ANDRÉS MANTECON RIBEIRO MARTANO

Diário Livre: co-criação de uma ferramenta
para publicação de um diário oficial em
formato aberto

São Paulo

2015

ANDRÉS MANTECON RIBEIRO MARTANO

**Diário Livre: co-criação de uma ferramenta para
publicação de um diário oficial em formato aberto**

Dissertação apresentada à Escola de Artes,
Ciências e Humanidades da Universidade de
São Paulo para obtenção do título de Mestre
em Ciências pelo Programa de Pós-graduação
em Sistemas de Informação.

Área de Concentração: Sistemas de In-
formação

Versão corrigida contendo as alterações
solicitadas pela comissão julgadora em 5
de novembro de 2015. A versão original
encontra-se em acervo reservado na Biblio-
teca da EACH-USP e na Biblioteca Digital
de Teses e Dissertações da USP (BDTD), de
acordo com a Resolução CoPGr 6018, de 13
de outubro de 2011.

Orientador: Profa. Dra. Gisele da Silva Cra-
veiro

São Paulo

2015

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Martano, Andrés Mantecon Ribeiro

Diário Livre : co-criação de uma ferramenta para publicação de um diário oficial municipal em formato aberto / Andrés Mantecon Ribeiro Martano ; orientadora, Gisele da Silva Craveiro. – São Paulo, 2015

141 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo
Versão corrigida

1. Governo eletrônico. 2. Sistemas de informação - Implantação. 3. Arquivos abertos. 4. Software livre. I. Craveiro, Gisele da Silva, orient. II. Título

CDD 22.ed.- 350.0285

Dissertação de autoria de Andrés Mantecon Ribeiro Martano, sob o título **Diário Livre: co-criação de uma ferramenta para publicação de um diário oficial em formato aberto**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Sistemas de Informação, aprovada em 5 de novembro de 2015 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Gisele da Silva Craveiro

Presidente

Escola de Artes Ciências e Humanidades - Universidade de São Paulo

Prof. Dr. Nicolau Reinhard

Faculdade de Economia Administração e Contabilidade - Universidade de São Paulo

Prof. Dr. João Porto de Albuquerque

Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo

À Larissa, Marcia e Norberto, pelo inesgotável apoio, compreensão e amor.

Agradecimentos

Tendo em vista que tudo se relaciona de alguma forma, e que o hoje só é hoje por conta do ontem, agradeço à tudo, à todas e todos.

Além disso, não teria como não destacar meus agradecimentos especiais à minha família (o que inclui, é claro, minha noiva e sua família). À minha orientadora e demais membras e membros do Colab. Ao Rafael, Perez, e às membras e membros da COPI, SEMPLA e LabPRODAM, que tão essenciais foram para a realização deste trabalho. E à toda comunidade desenvolvedora de software livre, fornecedora das ferramentas para que ele fosse cumprido.

“O mundo não é. O mundo está sendo. Como subjetividade curiosa, inteligente, interferidora na objetividade com que dialeticamente me relaciono, meu papel no mundo não é só o de quem constata o que ocorre mas também o de quem intervém como sujeito de ocorrências. Não sou apenas objeto da História mas seu sujeito igualmente. No mundo da História, da cultura, da política, constato não para me adaptar mas para mudar. No próprio mundo físico minha constatação não me leva à impotência. O conhecimento sobre os terremotos desenvolveu toda uma engenharia que nos ajuda a sobreviver a eles. Não podemos eliminá-los mas podemos diminuir os danos que nos causam. Constatando, nos tornamos capazes de intervir na realidade, tarefa incomparavelmente mais complexa e geradora de novos saberes do que simplesmente a de nos adaptar a ela. É por isso também que não me parece possível nem aceitável a posição ingênua ou, pior, astutamente neutra de quem estuda, seja o físico, o biólogo, o sociólogo, o matemático, ou o pensador da educação. Ninguém pode estar no mundo, com o mundo e com os outros de forma neutra. Não posso estar no mundo de luvas nas mãos constatando apenas. A acomodação em mim é apenas caminho para a inserção, que implica decisão, escolha, intervenção na realidade. Há perguntas a serem feitas insistentemente por todos nós e que nos fazem ver a impossibilidade de estudar por estudar. De estudar descomprometidamente como se misteriosamente de repente nada tivéssemos que ver com o mundo, um lá fora e distante mundo, alheado de nós e nós dele.”

Freire (2007)

Resumo

MARTANO, Andrés Mantecon Ribeiro. **Diário Livre: co-criação de uma ferramenta para publicação de um diário oficial em formato aberto**. 2015. 141 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015.

Muitos são os potenciais benefícios e obstáculos envolvendo a implantação de iniciativas de dados governamentais abertos (DGA). Dentro do universo de dados passíveis de serem abertos, os Diários Oficiais (DOs) constituem um exemplo ainda pouco estudado. Tendo em vista que os atos públicos se tornam válidos através da publicação em DOs, fica evidente a importância das informações contidas nesses documentos. Essas informações podem vir então a ser melhor reutilizadas uma vez que sejam publicadas seguindo os preceitos de DGA. Este trabalho teve como objetivos desenvolver uma ferramenta para a publicação de um DO em formato aberto, incluindo diversos grupos na definição de seus requisitos e documentando todo o processo. Utilizando uma metodologia inspirada na Design Science Research e na Pesquisa-Ação, foi firmada uma parceria com um órgão público, atuando no processo de abertura dos dados. O processo buscou permitir algum nível de participação através de eventos e questionários. O levantamento de requisitos considerou as necessidades dos gestores envolvidos, a literatura relacionada e as colocações feitas por participantes dos eventos e questionários, buscando assim uma aferição mais completa das necessidades relacionadas. O acompanhar do processo também permitiu algumas considerações sobre os interesses e desafios envolvidos no mesmo. A ferramenta desenvolvida tem recebido um número crescente de acessos e boas avaliações por parte de usuários tanto internos como externos à administração pública. Algumas tentativas de reuso automático dos dados publicados já vêm ocorrendo, mostrando os desafios e potencialidades dos mesmos.

Palavras-chaves: Dados Abertos Governamentais. Transparência. Processos de abertura de dados. Software livre. Pesquisa-Ação. Design Science Research. Diário Oficial. Implantação de sistemas de informação.

Abstract

MARTANO, Andrés Mantecon Ribeiro. **Diário Livre: the co-creation of a tool for a municipal official gazette publication as open data**. 2015. 141 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2015.

Many are the potential benefits and obstacles related to open government data (OGD) initiatives. Among all the data that could be opened, the Official Gazettes (OGs) are one example still demanding better studies. Noting that the public acts only become valid through publication in OGs, it is evident the importance of the information within these documents. This information can then be better reused if published following OGD principles. This work, adopting a methodology inspired by action research, and in partnership with a public body, acted in an OGD initiative and in the development of a tool complying with the requirements of its publication. The process sought to allow some level of participation through public events and questionnaires. The requirement analysis took into consideration the demands of the involved civil servants, related literature and the opinions of the participants of the events and questionnaires, seeking a more complete view about the related demands. Throughout the entire process the challenges and interests were explored, pursuing a better understanding of matter. The tool has received an increasing number of accesses and good reviews by users from the public administration and in general. Some attempts were made to automatically reuse the published data, demonstrating the potentialities and challenges of them.

Keywords: Open Government Data. Transparency. Free Software. Action Research. Design Science Research. Official Gazette.

Lista de figuras

Figura 1 – Organograma da CGM	50
Figura 2 – Transformações pelas quais os dados terão de passar em cada abertura.	54
Figura 3 – Fluxo de publicação do DOM	60
Figura 4 – Arquitetura do Diário Livre em camadas e o fluxo da informação.	65
Figura 5 – Fluxo de publicação do DOM e Diário Livre	77
Figura 6 – Um exemplo de tela do atual DL	78
Figura 7 – Número de visitantes únicos por dia. A oscilação de maior frequência é devido à redução no número de visitas nos fins de semana.	81
Figura 8 – Média de visitantes únicos por dia da semana.	81
Figura 9 – Número total de visitantes e pesquisas por semana. Contam-se aí as pesquisas feitas pela interface interna do site.	82
Figura 10 – Comparação por dispositivo da quantidade total de visitantes e número de visitantes que saíram após visualizar apenas uma página.	82
Figura 11 – Quantidade total de visitantes por tipo de referência de entrada.	83
Figura 12 – Média de ações por visita por tipo de referência de entrada.	83
Figura 13 – Número de questionários online respondidos ao longo do tempo. As cores diferenciam entre os questionários que foram preenchidos até o fim, e aqueles que, por algum motivo, não foram completados. Para as análises foram utilizados apenas os questionários completados.	85
Figura 14 – Respostas para a pergunta: “ Você teve dificuldade para utilizar o site do Diário Oficial Experimental? ”	86
Figura 15 – Respostas para a pergunta: “ Você já teve dificuldade para utilizar o site do Diário Oficial da Imprensa Oficial? ”	86
Figura 16 – Respostas para a pergunta: “ Você ficou satisfeito com o sistema de busca por artigos do Diário Oficial Experimental? ”	87
Figura 17 – Respostas para a pergunta: “ Que dispositivos você utilizou para acessar o site? ”	87
Figura 18 – Respostas para a pergunta: “ Qual seu interesse no Diário Oficial? ”	88
Figura 19 – Respostas para a pergunta: “ Você realizou o download integral da base de dados dos Diários Oficiais fornecidos? ”	88

Figura 20 – Respostas para a pergunta: “ Com que frequência você consulta o Diário Oficial? ”	89
Figura 21 – Respostas para a pergunta: “ Como você ficou sabendo do site experimental do diário oficial? ”	89
Figura 22 – Respostas para a pergunta: “ Você sentiu falta de alguma funcionalidade no site do Diário Oficial Experimental? Ou haveria alguma funcionalidade que você gostaria de sugerir? ”	90
Figura 23 – Respostas para a pergunta: “ Escolaridade ”	100
Figura 24 – Grupo a qual pertencia cada participante do evento. Os grupos foram inferidos a partir das respostas sobre profissão e cargo fornecidas por cada participante.	100
Figura 25 – Respostas para a pergunta: “ Você acha que esse projeto trará benefícios para a sociedade como um todo? ”	101
Figura 26 – Respostas para a pergunta: “ Quem você acha que se beneficia com um projeto desses? ”	102
Figura 27 – Respostas para a pergunta: “ Quem você acha que deveria ser convidado para um evento desse tipo? ”	102
Figura 28 – Respostas para a pergunta: “ Você acha que a abertura de dados governamentais é importante? ”	103
Figura 29 – Respostas para a pergunta: “ Você julga que a abertura dos dados públicos é dispendiosa para o estado? ”	103
Figura 30 – Respostas para a pergunta: “ Você acha que os benefícios da abertura de dados públicos compensam o esforço dispendido? ” . . .	104
Figura 31 – Respostas para a pergunta: “ Quais grupos da sociedade você acha que atualmente se interessam por esses dados? ”	104
Figura 32 – Respostas para a pergunta: “ Você acha que a abertura dos dados aumenta a confiança da população nos governos? ”	105
Figura 33 – Respostas para a pergunta: “ Você acha que a mídia usa dados governamentais para denunciar o governo de forma irresponsável? ”	105
Figura 34 – Respostas para a pergunta: “ Você acha que gestores públicos podem ter medo da abertura de certos dados governamentais? ” .	106

Figura 35 – Respostas para a pergunta: “ Você acha que a legislação atual sobre dados abertos é adequada? ”	107
Figura 36 – Respostas para a pergunta: “ Você acha que o prefeito de São Paulo apoia iniciativas de abertura de dados? ”	107
Figura 37 – Respostas para a pergunta: “ Você acha que a presença de gestores em um evento público sobre abertura de dados governamentais pode colocá-los em situação vulnerável? ”	108
Figura 38 – Respostas para a pergunta: “ Você acha que se posicionar contra ou a favor da abertura de dados pode colocar um(a) gestor(a) em situação vulnerável? ”	108
Figura 39 – Respostas para a pergunta: “ Esse evento atendeu às suas expectativas? ”	109

Lista de tabelas

Tabela 1 – Comparações entre as recomendações para processos de abertura de dados governamentais feitas por três organizações que advogam DGA. Os itens marcados com “✓” indicam pontos defendidos por aquele documento. Quando necessário, está explicitada a forma como o documento trata o item. Quando o documento não aborda o item em questão, o símbolo “—” é usado	41
--	----

Lista de abreviaturas e siglas

API	<i>Application programming interface</i>
CBN	Central Brasileira de Notícias
CGM	Controladoria Geral do Município
CKAN	<i>Comprehensive Knowledge Archive Network</i>
COPI	Coordenadoria de Promoção da Integridade
CSV	<i>Comma-Separated Values</i>
DGA	Dados Governamentais Abertos
DL	Diário Livre
DO	Diário Oficial
DOM	Diário Oficial Municipal de São Paulo
DOU	Diário Oficial da União
DSR	Design Science Research
EACH	Escola de Artes Ciências e Humanidades
FTP	<i>File Transfer Protocol</i>
HTTP	<i>Hypertext Transfer Protocol</i>
HTTPS	<i>Hypertext Transfer Protocol Secure</i>
JSON	<i>JavaScript Object Notation</i>
LAI	Lei de Acesso à Informação
MODS	<i>Metadata Object Description Schema</i>
ONG	Organização Não Governamental
OSC	Organização da Sociedade Civil

PA	Pesquisa-Ação
PDF	<i>Portable Document Format</i>
RSS	<i>Rich Site Summary</i>
SEMPLA	Secretaria Municipal de Planejamento, Orçamento e Gestão
SMDU	Secretaria Municipal de Desenvolvimento Urbano
SQL	<i>Structured Query Language</i>
TIC	Tecnologia de Informação e Comunicação
TXT	Arquivo de texto sem formatação
URI	Uniform Resource Identifier
USP	Universidade de São Paulo
UTF-8	<i>UCS Transformation Format-8-bit</i>
WebDAV	<i>Web Distributed Authoring and Versioning</i>
XML	<i>Extensible Markup Language</i>

Sumário

1	Introdução	25
1.1	Justificativa	28
1.2	Objetivos	28
1.3	Estrutura Geral do Trabalho	29
2	Revisão Bibliográfica	31
2.1	Diários Oficiais	31
2.2	Dados Governamentais Abertos	33
2.2.1	Desafios da Abertura de Dados Governamentais	35
2.2.2	Requisitos para a Abertura de Dados	40
3	Metodologia	43
3.1	Pesquisa-Ação e Design Science Research	43
3.2	Metodologia Aplicada	46
4	Desenvolvimento	49
4.1	Contexto	49
4.1.1	Escolha da Base de Dados	51
4.1.2	Parceria	55
4.1.3	Arcabouço Legal	55
4.2	Diário Oficial Municipal de São Paulo	57
4.2.1	Publicação de artigos	59
4.2.2	Estrutura dos arquivos	60
4.3	Diário Livre	62
4.3.1	Levantamento de Requisitos	62
4.3.2	Arquitetura	64
4.3.3	Implementação	65
5	Discussão e Resultados	75
5.1	Ferramenta	76
5.1.1	Estatísticas de Acesso	79
5.1.2	Questionário Online	84

5.1.3	Críticas e Limitações	91
5.1.4	Considerações	93
5.2	Interesses e Desafios	94
5.2.1	Questionário do Primeiro Evento	99
5.3	Participação	109
5.4	Apropriações Externas	110
6	Considerações Finais	113
	Referências ¹	115
	Apêndice A–Questionário Impresso	121
	Apêndice B–Primeiro Evento	129
	Apêndice C–Segundo Evento	133
	Apêndice D–Entrevista	135
	Apêndice E–Detalhamento das Outras Bases	137
E.1	Frota Municipal	137
E.2	Imóveis	138
E.3	Convênios	141

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Os Diários Oficiais (DOs) são fontes extremamente ricas de informação governamental, pois é através deles que os atos oficiais não somente se tornam públicos como também passam a ser considerados válidos. Ou seja, só são aplicáveis a partir do momento em que são publicados ou, em casos especificados nos próprios atos, em prazos que se contam a partir dessas publicações. No Brasil essa prática remonta a 1808 e, desde então, as publicações diárias buscam garantir o acesso universal do cidadão aos atos públicos e também o registro histórico dos mesmos.

