. Acesso em: 6 de Agosto de 2014. Citado na página 18.

TRANSPARÊNCIA RS. 2014. Disponível em: <http://www.transparencia.rs.gov.br/>. Acesso em: 6 de Agosto de 2014. Citado na página 19.

HACKER, C. T. (Ed.). *Manual dos dados abertos: governo*. [S.l.], 2011. Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Citado 2 vezes nas páginas 17 e 25.

WHAT is a Search Engine - Webopedia Definition. 2016. Disponível em: <http://www.webopedia.com/TERM/S/search_engine.html>. Acesso em: 13 de Junho de 2016. Citado na página 20.