Naturalmente esse tipo de publicação sofreu o impacto da chegada de novas tecnologias de informação e comunicação, de forma que boa parte das empresas que editoram e publicam DOs no Brasil os oferecem hoje tanto impressos como online ([ABRAMO, 2007](#)). Porém, a versão digital desses documentos geralmente é oferecida em formato PDF, o que dificulta a reutilização automatizada da informação publicada.

É justamente nesse aspecto que pode ser útil o conceito de Dado Governamental Aberto (DGA), uma vez que busca permitir novas formas de apresentação e apropriação dos dados públicos pela sociedade. Para isso são defendidos uma série de princípios de publicação de dados ([BERNERS-LEE, 2010](#); [THE 8 PRINCIPLES OF OPEN GOVERNMENT DATA, 2007](#); [SUNLIGHT FOUNDATION, 2014](#); [OPEN KNOWLEDGE FOUNDATION, 2012](#); [WORLD BANK'S OPEN GOVERNMENT DATA WORKING GROUP, 2013](#)), na esperança de que, ao segui-los, impactos positivos sejam gerados.

Muitos são os potenciais benefícios provenientes dos DGAs. Entre eles o aumento da transparência, aprimoramento de serviços públicos, combate à corrupção, possibilidade de inovação e aumento de eficiência governamental, abertura para maior participação da população, desenvolvimento de novos mercados utilizando os dados disponibilizados, melhorias na tomada de decisões etc ([UBALDI, 2013](#); [HALONEN, 2012](#); [JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012](#)). Apesar do movimento de DGA ter ganhado força nos últimos anos, influenciando governos pelo mundo todo e favorecendo a criação de inúmeras iniciativas relacionadas, os impactos reais produzidos ainda são pouco conhecidos ([HALONEN, 2012](#); [DAVIES; PERINI; ALONSO, 2013](#)). Os estudos preliminares indicam, contudo, que tais impactos podem ser tanto positivos como negativos, contrariando algumas expectativas mais otimistas ([GURSTEIN, 2011](#)).

No mesmo sentido, [Janssen, Charalabidis e Zuiderwijk \(2012\)](#) debatem que a abertura dos dados também é um ponto pouco explorado do tema. Eles atentam para uma série de mitos relacionados à questão, destacando que a disponibilização de dados geralmente não é interessante como um fim em si, mas sim como um meio de se produzir algum impacto social positivo, fato que dificilmente irá ocorrer se a sociedade não conseguir ou não possuir interesse em reutilizar os dados disponibilizados.

De maneira semelhante, [Denis e Goëta \(2014\)](#) defendem que o esforço dispendido dentro da máquina pública para disponibilizá-los é ainda uma face pouco pesquisada da mesma questão. Entrevistando gestores públicos na França eles defendem que, antes dos dados poderem ser disponibilizados, é necessário que haja uma série de processos internos ao órgão público: descobrir quais bases ele possui (exploração), extrair essas bases dos sistemas internos (extração) e, por fim, tratá-las de forma a se tornarem genéricas o bastante para serem usáveis por pessoas externas ao órgão (fase a qual deram o nome de ‘*Rawification*’).

O artigo deles, entretanto, não chega a abordar o que poderíamos chamar de um quarto passo: a disponibilização dos dados em si. Este quarto passo, como pode ser visto nos trabalhos de [Gurstein \(2011\)](#) e [Benjamin, Bhuvaneswari e Rajan \(2007\)](#), é fruto de uma série de escolhas não triviais e, muitas vezes, políticas. Membros de movimentos sociais, gestores públicos, jornalistas, acadêmicos e *hackers*¹ são usuários frequentes de dados governamentais, apesar disso, eles costumam ter necessidades distintas quanto à disponibilização dos mesmos. Uma cópia de uma base em *SQL* pode ser perfeita para *hackers*, mas totalmente inútil para o cidadão comum. O contrário pode ocorrer com a disponibilização de dados em formato PDF.

No trabalho de [Oliveira \(2013\)](#) pode ser encontrado um exemplo de tal problema. Ao entrevistar gestores, responsáveis por um portal de transparência, e cidadãos potenciais usuários do mesmo portal, revelou-se que, apesar dos gestores considerarem o portal acessível, cerca da metade dos cidadãos entrevistados possuía opinião contrária. De maneira semelhante, [Davies e Frank \(2013\)](#), acadêmicos, encontraram problemas ao tentar reutilizar

¹ Neste trabalho será utilizada a seguinte definição para o termo hacker: “Hacker é o ator dotado de conhecimento técnico – acima da média – sobre o uso das TICs, que através deste seu conhecimento trabalha de forma colaborativa, em busca de soluções para problemas tecnológicos” ([ALBANO, 2014](#)). Ressalta-se que essa não é a única existente, porém, parece ser o sentido mais usado para o termo dentro da área de Dados Abertos.

dados disponibilizados pelo governo como abertos, e se utilizaram da situação para explorar o processo que levou àquela disponibilização, incompleta ao uso que almejavam dar.

Um outro tipo de desafio enfrentado pela abertura de dados não diz respeito aos dados em si, mas à vontade política de abri-los. Entrevistando gestores, [Alfano \(2011\)](#) discute os desafios organizacionais internos aos órgãos públicos, quando os esforços de abertura se chocam com a cultura burocrática vigente.

As iniciativas de DGA geralmente visam à publicação de dados numéricos e não especificamente documentos, como aqueles que constituem os DOs. Apesar de existirem na literatura alguns trabalhos que tratam de DGA aplicados aos DOs ([RICHARDS, 2010](#); [CIFUENTES-SILVA; SIFAQUI; LABRA-GAYO, 2011](#); [BRANDAO et al., 2013](#)), esses trabalhos são escassos e não abordam a questão em sua totalidade sociotécnica. É com o preenchimento dessa lacuna que este projeto busca contribuir.

Dessa forma, visando a uma abertura que atente para as necessidades da sociedade e possibilite o reuso dos dados, esse trabalho se inseriu e colaborou com um processo de abertura de dados governamentais, buscando também espaços de participação com a sociedade. Pelo fato desse projeto ter sido desenvolvido em parceria com o órgão público responsável pela abertura das bases, acredita-se que foi possível uma visão mais aprofundada dos desafios técnicos e organizacionais envolvidos no processo, permitindo uma análise mais detalhada do mesmo.

Tendo-se em vista também, como mostrado por [Benjamin, Bhuvaneswari e Rajan \(2007\)](#), que a abertura de dados governamentais pode envolver o interesses de diversos grupos sociais, a análise desse contexto torna-se fundamental na implantação de um sistema para a disponibilização dos dados. Este projeto, através da análise do contexto, pretende trazer luz também para esse aspecto do tema.

Para possibilitar a publicação dos dados em formato aberto ao longo do processo foi desenvolvida uma ferramenta de código aberto que atendesse aos requisitos levantados pelos grupos envolvidos. A ferramenta, batizada de Diário Livre, vem recebendo um número crescente de acessos e boas avaliações por parte dos usuários. Ao fim do processo foram avaliados os resultados produzidos, assim como críticas e questionamentos levantados ao longo do mesmo.

1.1 Justificativa

Os DGA, apesar de não serem novos, tem ganhado mais espaço recentemente, começando a influenciar governos pelo mundo todo ([HALONEN, 2012](#)). Os impactos de tal fato ainda são pouco conhecidos, mas estudos apontam que podem ser positivos ou negativos e afetar populações inteiras ([BENJAMIN; BHUVANESWARI; RAJAN, 2007](#)).

Apesar disso, não foram encontrados estudos que documentassem em detalhe os processos de abertura de bases de dados governamentais, tanto do ponto de vista tecnológico como social, analisando inclusive o contexto e interesses envolvidos, fatores que podem influenciar os impactos finais do projeto ([GURSTEIN, 2011](#)). A literatura também não fornece uma metodologia própria para auxiliar no desenvolvimento de tais processos, fazendo desse tema um contexto de implementação de sistemas de informação ainda pouco explorado, e tornando esse projeto relevante.

Além disso, pode-se dizer que uma hipótese motivadora deste trabalho é facilitar o reuso dos dados, no caso, do Diário Oficial. Outro fator que aumenta o grau de relevância deste trabalho é o fato dele ter ganho a possibilidade de estudar o tema da perspectiva interna da administração pública enquanto a abertura de dados está sendo iniciada, algo que, novamente, não foi encontrado na literatura.

1.2 Objetivos

Esse projeto se dedica a estudar o processo de implantação de um sistema de informação para a publicação de um Diário Oficial (DO) em formato aberto. O estudo se deu participando do processo de abertura de dados de um DO, tanto do ponto de vista organizacional, como técnico. Como objetivos específicos pretendeu-se:

- Desenvolver uma ferramenta que permita a publicação do DO em formato aberto.
- Envolver atores de diversos segmentos da sociedade na etapa de levantamento de requisitos.
- Narrar o processo desenvolvido, os interesses envolvidos e obstáculos encontrados, especialmente aqueles relacionados à implantação de dados abertos pela administração pública.

1.3 Estrutura Geral do Trabalho

Buscando auxiliar a leitura deste documento, esta seção faz um breve esboço da estrutura do mesmo. O [Capítulo 1](#), no qual esta seção se encontra, apresenta o tema a ser tratado, procurando mostrar também a relevância do mesmo na cena global. Neste mesmo capítulo estão apresentados os objetivos que pautaram este trabalho. No [Capítulo 2](#), são descritos alguns conceitos fundamentais e trabalhos que influenciaram este projeto. Em seguida, o [Capítulo 3](#) descreve a metodologia a ser utilizada e os motivos de sua escolha.

No próximo bloco, [Capítulo 4](#), é narrado o desenvolvimento do projeto em si. Inicialmente com uma contextualização, buscando apresentar os atores envolvidos nas primeiras fases do trabalho e como se deu a relação entre eles. Passa-se então para um maior detalhamento do cenário encontrado e da ferramenta desenvolvida. Já o [Capítulo 5](#) concentra as análises dos dados coletados e discute as limitações e os resultados encontrados ao longo do projeto. Por fim, o [Capítulo 6](#) tece algumas considerações finais sobre o presente trabalho.

2 Revisão Bibliográfica

Neste capítulo, a [seção 2.1](#) traz um breve resumo da história dos DOs, do papel deles e trabalhos acadêmicos relacionados a eles e a DGA. A [seção 2.2](#), por sua vez, conceitua o termo DGA, discute as consequências da publicação desses dados ([subseção 2.2.1](#)) e as recomendações direcionadas a iniciativas do tipo ([subseção 2.2.2](#)).

2.1 Diários Oficiais

A publicidade das leis e atos governamentais é um dos pressupostos mais básicos da democracia, uma vez que é necessária para permitir uma possível contestação dos mesmos ([HASWANI, 2012](#)). Além disso, a publicação das leis também é fundamental para a eficácia destas, uma vez que torna-se mais difícil ainda aplicá-las sem um consenso sobre seu conteúdo, ou mesmo sobre sua existência ([XAVIER, 2012](#)).

Para minimamente cumprir esse papel, governos ao redor do mundo costumam publicar diariamente documentos chamados de diários oficiais. A história destes remonta, em alguns casos, ao século XVII ([THE GAZETTE, 2014](#)).

Conforme narra [Sodré \(1998\)](#), no Brasil, o primeiro diário oficial foi criado em 10 de setembro de 1808, fruto da fuga da corte de Portugal. Denominado de Gazeta do Rio de Janeiro, ele marca também o início da imprensa no país, uma vez que foi o primeiro jornal impresso. Ele tratava não apenas da publicação de atos e leis, mas também de notícias sob a perspectiva da coroa.

Hoje o *descendente* desse jornal se chama Diário Oficial da União (DOU), é publicado pela Imprensa Nacional e regulamentado por legislação própria ([BRASIL, 2002](#)). A mesma entidade publica também o Diário da Justiça, relacionado aos processos judiciais ([BRASIL, 2006](#)).

Os estados e municípios possuem seus próprio diários. No caso do Diário Oficial Municipal de São Paulo (DOM), segundo a gestora, são publicados cerca de 800 artigos por dia, distribuídos em aproximadamente 90 páginas, e estima-se que em sua existência cerca de 30 milhões de páginas tenham sido publicadas.

Legalmente, assume-se que os cidadãos passam a ter conhecimento de uma lei ou ato quando estes são publicados em um diário oficial. Porém, tal publicação não garante que sejam conhecidos pelos cidadãos, como mostra [Haswani \(2012\)](#). O grande

volume de informações publicadas diariamente e a linguagem adotada por esses meios são barreiras praticamente inviáveis de serem transpostas pela maior parte da população. É justamente nesse aspecto que pode ser útil o conceito de Dado Governamental Aberto (DGA), desenvolvido na seção seguinte, uma vez que busca permitir novas formas de apresentação e apropriação dos dados públicos pela sociedade.

Corroborando a relevância deste trabalho, não foram encontrados muitos estudos sobre iniciativas de DGA envolvendo DOs. Um dos poucos trabalhos encontrados comenta a iniciativa de dados abertos das Filipinas. Citando um relatório da *Open Government Partnership*, o autor afirma que os primeiros dados a serem publicados pela iniciativa foram versões digitais do DO daquele país, porém, a qualidade dos metadados foi criticada, uma vez que não possuía um padrão claro (DAVIES, 2014).

Já em outro artigo, os autores foram obrigados a minerar os PDFs do DOU brasileiro para poder anotar e relacionar artigos entre si (BRANDAO et al., 2013). Um dos resultados obtidos foi identificar qual ato revoga ou altera outro, porém, o domínio abordado foi apenas o do Tesouro. Para identificar os atos no texto dos artigos, utilizaram a base do sistema LexML do Senado brasileiro, o que aparentemente auxiliou bastante no processo. Por outro lado, tiveram uma dificuldade maior na identificação dos assinantes dos atos, uma vez que não há um vocabulário controlado quando se trata de nomes de pessoas. Os autores não deixam claro os problemas encontrados na extração do texto dos PDFs ou se esse processo pode ter afetado os resultados. Porém, comentam que o texto no PDF estava em várias colunas, o que deve ter gerado dificuldades na extração.

Outro artigo encontrado discute a publicação via SPARQL de dados sobre as leis do Chile (CIFUENTES-SILVA; SIFQUI; LABRA-GAYO, 2011), adotando o enfoque do estudo de uma arquitetura para a publicação de dados governamentais ligados. Apesar de detalhar a arquitetura adotada no oferecimento e documentação dos dados, não fica claro qual o formato inicial da base utilizada e nem se o escopo abrange tudo que é publicado pelo respectivo DO. Analisando o site resultante do projeto, aparentemente os dados ligados fornecidos são referentes aos metadados mais gerais das leis (como data de publicação, número, língua e título), mas os textos das leis em si não chegam a ser decompostos em metadados.

É também bastante interessante a iniciativa premiada *Federal Register 2.0* (RICHARDS, 2010) que busca converter o material impresso para arquivos XML processáveis por máquina. O projeto tem como objetivo tornar o DO estadunidense mais acessível e

fácil de processar e compartilhar entre pessoas e máquinas. Apesar de ser citado em alguns trabalhos como um exemplo a ser seguido ([LINDERS, 2012](#); [DUDLEY, 2011](#)), não foram localizados artigos na literatura científica sobre detalhes de seu projeto e implementação.

Uma das poucas análises mais gerais sobre DOs no Brasil foi publicada em 2007, pela ONG Transparência Brasil, e abordou o DOU e os DOs dos Estados e do Distrito Federal. Chamado “Gutenberg em bits”, o estudo conclui o seguinte:

O resultado é uma modernidade cosmética, uma vez que o produto transformado em bits é tratado estaticamente, como se fosse um produto impresso. Na quase totalidade dos casos, a única diferença entre o DO impresso e o DO apresentado na rede é o suporte, que em vez do papel passou a ser a tela do monitor. Meramente se reproduz o produto impresso no meio eletrônico, sem que a informação seja organizada de modo a possibilitar uma recuperação por meio da definição de critérios de agregação (agrupamento de informações de mesmo tipo publicadas em dias diferentes) e filtragem ([ABRAMO, 2007](#)).

2.2 Dados Governamentais Abertos

Em dezembro de 2007, trinta pensadores e ativistas da Internet¹ se reuniram em Sebastopol, São Francisco, para definir o conceito de DGA. O encontro teve como plano de fundo as eleições estadunidenses para a presidência, e serviu como mais uma ação para pressionar os candidatos rumo a uma adoção mais ampla de DGA ([CHIGNARD, 2013](#)).

Neste trabalho, como definição de DGA, são usados os oito princípios lançados por alguns dos criadores do movimento DGA ([THE 8 PRINCIPLES OF OPEN GOVERNMENT DATA, 2007](#)). Estes princípios foram traduzidos abaixo:

1. **Completos:** Todos os dados públicos (não restritos pela legislação) devem estar disponíveis.
2. **Primários:** Os dados devem ser coletados da fonte e não serem agregados.
3. **Atuais:** Os dados devem ser disponibilizados em um tempo máximo que preserve o valor dos mesmos.
4. **Acessíveis:** Os dados devem ser acessíveis à maior variedade de usos e propósitos possível.
5. **Processáveis por máquina:** Os dados devem estar minimamente estruturados de modo a permitir processamento automático dos mesmos.

¹ Entre eles Tim O’Reilly, Lawrence Lessig, Tom Steinberg e Aaron Swartz.

6. **Não discriminatório:** O acesso deve ser aberto a qualquer indivíduo, sem necessidade de registro.
7. **Não proprietário:** O formato no qual os dados são disponibilizados não pode estar sob o controle de uma única organização.
8. **Licença Livre:** Os dados não podem ser regulados por licença fechada, patente etc.

Também são observados os novos princípios disponíveis no mesmo site:

- **Gratuitos e Online:** Os dados devem estar disponíveis na Internet e sem cobrança de acesso.
- **Permanentes:** O formato e local dos dados deve permanecer constante pelo maior tempo possível.
- **Confiáveis:** Os dados devem permitir algum tipo de verificação quanto à publicação original, observado que isso não impeça a alteração dos documentos pelo público.
- **Presumir-se a abertura:** O contexto legal deve presumir que, se não atestado o contrário, os dados são abertos.
- **Documentados:** Dados devem estar documentados quanto ao formato e significado dos mesmos.
- **Seguros de abrir:** Formato utilizado não pode permitir conteúdo executável, evitando o risco de conteúdo malicioso.
- **Projetado com consulta pública:** O público está na melhor posição para dizer quais tecnologias são adequadas na disponibilização, logo, consultá-lo é essencial para que os dados sejam reutilizados.

Além disso, é importante citar o modelo das 5 estrelas usado na área ([BERNERS-LEE, 2010](#)). Nele as estrelas são atribuídas da seguinte maneira:

- 1 Estrela:** Dados disponíveis sob uma licença aberta.
- 2 Estrelas:** Dados estruturados (ex.: tabela ao invés de PDF).
- 3 Estrelas:** Dados em um formato não proprietário (ex.: CSV ao invés de XLS).
- 4 Estrelas:** Uso de identificadores únicos (URI) para denotar elementos nos dados.
- 5 Estrelas:** Dados ligados a outros dados.

Esse projeto visará atender pelo menos aos oito primeiros princípios, e às 4 estrelas descritas acima.

2.2.1 Desafios da Abertura de Dados Governamentais

Muitos são os potenciais benefícios provenientes dos DGAs. Entre eles o aumento da transparência, aprimoramento de serviços públicos, combate à corrupção, possibilidade de inovação e aumento de eficiência governamental, abertura para maior participação da população, desenvolvimento de novos mercados utilizando os dados disponibilizados, melhorias na tomada de decisões etc (UBALDI, 2013; HALONEN, 2012; JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012). Esses pontos são fortes o bastante para justificar o surgimento de iniciativas de dados abertos ao redor de todo o mundo. Apesar disso, os resultados dessas iniciativas podem nem sempre ser positivos, o que ressalta a importância de uma análise mais crítica sobre o tema.

Sobre DGA, Janssen, Charalabidis e Zuiderwijk (2012) e Halonen (2012) defendem que pouco adianta a publicação dos dados se deles não resultarem impactos positivos para a sociedade. Em outras palavras, dados abertos são um meio, não um fim (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012; POLLOCK, 2011). Halonen (2012) também defende que muitas das expectativas de impactos positivos dos DGA ainda não se concretizaram.

A ideia, por exemplo, de que a população passaria a usar, em massa, dados abertos para fiscalizar o governo, tem-se revelado como mito, pois ainda poucas pessoas se dedicam a fiscalizar as contas das autoridades públicas. Geralmente quem o faz são grupos organizados da sociedade civil, como ONGs, jornalistas e empresas, ou são indivíduos que já estão envolvidos no tema por outros motivos. Logo, ao contrário de algumas expectativas, as leis de acesso à informação não engajaram mais pessoas para discutir a vida pública e fiscalizar os gestores e políticos; elas apenas forneceram melhores oportunidades para aqueles que já faziam isso (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012; HALONEN, 2012). O que, entretanto, pode ainda ser visto como um impacto positivo relevante.

Por parte dos gestores públicos, uma reclamação percebida por Halonen (2012) foi quanto ao trabalho extra necessário para viabilizar a publicação dos dados, trabalho este muitas vezes não contabilizado na implantação de iniciativas do tipo. Segundo estimativas, na Inglaterra, a resposta a um pedido de acesso à informação pode custar entre £200 e £800 (COLQUHOUN, 2010).

Outro medo recorrente entre gestores públicos é o uso dos dados de forma sensacionalista pela imprensa. Isso poderia ocorrer, por exemplo, com dados mais detalhados sobre os gastos públicos, especialmente os menores. Gestores afirmam que os pequenos

gastos, fora de contexto, poderiam ser usados pela mídia de forma oportunista, enquanto os grandes montantes não seriam fiscalizados. Essa cultura do “te peguei!” poderia contribuir então para uma sociedade com menos confiança entre população e gestores ([HALONEN, 2012](#); [GRIMMELIKHUIJSEN, 2012](#)). Tony Blair foi um dos que fez tal acusação, se dizendo arrependido de ter aprovado a *Freedom of Information Act*, que regulamenta o acesso à informação na Inglaterra ([ROSENBAUM, 2010](#)). Segundo ele, ela seria mais usada por jornalistas do que por cidadãos, sendo que os primeiros a usariam apenas para “bater” no governo, nas palavras do próprio ex-primeiro ministro.

Outra crítica feita às iniciativas de DGA atuais, é o fato de algumas delas visarem muito mais a permitir às empresas obterem lucro explorando dados abertos do que em favorecer a apropriação dos dados por outros setores da sociedade. Essa crítica é feita, por exemplo, por [Halonen \(2012\)](#) e [Davies \(2011\)](#) quando analisam o *Autumn Statement Open Data Measures* do governo inglês, e afirmam que a iniciativa não se preocupou em auxiliar a sociedade como um todo a ter capacidade de utilizar os dados. Além disso, segundo o último autor, a mesma iniciativa escolhe bases bastante questionáveis para serem abertas. Ela abre bases de dados sobre a saúde dos cidadãos, o que, pelo risco à privacidade, pode gerar desconfiança entre a população quanto a iniciativas de DGA de uma forma geral. Apesar disso, a iniciativa não abre, por exemplo, bases sobre pagamento de impostos por empresas, ou violações legais cometidas por elas ([DAVIES, 2011](#)).

Analisando agora mais a fundo um caso que envolveu a publicação de DGA, discute-se o projeto Bhoomi. Ele consistiu na implementação de um de sistema para controle dos registros de propriedade de terra em um estado da Índia. O projeto foi estudado por [Benjamin, Bhuvaneswari e Rajan \(2007\)](#), e afirmam que ele foi divulgado como uma forma de modernizar e facilitar a comercialização e o controle das terras. Assim, prometeu-se uma redução na corrupção e no tempo de consulta aos documentos. Para a implementação, o governo fez uma parceria com uma grande empresa de TI da Índia, justamente uma das criadoras e defensoras do projeto.

Os mais de 1500 tipos de propriedades de terra², herança histórica da formação nacional da Índia e da cultura local, foram simplificados para 256. O sistema passou a controlar 20 milhões de registros de propriedade, posse de 6,7 milhões de cidadãos, de pequenos a grandes proprietários de terras. O controle dos registros, antes feito por conselhos locais eleitos pelos cidadãos, passou a ser centralizado. Tal atitude foi tomada para evitar a corrupção, uma vez que os conselhos locais eram tidos como corruptos pela empresa que implementou o sistema. O projeto foi considerado um exemplo de boas práticas pelo Banco Mundial.

Apesar das promessas, segundo os autores do estudo, o resultado não foi o esperado (BENJAMIN; BHUVANESWARI; RAJAN, 2007). O tempo necessário para se conseguir uma cópia do registro aumentou de 2 a 3 dias para 3 a 4 meses, devido à centralização e às falhas constantes nos terminais locais destinados ao uso da população. Antes as propinas eram apenas para os oficiais locais, que geralmente negociavam o preço de acordo com o poder aquisitivo do cidadão demandante do serviço. Após o sistema e a centralização, mais pessoas participam do processo, aumentando o valor da propina. Além disso, os oficiais de fora da localidade estabelecem um preço independente do demandante, piorando a situação para os agricultores de menor poder econômico.

A digitalização foi feita sem se preocupar muito se os dados estavam certos ou não. A ideia era que, se estivessem errados, o proprietário pediria uma correção e o sistema iria sendo corrigido de forma iterativa. Entretanto, os agricultores mais pobres são os mais prejudicados, pois os registros de suas terras são justamente os que costumam apresentar problemas, e eles não dispõem de poder aquisitivo para descobri-los e entrar com o pedido de correção.

A disponibilização online dos dados também permitiu que os códigos de registro dos pequenos proprietários fossem usados pelos grandes e médios para se passarem pelos primeiros, conseguindo subsídios do governo destinados apenas para pequenos proprietários. Além disso, no processo de padronização dos tipos de registros de propriedades, alguns

² Para se ter uma ideia do significado de “tipo de propriedade”, os autores citam que é costume na Índia (e em alguns países árabes) situações como a seguinte: O proprietário A de um imóvel de dois quartos os aluga aos inquilinos B e C pelo valor X mensal. Uma pessoa D quer alugar o quarto onde B está pelo valor $X + Y$. Como as leis locais costumam beneficiar o inquilino em detrimento do proprietário, é difícil para o segundo convencer o primeiro a desocupar o imóvel. Logo é corrente a seguinte prática: D divide Y entre os envolvidos. O proprietário A recebe $X + 30\%$ de Y, o inquilino B (que sairá do quarto) recebe 60% de Y e o inquilino C recebe 10% de Y.

tipos, os usados por grupos mais marginalizados, foram excluídos do sistema como forma de deslegitimar ainda mais essas populações e facilitar a desapropriação de suas terras.

Resumindo: com a centralização, os pequenos agricultores perderam a capacidade de controlar o processo, mas os grandes ganharam, uma vez que estavam bem relacionados com as elites centrais. A centralização imposta pelo sistema tornou a comercialização de terras disponível apenas para os grandes atores. Na mesma linha, do ponto de vista do capital globalizado, ficou mais fácil a aquisição de terras para, por exemplo, mega empreendimentos. Porém, para a população local, houve aumento de custo e prazos (BENJAMIN; BHUVANESWARI; RAJAN, 2007).

Os mesmos autores também criticam algumas análises feitas sobre o projeto de digitalização do sistema e que o avaliaram como positivo. Tais análises seriam muito “positivistas”, levando em consideração apenas aspectos gerais do sistema, mas não os contextos e impactos nas comunidades locais. O relatório disponível em um site da Microsoft, por exemplo, seria baseado em um questionário típico de análise de mercado, enquanto o estudo do autor realiza entrevistas a diversos atores para revelar os meandros políticos e econômicos da questão.

De tal forma, o projeto de digitalização se vende como uma “boa prática”, que neutraliza a “política” através da “governança” e reduz complexos processos sociais em nome da “reforma administrativa”. Essa visão é tida pelos autores como positivista e limitada, não levando em consideração quais são os grupos que, relativamente, se beneficiam com as mudanças implantadas. Ou seja, a tecnologia nem sempre exerce um papel positivo. Quando usada em contextos já desbalanceados, ela pode ampliar as diferenças e não diminuí-las.

Ainda segundo os autores, a questão da terra envolve interesses de grupos poderosos, como grandes empresas encabeçando megaempreendimentos, que usam de “parcerias” com a elite judicial e administrativa para consolidar e regularizar terras em seu favor. Nesse contexto, é ingênuo acreditar que um sistema de informação, mesmo que bem intencionado, conseguiria resolver questões estruturais tão fortes (BENJAMIN; BHUVANESWARI; RAJAN, 2007).

É importante notar que o estudo descrito acima não se trata apenas de um sistema de publicação de DGA, uma vez que o sistema implantado alterou também a coleta e controle dos dados. Porém, é importante ter em mente as críticas feitas, uma vez que

desmistificam uma série de crenças relacionadas à implantação de sistemas de informação por governos.

Tratando agora de experiências positivas com DGA, [Gurstein \(2011\)](#) discute o processo judicial de Zanesville, Ohio, caso inclusive utilizado por Tim Berners-Lee em uma apresentação que promoveu dados abertos ([BERNERS-LEE, 2009](#)). Nele, uma comunidade constituída majoritariamente por negros, processa a prefeitura por discriminá-la na distribuição de água, uma vez que a administração pública vinha atendendo principalmente as comunidades predominantemente brancas. O veredicto puniu a prefeitura em \$10,9 milhões.

A acusação se baseou em dados abertos, gerando mapas para provar as distorções. Porém, o uso dos dados abertos nesse caso só foi possível porque eles tiveram ajuda de diversos especialistas em processamento de dados geoespaciais, e um advogado qualificado para usá-los. [Gurstein \(2011\)](#) levanta o ponto de que a maioria dos cidadãos não possui a mesma “sorte”, ao exemplo do caso Bhoomi já discutido anteriormente, e fala de indícios de que o mesmo problema envolvendo propriedades de terra ocorre em Nova Scotia, Canada.

Outro caso significativo envolve o Centro de Pesquisas Públicas para a Saúde, na Califórnia. A instituição realiza, duas vezes por ano, uma pesquisa para analisar a saúde dos californianos, e a base de dados é disponibilizada online. Porém, além de disponibilizar os dados, o centro oferece cursos gratuitos para que qualquer cidadão possa aprender a analisá-los. Usando tal oportunidade, um grupo de moradores de Solano County, California, se apropriou dos dados e conseguiu demonstrar uma alta incidência de asma em sua localidade, argumentando contra a construção de outra parada de caminhões no local ([GURSTEIN, 2011](#)).

Comparando os três casos, [Gurstein \(2011\)](#) defende que a disponibilização dos dados em si não é suficiente para produzir impactos positivos, podendo inclusive ter consequências negativas. Ainda segundo o autor, um dos fatores principais que afeta os impactos produzidos, é a capacidade da comunidade de se apropriar dos dados disponibilizados. Se ela não tiver, por exemplo, conhecimento técnico, e o publicante dos dados não fornecer maneiras de superar essa barreira, é provável que ela fique dependente de um terceiro que se disponha a ajudar, o que pode não acontecer ([HALONEN, 2012](#)). [Gurstein \(2011\)](#) defende então que um publicante responsável deve tomar medidas para que os possíveis beneficiários dos dados tenham condições concretas para fazê-lo.

Na próxima seção são discutidas, entre outros assuntos, quais seriam as questões as quais um publicante deveria estar atento quando disponibiliza dados.

2.2.2 Requisitos para a Abertura de Dados

Como discutido anteriormente, uma diversidade de fatores, entre eles barreiras tecnológicas, podem fazer com que apenas setores da sociedade mais favorecidos utilizem os DGA. Tais assimetrias de acesso poderiam então ampliar as desigualdades ao invés de reduzi-las ([BENJAMIN; BHUVANESWARI; RAJAN, 2007](#); [GRIMMELIKHUIJSEN, 2012](#); [GURSTEIN, 2011](#)). Além disso, parte do movimento de DGA enxerga tais dados apenas como mais um grande mercado a ser explorado pelas empresas, não necessariamente preocupados com a expansão dos benefícios oriundos da abertura dos dados à toda a população ([HALONEN, 2012](#)). Nesse contexto, fica evidente que as escolhas de implementação de um sistema de informação, que possibilite a abertura dos dados, podem ser também escolhas políticas que influenciarão na apropriação dos mesmos.

Dessa forma, algumas das medidas apontadas para a publicação de DGA, buscando promover o uso dos mesmos e produzir impactos positivos, são uma maior participação da sociedade no processo de abertura e apoio para que grupos não capacitados tecnologicamente também possam fazer uso dos dados ([HALONEN, 2012](#); [GURSTEIN, 2011](#)).

No mesmo sentido, [Halonen \(2012\)](#) ressalta a importância de empoderar aqueles que já estão fazendo algum tipo de uso de DGA, assim como conectar usuários, desenvolvedores e gestores públicos, pois os gestores geralmente não conseguem visualizar quais as melhores formas de se utilizar os dados. Tal interação entre esses grupos também é fundamental para a escolha de quais dados devem ser abertos e de que forma publicá-los. Mas, infelizmente, esse tipo de interação não é regra, o que permite, por exemplo, que muitos gestores publiquem dados de uma forma que eles consideram aberta, mas que na verdade não é, pois não permite uma série de usos.

Almejando vislumbrar um panorama mais amplo, foram procuradas recomendações feitas por organizações que advoguem a publicação de DGA. Os três conjuntos de recomendações encontrados estão sistematizados na [Tabela 1](#) ([SUNLIGHT FOUNDATION, 2014](#); [OPEN KNOWLEDGE FOUNDATION, 2012](#); [WORLD BANK'S OPEN GOVERNMENT DATA WORKING GROUP, 2013](#)).

Tabela 1 – Comparações entre as recomendações para processos de abertura da dados governamentais feitas por três organizações que advogam DGA. Os itens marcados com “✓” indicam pontos defendidos por aquele documento. Quando necessário, está explicitada a forma como o documento trata o item. Quando o documento não aborda o item em questão, o símbolo “–” é usado .

Organização	Sunlight Foundation	Open Knowledge	World Bank
Documento	Open Data Policy Guidelines	Open Data Handbook	Open Government Data Toolkit
Preocupação com a viabilização política	Basear-se em leis e valores existentes	Trabalhar medos e criar protótipos	Debates internos e protótipos
Uso de licença aberta para os dados	Domínio Público	✓	✓
Permitir exploração comercial dos dados	✓	✓	✓
Participação social na escolha dos dados a serem abertos	Participação direta, mas não apenas	Voltada a intermediários, feita através de consulta pública a eles	–
Criação de um catálogo de dados	✓	✓	–
Publicação dos dados em formatos abertos	✓	✓	–
Uso de software livre na disponibilização	✓	✓	–
Grupos com os quais realizar parcerias para auxiliar no processo	Pesquisadores e ONGs	Intermediários	Desenvolvedores e jornalistas
Formas de incentivar o consumo dos dados	Parcerias	Eventos e hackathons	Eventos, hackathons e cursos

Fonte: autoria própria

Os três documentos abordam a questão da viabilização política, o que mostra a importância da mesma. As recomendações relacionadas vão no sentido de trabalhar a cultura já vigente dentro da organização, utilizando projetos protótipos para a experimentação com dados abertos e demonstrar seus benefícios.

Quanto às licenças para os dados, há consenso de que sejam abertas. Esse aspecto está, em parte, relacionado ao fato dos três documentos defenderem também a possibilidade de exploração comercial desses dados, algo que licenças mais restritivas poderiam limitar.

Dois dos documentos parecem defender explicitamente um processo de abertura mais participativo. Tal aspecto se manifesta em vários pontos. Primeiramente ao defenderem que a escolha dos dados a serem abertos passe por uma consulta a grupos externos à administração pública. Em seguida ao sugerirem um catálogo de dados, permitindo à comunidade conhecer os dados existentes e atuar com maior propriedade na priorização da abertura.

Outro ponto importante, que também favorece a participação, é a defesa do uso de formatos abertos e software livre, possibilitando alguma colaboração no desenvolvimento das ferramentas utilizadas e uma maior interação e apropriação dos dados disponibilizados. Além disso, também relacionado à participação, os três documentos concordam em realizar algum tipo de parceria com intermediários para auxiliar no processo como um todo ou em alguma fase específica dele.

Por fim, mais especificamente sobre o consumo dos dados, os documentos também discutem fomentá-lo através de eventos (como hackathons ou cursos) ou parcerias. Pode-se notar então que, como um todo, não há grandes divergências entre os três documentos, apenas algumas diferenças de enfoque, mas que podem funcionar de maneira complementar.

3 Metodologia

Nesse capítulo são inicialmente discutidos os conceitos Pesquisa-Ação e Design Science Research, uma vez que eles balizaram o desenho da metodologia e permearam o projeto como um todo. Na seção consecutiva é justamente detalhada a metodologia adotada neste trabalho, descrevendo suas etapas, ferramentas e limitações.

3.1 Pesquisa-Ação e Design Science Research

A metodologia empregada neste trabalho se inspira nos moldes da Pesquisa-Ação (PA) e da Design Science Research (DSR). Tal escolha se deu principalmente porque essas metodologias prevêm e defendem a intervenção no processo, algo que ocorreria neste trabalho devido à parceria.

A PA não é uma única metodologia, mas uma família delas, usadas em diversas áreas do conhecimento. Não parece haver um consenso na terminologia usada para indicar as vertentes dentro dessa família, mas Pesquisa-Ação Participativa ou apenas Pesquisa Participativa, costuma ser usada para as metodologias que permitam uma maior inclusão dos envolvidos na pesquisa ([BASKERVILLE, 1999](#)). Apesar disso, [Reid \(2004\)](#) parece utilizar o termo Pesquisa-Ação de forma equivalente a Pesquisa Participativa.

De forma semelhante à PA, também não parece haver consenso sobre a definição de DSR ([LACERDA et al., 2013](#)). [Järvinen \(2007\)](#) chega inclusive a afirmar que DSR e PA são equivalentes. Segundo [Vaishnavi e Kuechler \(2004\)](#), a DSR se diferencia do design convencional pois busca produzir conhecimento novo, abordando lacunas geralmente evitadas pelo design convencional.

Segundo [Baskerville \(1999\)](#), a PA é fundamentada na prática, porém há um compromisso em produzir conhecimento acadêmico. Além disso, essa metodologia parte do pressuposto que sistemas humanos não podem ser reduzidos para serem estudados, já que perderiam seu significado real quando isso é feito. Dessa forma, essa categoria de problemas precisa ser abordada em sua complexidade, não tentando simplificá-la e reduzi-la em partes.

Segundo [Vaishnavi e Kuechler \(2004\)](#), a DSR também é fundamentada na prática, tendo um foco na criação de artefatos que funcionem, mais do que propriamente em entendê-los.

A PA costuma levar a análises: idiográficas, interpretativistas (ou antipositivistas) e qualitativas, característica aparentemente menos frequentes à DSR ([BASKERVILLE, 1999](#); [LACERDA et al., 2013](#)). O ponto de vista interpretativista defende que, quando o pesquisador tenta entender o que observa, sua visão pessoal dos fatos acaba influenciando essa análise. Dessa forma, o contexto social onde o estudo ocorre, e o do próprio pesquisador, afetam os resultados obtidos. Já a perspectiva idiográfica assume que cada situação possui um conjunto de especificidades únicas, que deve também ser levado em consideração nas análises, e quando os resultados forem comparados ou aplicados a outras situações. Ou seja, o pesquisador aborda um problema em um contexto social específico, mas de forma a contribuir com o desenvolvimento da teoria geral. Uma das consequências de tais abordagens é o uso e produção de dados qualitativos.

Metodologias de PA sempre envolvem uma equipe, e fazem com que o pesquisador observe e participe do fenômeno estudado. Não há o conceito de um pesquisador distanciado, pois não se acredita que há como existir esse distanciamento, uma vez que análises feitas por um ser humano sempre estarão carregadas de certa subjetividade. Ainda segundo [Baskerville \(1999\)](#), as primeiras formas de PA possuíam dois estágios:

1. Diagnosticar colaborativamente a realidade formulando teorias.
2. Intervir na realidade e observar os efeitos resultantes.

Ao longo dos anos surgiram muitas variedades de PA, mas, de uma forma geral, elas compartilham alguns pontos centrais:

- Orientação à ação.
- Foco na resolução de problemas.
- Processo orgânico envolvendo estágios iterativos.
- Colaboração entre os participantes.

Ainda segundo o autor, a abordagem mais consolidada possui um momento inicial de estabelecimento do ambiente de pesquisa e então a iteração cíclica de 5 fases:

- **Diagnóstico:** formulação de teorias sobre as causas dos problemas da organização.
- **Planejamento da Ação:** definição das ações a serem tomadas para combater os problemas, baseando-se em um arcabouço teórico, e registro dos objetivos esperados.
- **Ação** - implementação das ações planejadas, pessoalmente ou através de terceiros.

- **Avaliação** - em caso de sucesso avaliar se mudanças realmente contribuíram, em caso de falha adaptar o arcabouço e as hipóteses.
- **Aprendizado Específico**: estruturação do conhecimento adquirido, seja com uma experiência bem sucedida ou não.

Segundo [Vaishnavi e Kuechler \(2004\)](#) a DSR teria as seguintes fases (que se assemelham bastante com as da PA):

- **Descoberta do problema**: definição do objeto de pesquisa.
- **Sugestão**: criação de uma abordagem a ser usada para tentar resolver ou conhecer melhor o problema.
- **Desenvolvimento**: realização das atividades ou construção dos artefatos planejados.
- **Avaliação**: dos resultados obtidos frente às hipóteses levantadas. Hipóteses são raramente descartadas, mas sim adaptadas frente a luz dos novos dados obtidos.
- **Conclusão**: fim de um esforço de pesquisa, quando os resultados são considerados suficientes para o momento, mas raramente esgota o assunto pesquisado.

Dessa forma pode-se dizer que as etapas de ambas as metodologias possuem semelhanças.

No domínio de Sistemas de Informação, metodologias de PA foram aplicadas por [Baskerville \(1999\)](#) e [Kim \(2009\)](#), ambos afirmando obter bons resultados. O primeiro trabalho aponta ainda que, dentro dessa área de pesquisa, metodologias de PA se justificam especialmente onde organizações humanas interagem com sistemas de informação, dada a complexidade dessa interação.

Porém, ambos os trabalhos não dedicam uma atenção especial ao contexto político e às relações de poder envolvidas na implementação dos sistemas, fatores bastante discutidos por [Reid \(2004\)](#). Para esta autora, a PA não deve buscar apenas resultados melhores embasados na prática, mas também atuar de forma a expor as relações de poder, possibilitando sua subversão. Desta forma, é uma metodologia que, através da real inclusão dos participantes no processo, permite o empoderamento deles.

Por fim, pode-se dizer que a metodologia adotada por esse trabalho mistura elementos de DSR e de PA. Da primeira ao produzir um artefato de software bem definido e ter apenas uma iteração do ciclo principal de intervenção e avaliação. Da segunda ao dedicar uma atenção especial ao contexto buscando compreendê-lo e ao assumir uma pers-

pectiva idiográfica e interpretativista. Porém, nenhuma dessas características é exclusiva de cada uma dessas metodologias, uma vez que as próprias metodologias também não estão rigidamente definidas na literatura.

3.2 Metodologia Aplicada

Tendo em vista o exposto na seção anterior, esse projeto buscou aplicar uma metodologia que mescla aspectos da PA e da DSR. Nas primeiras reuniões entre o grupo acadêmico e o grupo da prefeitura, foi determinado o escopo inicial do projeto. Através da parceria, foi constituída a equipe que o desenvolveria. Esta, ao longo de diversas reuniões, cuidou do ciclo de planejamento, ação e reavaliação. Dessa forma o projeto ocorreu em: uma etapa de levantamento de requisitos, uma de construção do protótipo, uma de lançamento do mesmo e, por fim, uma etapa de coleta das impressões dos usuários e demais atores envolvidos. Ressalta-se que, devido às características da metodologia adotada, essas etapas não foram rígidas, havendo interações entre elas. Pode-se dizer que essas etapas, e o processo como um todo, foram influenciados principalmente por:

- Discussões realizadas durante as reuniões periódicas da equipe.
- Contexto interno da prefeitura (articulações com secretarias, atritos entre os grupos internos, disponibilidade dos dados etc).
- Capacidades da equipe (disponibilidade de tempo, conhecimento técnico etc).
- Expectativas dos membros da equipe.
- Contribuições feitas durante o primeiro evento.

De maneira geral as decisões também foram orientadas pelas recomendações encontradas na literatura sobre iniciativas de DGA, estas sistematizadas na [Tabela 1](#). Além disso, como forma de conhecer melhor o contexto, os resultados, orientar as ações do projeto e possibilitar algum grau de participação, foram realizados:

- Questionário presencial no primeiro evento.
- Questionário online no site da ferramenta, aberto durante todo o processo.
- Coleta de estatísticas de acesso da ferramenta.
- Entrevista com gestores responsáveis pela base de dados.
- Um segundo evento de lançamento da ferramenta e coleta de impressões.

Como orienta a PA, buscou-se documentar o processo o máximo possível (gravações de áudio, registro de documentos e e-mails produzidos, diário de pesquisa etc), justamente para facilitar as análises ao longo do mesmo.

Tendo em vista o apresentado aqui, é importante que seja feita uma ressalva. Este trabalho foi desenvolvido através de uma parceria com um órgão da administração pública. Se por um lado tal fato permite uma análise mais aprofundada do contexto interno da máquina pública, e automaticamente adiciona um grupo às decisões do projeto (os gestores públicos); por outro, tal parceria possivelmente impôs restrições à participação de outros grupos, seja pela vontade dos gestores envolvidos, seja por outras restrições impostas a eles próprios. Tais fatores se somam às dificuldades de aplicação de PA nos moldes defendidos por Reid (2004). Outro ponto importante de ser frisado, é o fato de que, para aplicar a metodologia proposta, o autor do projeto interveio no processo, algo previsto e defendido pelas metodologia de PA e DSR. Porém, como dito anteriormente, não houve um controle total do mesmo.

4 Desenvolvimento

Este capítulo se dedica a detalhar o ambiente onde esse projeto foi desenvolvido e quais as ações realizadas. Primeiramente, na [seção 4.1](#), é apresentado o contexto que o envolveu, os atores participantes, a parceria firmada e as outras bases que também foram discutidas paralelamente à do Diário Oficial Municipal de São Paulo (DOM). Em sequência, na [seção 4.2](#), é apresentado o cenário encontrado: o sistema oficial de publicação do DOM, fluxo dos dados e estrutura dos mesmos. Por fim, na [seção 4.3](#), é apresentada a ferramenta fruto deste projeto. São descritos os requisitos que a motivaram, a arquitetura adotada e como se deu a implementação.

4.1 Contexto

Um dos principais atores deste trabalho é o [Colaboratório de Desenvolvimento e Participação da USP¹](#) (Colab-USP). Este é um grupo de pesquisa fundado por ex-integrantes do [Grupo de Pesquisa em Políticas Públicas para o Acesso à Informação²](#), GPPai, ambos vinculados à [Escola de Artes, Ciências e Humanidades da USP³](#), EACH. O Colab-USP é um grupo multidisciplinar, sendo que seus membros estão principalmente ligados ao curso de Sistemas de Informação ou ao de Gestão de Políticas Públicas, ambos da EACH. Isso permite que ele estude a computação sob múltiplas perspectivas, dentre elas a técnica e a social.

Um dos projetos desenvolvidos por seus membros é o [Cuidando do Meu Bairro⁴](#). Esta é uma ferramenta online que busca *geolocalizar* e dispor o orçamento municipal de São Paulo em um mapa. Ela faz isso utilizando dados disponibilizados pela própria prefeitura dessa cidade. Além desse projeto, o Colab-USP também possui diversas pesquisas e trabalhos na área de dados abertos, especialmente quando relacionados a governos.

Outro ator importante para este trabalho é a Controladoria Geral do Município (CGM). Esse órgão é bastante recente no município, tendo sido criado no primeiro ano do mandato do atual prefeito, em 2013, e tem como papel “prevenir e combater a corrupção na gestão municipal, garantir a defesa do patrimônio público, promover a transparência

¹ <http://colab.each.usp.br>

² <http://www.gpopai.usp.br>

³ <http://each.uspnet.usp.br>

⁴ <http://cuidando.org.br>

Figura 1 – Organograma da CGM

Fonte: Prefeitura de São Paulo

e a participação social e contribuir para a melhoria da qualidade dos serviços públicos” ([CONTROLADORIA GERAL DO MUNICÍPIO, 2013](#)). Segundo entrevista com gestor, as controladorias são algo novo nas prefeituras e muitas ainda não possuem uma. Na cidade de São Paulo, a CGM agrupa três setores anteriormente separados, mas que mantiveram suas mesmas funções:

- Ouvidoria: recebe demandas da sociedade.
- Auditoria: averigua demandas levantadas.
- Corregedoria: aplica punições.

Esses três setores foram agrupados para aumentar o diálogo entre eles e reduzir a burocracia envolvida. Passaram também a funcionar no mesmo prédio. Além deles, a CGM é composta por um novo setor: a Coordenadoria de Promoção da Integridade (COPI-CGM). A [figura 1](#) mostra o organograma da CGM.

Esse novo setor, a COPI-CGM, “tem como principal atribuição promover o incremento da transparência pública e fomentar a participação da sociedade civil na prevenção da corrupção” ([COORDENADORIA DE PROMOÇÃO DA INTEGRIDADE, 2013a](#)). Entre suas responsabilidades está a transparência ativa e passiva, mais especificamente atribuições como a garantia do cumprimento da lei de acesso à informação no município e a gestão do portal de transparência e dos dados abertos municipais. A COPI-CGM foi justamente a peça central que atuou como ponte para viabilizar a parceria dentro da administração pública e, uma vez firmada, garantir a coordenação com os demais grupos da prefeitura.

Outro ator relevante dentro da prefeitura foi a Secretaria Municipal de Planejamento, Orçamento e Gestão (SEMPLA). Essa secretaria agrega uma série de funções na administração municipal, entre elas a publicação do DOM. Por ser a guardiã dos dados a serem abertos, ela teve um papel importante na viabilização do processo de diversas maneiras, entre elas a transferência dos dados e das informações necessárias para interpretá-los.

Além disso, ela é a responsável pela contratação da atual publicante oficial do DOM, a Imprensa Oficial. E, caso a ferramenta produto desse projeto vier a ser internalizada pela prefeitura, seria a SEMPLA a responsável por isso, provavelmente através da contratação da PRODAM. Essa última é “uma empresa de economia mista, que apoia a prefeitura na elaboração das políticas de informação e de informática da Cidade” (PRODAM, 2013).

Por fim, o projeto descrito nesta dissertação tem seu início com uma aproximação entre a COPI-CGM e o Colab-USP. O segundo, por ter já utilizado uma base de dados da prefeitura para a construção de um aplicativo de visualização (o *Cuidando do Meu Bairro*⁵), havia se aproximado de certa forma da SEMPLA, provedora da base utilizada, e chamado a atenção da COPI-CGM. Essa última, possuía uma agenda de abertura de bases de dados, mas para que as bases realmente fossem abertas, uma série de requisitos precisavam ser atendidos, entre eles o desenvolvimento de plataformas que permitissem a disponibilização dos dados. Dado o alinhamento de valores entre os dois grupos, a parceria conseguiu ser efetivada.

4.1.1 Escolha da Base de Dados

Antes de procurar o Colab-USP para iniciar a parceria, a COPI-CGM fez uma exploração pelos departamentos da prefeitura à procura de bases a serem disponibilizadas. A prioridade foi dada à disponibilidade das mesmas e foram constatadas as seguintes bases, já existentes, que poderiam ser abertas:

- Arquivos contendo todos os textos e imagens publicados no DOM desde 2003.
- Arquivos em papel sobre todos os imóveis pertencentes à prefeitura.
- Dados sobre todos os imóveis pertencentes à prefeitura, mas atualmente cedidos a terceiros.

⁵ <http://cuidando.org.br>

Além disso, a prefeitura pretende desenvolver um sistema para monitoramento de convênios com Organizações da Sociedade Civil (OSCs), de forma que os dados fiquem abertos para a população, configurando mais uma base:

- Base dos convênios firmados entre a prefeitura e OSCs.

Por fim, um pedido de acesso à informação questionava a prefeitura no sentido de fiscalizar a utilização da frota municipal. Essa informação, no momento, não era coletada pela prefeitura, logo não poderia ser aberta imediatamente. Cogitou-se então construir um sistema para coletá-la e disponibilizá-la.

- Base sobre uso da frota da prefeitura.

Planejou-se também trabalhar na especificação e desenvolvimento de um novo portal de transparência para a prefeitura. De acordo com um gestor, as motivações de escolha desses pontos a serem tratados foram as seguintes:

- **Diário Oficial Municipal (DOM):** os dados já estavam coletados e a COPI-CGM poderia ter fácil acesso a eles, graças a um bom relacionamento com a secretaria gestora da base.
- **Imóveis:** já planejavam sistematizá-los.
- **Convênios:** alta relevância, já que frequentemente são recebidas denúncias de mau uso do dinheiro repassado.
- **Portal Transparência:** algo sob o domínio da própria COPI-CGM, já que ela própria é responsável pelo portal.

Outras bases foram cogitadas, mas preferiu-se não tentar disponibilizá-las por envolverem questões consideradas problemáticas:

- **Casas alugadas para mulheres vítimas de violência doméstica:** se os dados fossem disponibilizados poderiam colocar as envolvidas em risco.
- **Base com relatórios de vigilância sanitária sobre estabelecimentos comerciais ou industriais:** muitas vezes os relatórios contêm segredos sobre os processos de produção, logo a disponibilização dos mesmo poderia gerar problemas jurídicos.

Faz-se necessária a ressalva de que as bases a serem abertas não foram escolhidas através de uma consulta popular. A prioridade de abertura se deu de acordo com a viabilidade e relevância estimadas pelos próprios gestores.

É importante notar também que cada uma das bases escolhidas para abertura possui desafios próprios de disponibilização, mas, devido a restrições de tempo e esforço, esse projeto de mestrado se aprofundará apenas na abertura da base do DOM. Acreditando porém que os processos de abertura das outras bases também constituem uma parte relevante do contexto deste projeto, eles serão comentados em maior detalhe a seguir.

Para melhor visualizar as diferenças entre as bases, foi construída a [figura 2](#) que mostra as transformações pelas quais os dados terão de passar. Na mesma figura, uma diminuição no tamanho do retângulo dentro de uma mesma linha, assim como uma passagem via seta tracejada, indicam perda de dados na transformação.

No caso do DOM, primeira linha do diagrama, os dados já são colhidos de forma digital, representado pelo estado **A**, e são publicado em **B**, que representa o site atual que o publica. Como parte da informação presente em **A** (publicante e tipo de conteúdo) não são publicados em **B**, esse fluxo foi considerado com perda. O sistema planejado para esse caso, representado por **C**, disponibilizaria todas as informações fornecidas por **A**, e em formato aberto.

No caso dos imóveis, segunda linha do diagrama, a base não digitalizada está representada por **D**, que terá uma fração digitalizada em **E**. Uma vez feita essa digitalização, pretende-se disponibilizar a base integralmente em **F**. A base de imóveis cedidos, **G**, se somará então na disponibilização **F**.

No caso dos convênios, terceira linha do diagrama, a informação ainda não começou a ser coletada, **H**. Porém, uma vez que isso seja feito, **I**, diretamente em formato digital, ela será integralmente disponibilizada em **J**.

Por fim, no caso dos veículos, quarta linha, a informação **K** seria coletada em formulários de papel, **L**, para então ser integralmente transcrita em **M**. Alguns dados seriam então omitidos para a disponibilização final **N**.

A passagem da primeira coluna, **Existe**, para qualquer outra, foi considerada sempre com perda, uma vez que é impossível coletar todas as informações existentes no mundo real sobre um determinado tema. Mais informações sobre cada uma dessas outras bases, exceto DOM, podem ser encontradas no [Apêndice E](#).

Figura 2 – Transformações pelas quais os dados terão de passar em cada abertura.

Fonte: autoria própria

4.1.2 Parceria

Uma vez acordada a parceria entre a COPI-CGM e o Colab-USP, o projeto começou a ser delineado. Isso se deu através de reuniões, ocorrendo dentro do próprio ambiente de trabalho dos gestores, em seu horário de expediente. Nas primeiras reuniões de início de implantação do projeto, percebeu-se que a expectativa por parte da COPI-CGM, quanto ao desenvolvimento dos sistemas, era bem maior do que o Colab-USP estava disposto a fornecer.

A primeira esperava, por exemplo, o desenvolvimento por parte do segundo, de um sistema de controle de convênios, que, como pode ser visto no [Apêndice E](#), é algo de relativa complexidade e que exige uma manutenção e suporte técnico constante com a qual a pequena equipe do Colab-USP não poderia arcar.⁶ Uma falha em um sistema desse tipo poderia paralisar toda a prestação de contas das entidades conveniadas à prefeitura, e falhas de segurança poderiam ter consequências bastante graves.

O Colab-USP, por sua vez, pretendia atuar em questões mais simples do ponto de vista técnico, que exigissem menor desenvolvimento e comprometimento a longo prazo, uma vez que o projeto seria principalmente desenvolvido dentro de um mestrado, com data definida para terminar.

Ficou acordado então que o Colab-USP, quanto ao sistema de convênios, auxiliaria apenas na formulação de uma especificação do sistema, e não na implementação do mesmo, de forma análoga ao portal de transparência. Já o desenvolvimento necessário para a disponibilização das bases do DOM, Imóveis e Frota, seria feito pela própria equipe Colab-USP e COPI-CGM. Sendo que a última também contrataria dois estagiários para auxiliar no projeto.

4.1.3 Arcabouço Legal

Outra questão relevante para este trabalho é o arcabouço legal, pois os DOs, sendo instrumentos legais, também são regidos por legislação específica. No caso do Diário Oficial da União, um decreto determina o escopo do conteúdo a ser publicado ([BRASIL, 2002](#)).

⁶ Nesse momento, a equipe de desenvolvimento técnico do Colab-USP que participaria do projeto consistia apenas em uma pessoa, o autor dessa dissertação, que deveria, naturalmente, cuidar também da pesquisa e escrita deste texto.

Alguns atos oficiais, como atas e decisões de tribunais e de órgãos colegiados, devem ser publicados em resumo, porém os seguintes devem ser publicados na íntegra:

- Leis
- Tratados e convenções internacionais
- Medidas provisórias, decretos e outros atos normativos baixados pelo Presidente da República
- Atos dos Ministros de Estado
- Pareceres do Advogado Geral da União e respectivos despachos presidenciais
- Dispositivos e ementas das ações direta de inconstitucionalidade, das ações declaratórias de constitucionalidade e das arguições de descumprimento de preceito fundamental decorrente da Constituição
- Julgamentos do Tribunal de Contas da União
- Atos de caráter normativo do Poder Judiciário

Especificamente sobre DGA, segundo [Albano \(2014\)](#), o governo federal vem implantando uma política de abertura de dados desde 2009, com a publicação de um decreto ([BRASIL, 2009a](#)). Este “atribuiu aos órgãos da Administração Pública Federal a obrigação de publicar informações sobre os serviços prestados, bem como forma de atendimento e forma de prestação desses serviços, instituindo a carta de serviço ao cidadão” ([ALBANO, 2014](#)).

Outro marco importante seria a publicação da Lei Complementar 131 (conhecida como lei Capiberibe), também em 2009, que “estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências, a fim de determinar a disponibilização, em tempo real, de informações pormenorizadas sobre a execução orçamentária e financeira” ([BRASIL, 2009b](#)). Fica assim estipulada a publicação eletrônica de tais informações ([SANTANA M., 2014](#)).

Por fim, em 2011, é publicada a Lei de Acesso à Informação (LAI). Ainda segundo [Albano \(2014\)](#), esta lei “destina-se a assegurar o direito fundamental de acesso à informação, deve ser observada pela União, Estados, Distrito Federal e Municípios, com o fim de garantir o acesso a informações, conforme previsto na Constituição Federal Brasileira de 1988”. Ela também estipula alguns requisitos para os sítios de disponibilização de dados governamentais, entre eles: “possibilitar a gravação de relatórios em diversos formatos eletrônicos, inclusive abertos e não proprietários, tais como planilhas e texto, de modo

a facilitar a análise das informações” e “possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina” (BRASIL, 2011).

No nível municipal a legislação referente a DGA é composta principalmente pelos decretos 53.623/2012 e 54.779/2014 (COORDENADORIA DE PROMOÇÃO DA INTEGRIDADE, 2013b). O primeiro “[r]egulamenta a Lei Federal nº 12.527, de 18 de novembro de 2011, no âmbito do Poder Executivo, estabelecendo procedimentos e outras providências correlatas para garantir o direito de acesso à informação” (SÃO PAULO, 2012). O segundo decreto, por sua vez, altera a redação de 21 dos 84 artigos do primeiro e institui o Catálogo Municipal de Bases de Dados (SÃO PAULO, 2014).

4.2 Diário Oficial Municipal de São Paulo

“O Diário Oficial publica diariamente os atos oficiais no âmbito do governo municipal, tais como leis, decretos, portarias, editais de licitação, nomeações e exonerações dentre outros. O jornal é dividido em oito seções: Gabinete do Prefeito, Secretarias, Servidores, Concursos, Editais, Licitações, Câmara Municipal e Tribunal de Contas. [...] Todos os dias, o conteúdo do Diário Oficial da Cidade de São Paulo é publicado integralmente no DO.online. São milhões de páginas publicadas, com arquivos desde 2 de dezembro de 1975.” (IMPrensa Oficial, 2010)

A SEMPLA é a responsável na prefeitura municipal de São Paulo pela publicação do DOM. Ela, por sua vez, contrata a Imprensa Oficial (IO), a mesma empresa responsável pela publicação do DO Estadual, para realizar a publicação do DOM.

A IO publica os diários em formato digital na Internet, porém não permite o download integral da base. Além disso, o formulário de busca textual do site, que indexa o DOM desde 2005, não realiza busca por categoria e as páginas do DOM são disponibilizadas no site apenas em formato PDF.

No DOM constam informações que afetam o funcionamento da prefeitura, de empresas, da cidade e os cidadãos em geral, logo, existem diversos grupos interessados em possuir essa base completa ou simplesmente em poderem ter mais liberdade de busca nela.

Apesar da base integral não estar atualmente disponível para download no site da IO, é possível escrever programas para automatizar a extração dos PDFs. Deles pode-se então extrair os textos dos artigos. Porém tal processo é impreciso, uma vez que os PDFs não são estruturados. Tal procedimento é realizado, por exemplo, por empresas que vendem serviços minerando dados na base extraída (RADAR Oficial, 2012; UPLEXIS, 2011).

A equipe da SEMPLA que cuida da publicação do DOM demonstrou diversas vezes uma preocupação com segurança, desconfiando de que uma publicação como dado aberto pudesse ser segura. Afirmaram também que a publicação atual é segura pois utiliza PDF e certificação digital. Apesar disso, eles pareceram desconhecer que arquivos PDF podem ser facilmente editados.

É interessante notar que o site de disponibilização do DOM nem mesmo suporta HTTPS. O site da IO, por sua vez, apesar de possuir uma versão em HTTPS, não encaminha automaticamente o usuário para ela, mesmo em uma tela de cadastro de usuário.

Para fim de comparação pode ser citado o site que disponibiliza os DOs do governo federal dos Estados Unidos, *Federal Register*⁷. Porém, é importante que seja feito o destaque de que a qualidade de tal site é exceção e não a regra em matéria de publicação de DOs pelo mundo.

Primeiramente pode-se notar que a busca permite utilização de diversos critérios, e não apenas palavras-chave e data, como no DOM. Entre esses critérios destaca-se a filtragem por agência publicante, entidades afetadas pela lei e até mesmo a delimitação de uma região citada no texto.

Uma vez construído o filtro de busca é possível navegar pelos resultados na interface do site, própria para um usuário convencional, ou exportar a lista de resultados em um CSV ou JSON, formatos próprios para a utilização por programas. Nota-se que as listas de resultados exportados não contém os artigos em si, apenas os metadados dos mesmos.

Com a construção do filtro de busca é possível obter avisos de novos resultados. Isso pode ser feito via e-mail ou por RSS, sem qualquer cobrança. Sobre esse aspecto, o site da IO disponibiliza funcionalidade significativamente mais simples, mas semelhante, porém são cobrados R\$27/mês por palavra-chave utilizada na busca (IMPrensa Oficial, 2011a). O site do DO da União oferece serviço equivalente ao da IO, sem custo algum ao usuário (IMPrensa Nacional, 2011).

Ainda sobre o *Federal Register*, a visualização dos artigos é feita, por padrão, em HTML, porém há links para versões em XML, JSON, MODS e PDF. Todas, exceto o PDF, são estruturadas, contendo os diversos metadados dos artigos, como por exemplo data de publicação, publicante e resumo da matéria. É importante notar que os artigos publicados são identificados univocamente, e não apenas as páginas, como é feito no site

⁷ <https://www.federalregister.gov>

da IO. A página de cada artigo possui um aviso de que o único documento com valor legal é o em PDF, que é certificado. O site utiliza HTTPS, inclusive para a disponibilização dos artigos em formato XML, JSON e MODS.

Sobre os possíveis impactos negativos advindos da publicação de tais dados pode-se dizer que, como o DO já é um dado disponibilizado, não há preocupação quanto à violação de informações sigilosas. Porém, os dados do DO, muitas vezes informações sobre pessoas, são usados para mineração de dados por empresas de seguros, marketing etc (UPLEXIS, 2011). Melhorar a disponibilização dos DOs poderia facilitar esse uso. Por outro lado, esse tipo de exploração já ocorre e a melhoria da disponibilização permitiria talvez que outros grupos, com menos recursos, como acadêmicos ou ativistas, pudessem analisar a base também. Dessa forma, não são claros os impactos reais que podem vir a ser produzidos.

4.2.1 Publicação de artigos

Segundo orientações da própria IO, para um funcionário de um determinado órgão da prefeitura publicar no DOM, primeiro o texto deve ser escrito e salvo como TXT. O nome do arquivo TXT salvo deve conter um código, chamado de *retranca*, que representa o conteúdo do artigo e qual órgão público que o escreveu. Esse arquivo é então submetido ao Pubnet⁸, sistema da IO para coleta de artigos. Imagens, chamadas de *calhau*, seguem outro processo, sendo salvas como PDF e não como TXT, antes de serem submetidas para publicação (IMPrensa Oficial, 2011b). Por volta das 17h de cada dia útil, o sistema de coleta de artigos interno interrompe o recebimento e uma funcionária da prefeitura cuida, junto à IO, dos trâmites finais para a transferência de todo o material necessário à publicação do dia. A figura 3 representa o fluxo inicial de publicação do DOM.

Uma vez dentro da IO, o material é salvo e disponibilizado integralmente, através do protocolo FTP, de volta à prefeitura. Esse material, constituído de um arquivo ZIP⁹ por dia, é importante para permitir posterior conferência caso, por exemplo, a prefeitura julgue que houve um erro por parte da IO na publicação do DOM. O material segue também para editoração, onde é formatado para publicação. Esta última é feita em papel (jornais impressos) e online, disponibilizando os PDFs no site da IO.

⁸ <https://pubnet.imprensaoficial.com.br/pubnet>

⁹ Formato de arquivo que suporta compressão sem perdas.

Figura 3 – Fluxo de publicação do DOM

Fonte: autoria própria

Os PDFs mantêm a mesma estrutura visual do jornal impresso, bastante distinta da estrutura dos TXTs iniciais. Tal fato dificulta a extração automática do texto, porém facilita a comparação entre a versão impressa e a digital.

4.2.2 Estrutura dos arquivos

Como dito anteriormente, quando os arquivos são salvos para publicação, o nome utilizado para o arquivo deve seguir um padrão (SEMPLA, 2010). Esse padrão é composto por sete letras, que formam a *retranca*, e um número de três algarismos, ex.: **FAFA-ADM.004**. A primeira letra representa a seção do DOM onde o artigo deve ser publicado, sendo que existem sete possibilidades válidas:

- a-) **DESPACHOS**: onde são publicadas as matérias com decisões das autoridades municipais. Ex.: decisões em processos, autorizações, contratos.
- b-) **SERVIDORES**: onde são publicadas as matérias referentes a vida funcional dos servidores. Ex.: exonerações, nomeações, designações, licenças médicas, cursos, contagem de tempo, aposentadoria, férias, etc.
- c-) **CONCURSOS**: onde são publicadas as matérias referentes a concursos públicos. Por ex.: relação de inscritos, local de provas, gabaritos de provas, classificação dos candidatos, etc.
- d-) **EDITAIS**: onde são publicadas as matérias de conteúdo informativo. Por ex.: convocação de munição, demonstrativo de despesas, relação de cargos e salários, balancetes, etc.

e-) LICITAÇÕES: onde são publicadas as matérias sobre certames licitatórios. Por ex.: edital de abertura de convite, de concorrência pública, de tomada de preços, ata de registro de preços, ata de abertura do certame licitatório, classificação dos participantes, adjudicação do objeto, homologação do resultado, etc.

g-) CÂMARA MUNICIPAL: onde são publicados todos os atos da Câmara Municipal de São Paulo, exceto matérias sobre licitações.

h-) TRIBUNAL DE CONTAS: onde são publicados todos os atos do Tribunal de Contas do Município de São Paulo, exceto as matérias sobre licitações.

(SEMPLA, 2010)

A segunda, terceira, quarta e quinta letras da *retranca* representam respectivamente a secretaria, departamento, divisão e seção da administração municipal publicante do artigo. Essas letras são tabeladas em um arquivo chamado *espelho*. A sexta e sétima letras devem ser sempre **DM**, representando o Diário Municipal dentro do sistema Pubnet.

Por fim, o número de três algarismos deve ser usado para se evitar repetição de nomes de arquivos, uma vez que uma mesma entidade da administração pode publicar mais de um artigo para ser colocado na mesma seção do DOM, fazendo com que as sete letras do código sejam idênticas, restando apenas o número para diferenciação. Já sobre o conteúdo do arquivo, como está em formato texto “puro”, há quatro marcações que podem ser utilizadas:

- **Início do título:** ((TITULO))
- **Início do texto:** ((TEXTO))
- **Início de negrito:** ((NG))
- **Fim de negrito:** ((CL))

A informação ((TITULO)) deve anteceder a descrição do título da lauda[...]. Este parâmetro não é obrigatório (e) só deve ser utilizado quando houver uma linha de título.

A informação ((TEXTO)) deve ser colocada após a descrição do título e antes de se iniciar a descrição da lauda. Este parâmetro é obrigatório, ele identifica onde começa a matéria.

Exemplo de um texto pronto para transmissão: ((TITULO))COMISSÃO PERMANENTE DE LICITAÇÃO - CPL-2 ((TEXTO))((NG))1998-0.202.209-8((CL)) - Concorrência para registro de preços ((NG))131/98((CL)) - Fornecimento de fita adesiva de celulose transparente e canetas esferográficas - Decorrido o prazo recursal, deliberou a Comissão: ((NG))DESIGNAR((CL)) o dia 08.02.99 às 11 horas para abertura dos invólucros. ((NG))1998-0.179.332-5((CL)) - DEMAT.1 - Utilização da ata de registro de preços 68/97-DEMAT - Reti-ratificação de despacho - Face à insuficiência de recursos financeiros para consecução da despesa de 1998 ((NG))RETI-RATIFICO((CL)) o despacho exarado sob fl. 30 (publicado no DOM em 13.11.98) para o corrente exercício financeiro.

(SEMPLA, 2010)

De acordo com o aviso publicado no próprio DOM, “os comandos digitados erroneamente serão publicados como parte da matéria”, o que ocorre com certa frequência (IMPrensa Oficial, 2014a).

4.3 Diário Livre

Nesta seção discute-se o processo de criação da ferramenta fruto deste projeto, batizada de Diário Livre (DL). Primeiramente é discutido o levantamento de requisitos, realizado através de cerca de uma dezena de reuniões da equipe, comentários feitos no primeiro evento e literatura relacionada. Uma vez definidos esses requisitos é apresentada a arquitetura adotada para o projeto. E, por fim, a implementação do mesmo, sendo descritos seus detalhes técnicos.

Vale notar que, como a metodologia aplicada neste trabalho se baseia na Pesquisa-Ação, as etapas aqui descritas foram desenvolvidas em parceria com outros grupos. Prioritariamente com os gestores públicos membros da equipe, porém, no caso do levantamento de requisitos, também houve a participação de pessoas externas através de um evento.

4.3.1 Levantamento de Requisitos

O levantamento de requisitos foi realizado através de reuniões com os gestores envolvidos diretamente com o projeto, eventos públicos e literatura relacionada a DGA. Os principais deles estão descritos abaixo, divididos em três categorias. A primeira categoria consiste nos requisitos que já são atendidos pelo site oficial de publicação do DOM, e eram importantes de serem mantidos pelo novo sistema:

- **Atualizações diárias:** o DOM é publicado diariamente e o novo site deveria ser atualizado com os novos dados na mesma frequência.
- **Busca textual rápida:** a base, mesmo depois de processada, inclui mais de 10 gigabytes de texto. Uma ferramenta eficiente de busca textual precisava ser utilizada para obter os resultados em poucos segundos e não prejudicar a usabilidade do site.

A segunda categoria é formada pelos requisitos que são cumpridos pelo site oficial, mas deveriam ser melhorados na nova ferramenta:

- **Filtragem das buscas:** seria importante permitir filtragem dos dados a partir dos metadados disponíveis, como publicante e data de publicação.
- **Notificação aos usuários:** permitir que os usuários sejam avisados quando novas publicações contiverem determinadas palavras chaves. Assim um usuário pode ser avisado quando, por exemplo, seu nome ou algum assunto de seu interesse aparecer em um novo artigo no DOM. É importante notar que, apesar do site atual de disponibilização do DOM possuir essa funcionalidade, é cobrada uma taxa mensal proporcional ao número de palavras usado.

E por último, na terceira categoria, estão os requisitos inexistentes no site oficial, mas que deveriam ser implementados na nova ferramenta:

- **Visualização em texto:** muitos usuários reclamaram da dificuldade para copiar os texto dos PDFs da IO, ou perda de formatação ao colar o texto copiado. O site deveria então retornar os artigos como texto em uma página comum (HTML), facilitando o processo de cópia.
- **Acesso via API:** para permitir o fácil acesso aos dados através de outros aplicativos, era importante que o site disponibilizasse algum tipo de API.
- **Disponibilização integral da base:** permitir o uso da base como um todo, seja para pesquisas ou para aplicativos que exigissem um controle dos dados maior do que o oferecido pela API.
- **URLs únicas para artigos:** para permitir a fácil citação dos artigos, seria importante que cada um fosse acessível através de uma URL única.

O primeiro evento foi significativo para a construção destes requisitos, sendo que alguns dos comentários feitos nele estão representados na lista acima. Mais especificamente, como requisitos levantados nesse evento e que estão contemplados acima, pode-se citar:

- Evitar a perda de formatação decorrente da divisão em colunas (contemplado na visualização em texto).
- Fornecimento de mais metadados (contemplado na filtragem das buscas usando todos os metadados disponíveis).
- Notificações aos usuários.
- URLs únicas para permitir citações.
- Disponibilização em formato não PDF (contemplado na visualização em texto).

Apesar disso, nem todos os requisitos levantados no evento foram contemplados, uma vez que alguns deles demandariam esforço humano além do disponível para esse projeto. Um desses pedidos, feito por uma jornalista, foi o fornecimento de informações para possibilitar a interpretação dos dados disponibilizados, algo que facilitaria a ela, por exemplo, a produção de matérias a partir dos artigos do DOM. Mas considerou-se tal pedido inviável, uma vez que implicaria em analisar e adicionar, manualmente, informações a cada artigo.

Outro pedido foi a produção de resumos para cada artigo, facilitando a leitura pela população. Porém, a produção manual desses resumos seria muito custosa. Não foi avaliada a possibilidade de gerá-los automaticamente. Por fim, um outro pedido foi a capacitação da população para o uso dos dados, uma vez que muitas pessoas não compreendem o que está escrito nos artigos do DOM e qual a implicação deles em suas vidas. Nota-se que esses pedidos, os três relacionados com a interpretação e compreensão dos dados, também não são atendidos pelo site oficial, e estão relacionados com um problema maior de reúso do dados públicos de tamanho e complexidade considerável.

4.3.2 Arquitetura

Uma vez feito o levantamento dos requisitos, foi projetada a ferramenta Diário Livre buscando atendê-los. Sua arquitetura está representada na [figura 4](#). Na camada de extração, os dados são repassados pela administração pública e inseridos na máquina onde o sistema está hospedado, para que o processamento dos dados possa ser realizado.

Devido à política de segurança interna da prefeitura e questões de contrato entre ela e a IO, o sistema não poderia colher os dados de dentro da rede da prefeitura. Ao contrário, os funcionários da prefeitura é que precisariam fazer o repasse dos mesmos para o sistema de publicação. E como, em um primeiro momento, essa transferência tinha que ser realizada manualmente pelos próprios funcionários da prefeitura, a ferramenta utilizada nessa etapa precisava ser usável por não desenvolvedores.

Durante a transformação, os dados são padronizados e preparados para a indexação e publicação. Os dados fornecidos pela prefeitura estão compactados e incluem diversos formatos de arquivos. Esses precisam ser tratados e padronizados, identificando as codificações do conteúdo dos arquivos em texto e os metadados implícitos nos nomes desses arquivos, para só então ser possível a indexação dos dados.

Figura 4 – Arquitetura do Diário Livre em camadas e o fluxo da informação.

Fonte: autoria própria

Na terceira etapa, carregamento, os dados são indexados para permitir as buscas. Devido ao tamanho e características da base, era necessária uma ferramenta capaz de indexar grandes quantidades de texto, realização de buscas textuais em tempo hábil e filtragem pelos metadados.

Na quarta etapa os dados são finalmente publicados em diversos formatos. Estes formatos compreendem a visualização web convencional, o download integral da base e o acesso através de API, buscando maximizar o número de usos possíveis. As URLs usadas na visualização web convencional ou na API permitem acessar os artigos individualmente, e não por página, como ocorre no sistema oficial da IO.

4.3.3 Implementação

Nesta subseção se discute de forma mais aprofundada os elementos constituintes do processo de criação da ferramenta. Tendo em vista o debate sobre as atividades e o custo necessário para o processo de *Rawification*, como colocado por [Davies e Frank \(2013\)](#), nas subseções seguintes será explanado o estado inicial da base e as atividades que foram necessárias para sua disponibilização.

Extração

Uma vez que os artigos são enviados para publicação através do sistema Pubnet, a IO guarda a versão em TXT em uma pasta que pode ser acessada pela SEMPLA via FTP. Porém, os arquivos são deletados dessa pasta depois de um prazo que costuma variar entre uma semana e um mês. Logo, a base completa do DOM em TXT não poderia ser extraída.

Apesar disso, por acaso, uma funcionária da SEMPLA guardou em seu computador pessoal os arquivos desde 2003, possivelmente para se proteger de possíveis falhas que ocorressem no sistema. A funcionária contribuiu com o projeto fornecendo a base que possuía, além de acesso FTP para captura dos novos DOs em TXT que fossem publicados.

Limpeza

A base copiada possuía arquivos organizados em pastas na seguinte estrutura:

`<ano>/<mês>/<dia>.zip`

Havia cerca de um arquivo ZIP por dia útil desde janeiro de 2003. Dentro de cada arquivo ZIP havia algumas centenas de arquivos, contendo cada um o texto de um ou mais artigos publicados naquele dia. Os tamanhos dos ZIPs variavam entre 1 megabytes e 3 megabytes. A base como um todo, compactada, possuía cerca de 16 gigabytes.

O primeiro passo necessário foi a extração dos mais de 2500 arquivos ZIP distribuídos em mais de 120 pastas. Por ser uma tarefa trabalhosa, foi escrito um *script*¹⁰ em Python¹¹ para tal. O *script* também adiciona ao nome de cada arquivo extraído o ano, mês e dia, informações obtidas a partir do nome das pastas e do ZIP de onde o arquivo foi extraído. O dia é acrescido de 1 para igualar à data de publicação pela IO, uma vez que, aparentemente, o nome dos ZIPs contém a data em que a SEMPLA submeteu o arquivo para publicação, sendo que eles são sempre publicados no dia seguinte.

Além disso, o *script* tenta detectar se o arquivo a ser extraído é um TXT, um PDF, um DOC ou outro formato de arquivo. Essa verificação tornou-se necessária pois muitos arquivos dentro dos ZIPs, apesar de possuírem extensão “.TXT” ou “.txt”, eram na verdade arquivos PDF ou DOC, o que gerava problemas no processamento futuro da base. Essa verificação de tipo de arquivo se deu então utilizando a saída do comando de Linux **file**¹², com parâmetros **-bi**, como pode-se ver no código do próprio *script*.

Uma vez descompactada a base, foram obtidos 1.430.648 arquivos considerados TXT, totalizando cerca de 13 gigabytes. Sobre os demais, PDF (60.841 arquivos), DOC (230 arquivos) e outros formatos (14.647 arquivos) totalizavam 10 gigabytes. Analisando os arquivos não-TXT, observou-se que, em sua maioria, consistiam em PDFs contendo imagens

¹⁰ <https://github.com/andresrmr/mestrado/blob/master/descompactador.py>

¹¹ <https://www.python.org>

¹² <http://linux.die.net/man/1/file>

(os *calhaus* discutidos anteriormente). Porém, há vários arquivos de formatos e conteúdos não identificados. Visando produzir uma primeira versão do sistema que permitisse, pelo menos acesso e busca aos arquivos TXT, os demais arquivos foram desconsiderados.

O passo seguinte consistiu em utilizar a *retranca*, código no nome dos arquivos, para identificar o publicante e conteúdo dos artigos. Para executar esse processo foram escritos outros dois *scripts*^{13,14}. No segundo pode-se inclusive ver a tabela de *retrancas* utilizada, chamada de *espelho*, extraída de um arquivo XLS fornecido pela SEMPLA.

Um outro problema que surgiu ao processar os TXTs foi o fato deles estarem em diversas codificações diferentes, o que causava erros na hora de tentar converter os conteúdos para UTF-8. Como o formato TXT não contém informações sobre a codificação usada para o conteúdo, não há como determiná-la com certeza. Foram utilizados três programas para tentar descobrir a codificação de cada arquivo: *Chardet*¹⁵, *Uchardet*¹⁶ e *File* (esse último já citado anteriormente). O processo de identificação de codificação utiliza algumas heurísticas, mas geralmente não se pode ter certeza se ela foi identificada corretamente. Algumas das codificações que os programas indicaram encontrar foram: UTF-8, UTF-16, ISO-8859-1, WINDOWS-1253, IBM855, X-MAC-CYRILLIC, TIS-620 e EUC-KR. Algumas das codificações indicadas aparentemente estavam certas, pois uma vez feita a conversão, o texto do arquivo não apresentava caracteres defeituosos. Porém, nenhum dos três programas conseguiu identificar a codificação de todos os artigos, sendo que frequentemente discordavam quanto à codificação dos mesmos.

Por outro lado, verificou-se que a própria IO também não conseguiu resolver a codificação de todos os artigos, uma vez que são encontradas páginas do DOM com caracteres errados, como por exemplo, “OCUPAR IMãVEL COM USO NPO RESIDENCIAL SEM A LICEN A DE FUNCIONAMENTO” (*IMPrensa Oficial*, 2014b). Tendo isso em vista, os artigos mais problemáticos quanto codificação, foram descartados da base a ser disponibilizada. Todos os demais foram convertidos para UTF-8.

O último *script* citado constrói, com o conteúdo dos artigos e respectivos metadados, um CSV com as seguinte colunas:

- **ID:** campo criado para permitir identificar unicamente cada artigo.
- **Data:** data de publicação do artigo.

¹³ <https://github.com/andresmrm/mestrado/blob/master/classificador.py>

¹⁴ <https://github.com/andresmrm/mestrado/blob/master/retrancas.py>

¹⁵ <https://pypi.python.org/pypi/chardet>

¹⁶ <https://code.google.com/p/uchardet>

- **Retranca:** código de 5 letras inicial do qual foi extraído o tipo de conteúdo e publicante do artigo.
- **Tipo do Conteúdo:** seção do DOM em que o artigo deve ter sido publicado.
- **Secretaria:** secretaria publicante.
- **Orgão:** órgão publicante.
- **Texto:** o artigo em si.

Como existem artigos com mais de um megabyte de tamanho, existem células da coluna **Texto** igualmente com mais de um megabyte de tamanho, o que não é convencional em um arquivo CSV. Levando-se em consideração os artigos publicados entre janeiro de 2003 a abril de 2014, quando a primeira versão da base foi coletada para processamento, o CSV produzido teve cerca de 9 gigabytes de tamanho e 1.412.232 artigos (cerca de 94% de todos os arquivos extraídos da base original).

Além dos problemas já citados, é importante destacar que o *espelho*, que identifica os publicantes, foi alterado ao longo dos últimos anos. Isso não foi levado em consideração no preenchimento dos metadados dos artigos, uma vez que ainda não se conseguiu o histórico de alteração do *espelho*.

Publicação

Uma vez que a base estava minimamente limpa e padronizada, faltava-se construir um sistema para disponibilizá-la. Tendo em vista o tamanho da mesma, era importante se preocupar com a velocidade de busca textual da ferramenta que fosse ser utilizada¹⁷.

Duas ferramentas livres foram cogitadas inicialmente para a disponibilização: o CKAN¹⁸ e o Solr¹⁹. O primeiro é o sistema de disponibilização de DGA utilizado pelos principais portais de dados abertos do mundo, o que inclui o portal nacional brasileiro, estadunidense e inglês (KUČERA; CHLAPEK; MYNARZ, 2012). O segundo, Apache Solr, é uma tecnologia NoSQL, ou seja, não relacional, que é otimizada para resolver uma classe específica de problemas para tipos específicos de dados. Solr é escalável e otimizado para grandes volumes de dados (dados centrados em texto), e retorna resultados ordenados por

¹⁷ Antes da base passar por esse processo, um dos gestores da COPI-CGM tentou construir um sistema simples para busca em parte da base, mesmo sem a extração dos metadados. Porém, o tamanho dela fazia com que uma busca levasse minutos para ser completada.

¹⁸ <http://ckan.org>

¹⁹ <https://lucene.apache.org/solr>

relevância. Ele é construído usando o [Lucene](http://lucene.apache.org)²⁰ como motor de busca, e é desenvolvido pela própria comunidade do Lucene (SMILEY; PUGH, 2009; GRAINGER; POTTER; SEELEY, 2014).

Tendo em vista a falta de literatura a respeito de quais são as soluções e desafios técnicos relacionados à disponibilização de bases de dados governamentais, as subseções seguintes documentam como a questão foi tratada neste projeto.

CKAN

Observando que o CKAN é desenvolvido pensando-se na disponibilização de DGA, optou-se primeiramente por ele. Essa ferramenta possui diversas funcionalidades, como catálogo de bases de dados, repositório de bases para download, visualização das bases colocadas no repositório e API para manipulação das mesmas.

Na instalação do CKAN percebeu-se que ele próprio utiliza o Solr para a busca interna entre bases de dados. Porém, através de conversas na lista de e-mails dos desenvolvedores, descobriu-se que, para a busca dentro de uma base de dados, é utilizada a ferramenta [PostgreSQL](http://www.postgresql.org)²¹. Essa última, apesar de ser um gerenciador de base de dados bastante robusto, não foi desenvolvido com a busca textual como foco.

O CKAN faz essas escolhas pois foi pensado para situações onde pode-se até possuir um grande número de bases de dados a serem disponibilizadas, porém cada uma das bases possui no máximo algumas dezenas de megabytes. No caso abordado aqui há apenas uma base de dados, mas que possui mais de 9 gigabytes.

Uma alternativa seria o mapeamento de cada artigo como uma base de dados, porém o conteúdo dos artigos teria de ser inserido no CKAN como metadado do próprio artigo, já que a busca entre base de dados, feita via Solr, só analisa os metadados e não os conteúdos de cada base de dados. Isso aumentaria o tamanho total de espaço utilizado pelo sistema, já que os conteúdos estariam duplicados, além de ser um uso não convencional do CKAN, o que poderia gerar outros problemas. Preferiu-se manter o mapeamento inicial, uma base contendo todos os artigos, e testar a capacidade de busca textual do PostgreSQL nessa situação.

²⁰ <http://lucene.apache.org>

²¹ <http://www.postgresql.org>

Uma vez instalado o CKAN em um servidor²², a base do DOM foi submetida e processada por ele. Porém, quando o teste foi feito, as buscas via API do CKAN não chegavam nem mesmo a serem completadas, uma vez que levavam mais de um minuto e eram interrompidas pelo próprio CKAN.

Pesquisando-se a respeito descobriu-se que o PostgreSQL poderia ganhar muito mais velocidade em buscas textuais do tipo se fosse configurado para indexar a base. Porém os comandos de indexação não conseguiram ser processados. Aparentemente a indexação feita pelo PostgreSQL possui algumas restrições quanto ao número de palavras únicas dentro de um mesmo campo da base de dados. Todavia, na base em questão existem campos com, por exemplo, listas de nomes de pessoas inscritas para um concurso público. Tais listas podem conter dezenas de milhares de indivíduos, sendo que nomes, sobrenomes e números de inscrição são elementos com baixa quantidade de repetições se comparados a um texto comum. Dessa forma, tratava-se de um caso excepcional para o qual o PostgreSQL não foi projetado.

Solr

Paralelamente à tentativa de uso do CKAN, Rafael Santana (um aluno próximo ao Colab-USP) estava estudando como TCC (orientado por José de Jesús Pérez-Alcazar) a possibilidade de uso do Solr para o mesmo fim. Uma vez que ele obteve sucesso na indexação de uma parte da base, utilizando uma instalação local de Solr, a tentativa via CKAN foi suspensa e optou-se pelo uso do Solr. Configurou-se então o Solr no servidor citado anteriormente. O Solr já se encontrava instalado naquele servidor, uma vez que é uma dependência para a instalação do CKAN.

Para facilitar a indexação pelo Solr, a base do DOM foi dividida, obtendo-se um CSV por dia. Tal divisão também facilitaria a adição de novos arquivos conforme fossem publicados. O Solr foi então utilizado para indexar os 2.704 CSVs. Por fim, a busca textual estava finalmente funcionando e levando menos de 1 segundo por requisição.

A interface do servidor de Solr, apesar de poder ser visualizada por um navegador, não é adequada para usuários comuns, mas apenas para administração. Tornou-se então necessária a construção de um site que usasse o Solr no *backend* para buscas. Optou-se

²² O servidor possuía cerca de 200 gigabytes livres no disco rígido, 8 gigabytes de memória e processador *Intel(R) Xeon(R) CPU E7 - 2870 @ 2.40GHz*

por uma ferramenta chamada [BlackLight](http://projectblacklight.org)²³, uma vez que ela já oferecia uma interface *amigável* para as buscas.

BlackLight

BlackLight é uma ferramenta feita em [Ruby on Rails](http://rubyonrails.org)²⁴ para servir de interface ao Solr. Por ser feita utilizando Rails, um arcabouço web bastante popular e versátil, a ferramenta pode ser moldada de acordo com as necessidades do site final desejado (DUPLAIN; BALSER; RADZIWILL, 2010; SADLER, 2009).

Essa ferramenta foi instalada no mesmo servidor citado anteriormente e configurada para utilizar a instância do Solr na mesma máquina. A aparência do sistema também foi personalizada para fornecer uma identidade visual própria ao site. Além disso, foi adicionada uma página onde é possível baixar as bases completas dos diários em dois formatos: em estado *bruto* ou depois de serem convertidos em CSV.

Atualização dos dados

Para que o processo de atualização diário fosse viável, ele precisava ser automatizado. Com esse propósito foram utilizados os scripts em Python, discutidos anteriormente, para a extração dos arquivos dos ZIPs, *limpeza* dos dados, organização dos mesmos em um CSV por dia, e indexação pelo Solr. Tais scripts cuidaram da automatização do processo uma vez que os ZIPs estavam no servidor do Colab-USP. Restava automatizar também a transferência dos ZIPs do FTP da IO para esse servidor.

Porém, por questão contratual entre a prefeitura e a IO, o Colab-USP não podia ter acesso à senha do FTP. Como forma temporária de permitir a transferência dos arquivos pelos próprios funcionários da COPI-CGM, mesmo que manualmente, foi hospedada uma instância do programa [OwnCloud](https://owncloud.org)²⁵ no servidor do Colab-USP. Esse software livre cumpre o papel de um site para compartilhamento de arquivos. Uma vez fornecido usuário e senha para os funcionários da COPI-CGM, eles podiam entrar no sistema e submeter os arquivos ZIPs. Como o OwnCloud salva os arquivos submetidos em uma pasta no servidor

²³ <http://projectblacklight.org>

²⁴ <http://rubyonrails.org>

²⁵ <https://owncloud.org>

hospedeiro, os scripts, rodando na mesma máquina, poderiam acessar esses arquivos e cuidar do restante do processo.

Porém, os funcionários da COPI-CGM ainda precisavam submeter arquivos manualmente, o que gerava atrasos, e algumas reclamações em listas de e-mail onde o DL foi divulgado. Um usuário entusiasta do DL ofereceu ajuda para desenvolver um script de automatização de FTP em *Windows*. Rodando o script em um servidor da COPI-CGM, ele poderia ter acesso à senha do FTP da IO e automatizar a transferência para o servidor do Colab-USP. Mas a ajuda acabou não se concretizando.

É importante notar aqui que a equipe da COPI-CGM não dispunha de uma equipe técnica capaz de resolver o problema. Além disso, o fato de só disporem de máquinas *Windows*, dificultava a ajuda por parte do programador do Colab-USP que havia automatizado o restante do processo (o autor deste texto), acostumado ao desenvolvimento em Linux.

Após cerca de um semestre, a COPI-CGM conseguiu ajuda do LabProdAm. Este é um laboratório de inovação da PRODAM que havia sido recentemente criado. O desenvolvedor desse grupo realizou a automatização restante através da interface [WebDAV](#)²⁶ da instância do OwnCloud hospedada no servidor do Colab-USP. A atualização passou então a ser diária.

Disponibilização Integral da Base

Além da disponibilização via interface *web* convencional, buscou-se disponibilizar também o download integral da base. Tanto dos arquivos *brutos*, no formato original repassado pela SEMPLA, como dos arquivos tratados e organizados em CSVs. Apesar do segundo formato possivelmente facilitar a manipulação da base, uma vez que já possui certa estruturação, ele não continha os arquivos PDF e parte dos TXTs, os que não conseguiram ser processados para a indexação.

Tendo em vista que os arquivos referentes aos anos de 2003 a 2013 já estavam consolidados e não sofreriam atualizações, foram criados dois arquivos compactados contendo:

- Os arquivos referentes a esses anos em formato original, totalizando 11,82 gigabytes.

²⁶ <http://www.webdav.org>

- Os arquivos referentes a esses anos em formato CSV, totalizando 0,74 gigabytes.

O segundo é significativamente menor que o primeiro principalmente por não conter os PDFs, já que a grande maioria deles são imagens. Os arquivos foram compactados utilizando o algoritmo *7z*²⁷ no maior nível de compressão suportado.

Uma vez que se tratava de arquivos razoavelmente grandes (principalmente o primeiro), buscou-se disponibilizá-los através do protocolo *Torrent*²⁸. Esse permitiria escalar mais facilmente a distribuição dos arquivos. Porém, após diversas dificuldades no oferecimento dos arquivos, acabou-se por adotar a abordagem mais simples: download comum por HTTP. Aparentemente, via *Torrent*, a base chegou a ser baixada apenas uma vez.

Para o ano atual, cujo arquivo ainda estava sendo consolidado e precisava ser atualizado diariamente, optou-se diretamente pelo oferecimento via HTTP, uma vez que o protocolo *Torrent* não suporta atualização dos arquivos transferidos. Dessa forma os novos dados vão sendo diariamente e automaticamente incluídos nos arquivos *7z*, que vão aumentando de tamanho ao longo do ano e ficam disponíveis para download.

Ao final do ano de 2014, o arquivo com os dados *brutos* somava cerca de 2 gigabytes, enquanto o de CSVs possuía cerca de 50 megabytes. A partir do início de 2015 os dados começaram a ser totalizados em novos arquivos.

Para oferecer algum tipo de verificação de integridade dos dados baixados, o site disponibiliza uma lista com as *hashes* *SHA1*²⁹ referentes a cada arquivo. Tal arquivo é atualizado diariamente e automaticamente conforme novos dados são inseridos no arquivo do ano atual.

²⁷ <https://en.wikipedia.org/wiki/7z>

²⁸ <https://pt.wikipedia.org/wiki/BitTorrent>

²⁹ <https://en.wikipedia.org/wiki/SHA-1>

5 Discussão e Resultados

Neste capítulo são apresentados e debatidos os frutos deste trabalho. Sobre o processo como um todo, observando-se a partir da perspectiva das recomendações sistematizadas na [Tabela 1](#), pode-se dizer que diversas delas foram cumpridas (porém não todas):

- O projeto buscou trabalhar os medos dos gestores públicos, através de eventos e articulações internas.
- Desenvolveu um protótipo.
- Adotou licenças abertas para o código e dados.
- Utilizou software livre.
- Publicou os dados em formatos abertos.
- Foi desenvolvido através de uma parceria com a academia.

Sobre os pontos não citados a cima:

- A prefeitura realizou um catálogo de dados, mas não como parte desta parceria.
- A escolha dos dados, ao contrário do recomendado por alguns documentos, não foi feita com consulta pública, mas sim a partir da disponibilidade dos mesmos e da articulação política entre as áreas da administração pública municipal envolvidas.

Nas seções seguintes são abordados em maior profundidade os resultados obtidos relacionados a cada um dos objetivos iniciais. Primeiramente, a [seção 5.1](#) trata do que foi alcançado com a ferramenta desenvolvida, e faz um balanço crítico da mesma. Em segundo, na [seção 5.2](#), são discutidos os temores, anseios e obstáculos encontrados no processo, especialmente sob a perspectiva da administração pública. Em seguida, relacionada ao terceiro objetivo inicial, na [seção 5.3](#) o projeto é avaliado sob a ótica da participação proposta para o mesmo. A [seção 5.4](#), por sua vez, busca avaliar como está se dando, e pode vir a se dar, a apropriação do projeto por atores externos a ele, uma preocupação transversal deste trabalho, mas não explicitamente um objetivo do mesmo.

5.1 Ferramenta

No momento de escrita deste texto o [site resultante](#)¹ e o [código fonte](#)² estão disponíveis online, estando o último licenciado com uma licença aberta, a *GNU Affero General Public License*³. Os dados disponíveis também estão licenciados com uma licença aberta, a *Creative Commons 4.0*⁴. Eles estão em formatos não proprietários e processáveis por máquina. Cada artigo, em cada formato, possui sua própria URI.

Tendo em vista todos esses fatores, pode-se dizer que são dados abertos dignos de 4 estrelas segundo o modelo de [Berners-Lee \(2010\)](#). O projeto não teve como plano atingir a quinta estrela do modelo, pelo menos não para essa primeira fase. Nesse aspecto pode-se dizer que foi seguida a recomendação de um dos participantes do primeiro evento, ao afirmar que não se buscasse, logo no primeiro momento, atingir “o melhor dos mundos”.

A [figura 5](#) representa o fluxo atual de publicação. Nessa, diferentemente da [figura 3](#), está incluído o DL. O fluxo antigo entre a prefeitura e a IO continua como antes, porém os dados são repassados diariamente e automaticamente para o Colab-USP pela SEMPLA, atendendo ao requisito de atualizações em tempo real. Esses dados são então descompactados, os TXTs são convertidos para CSVs e os PDFs descartados. Os CSVs são indexados pelo *Solr*, que permite buscas através da ferramenta *BlackLight*. Tendo em vista a eficiência de busca textual do *Solr*, outro requisito é atendido.

O *BlackLight* funciona como um site convencional e possibilita a visualização das buscas e artigos em HTML, XML, JSON, RSS e Atom, atendendo aos requisitos de visualização textual e através de API. Dessa forma é possível tanto para um usuário convencional, como para um programa acessando através de API, visualizar e buscar artigos. Os dois últimos formatos, RSS e Atom, possibilitam atender ao requisito de notificações quando um determinado termo for encontrado em um novo artigo publicado.

Pela maneira como o *BlackLight* foi utilizado, cada artigo possui um URL próprio para referenciá-lo, atendendo um dos requisitos do projeto. A mesma ferramenta também permite filtragem dos resultados de acordo com seus metadados, satisfazendo mais um requisito.

¹ <http://devcolab.each.usp.br/do>

² <https://github.com/rafaeusantana/docsolr>

³ <https://gnu.org/licenses/agpl.html>

⁴ <https://creativecommons.org/licenses/by/4.0/>

Figura 5 – Fluxo de publicação do DOM e Diário Livre

Fonte: autoria própria

Figura 6 – Um exemplo de tela do atual DL

The screenshot shows the Diário Livre website interface. At the top, there's a navigation bar with links: Favoritos, Histórico, Baixar, Questionário, and Acessar. Below this is a search bar with the text "Busca..." and a "Buscar" button. On the left, there's a "Filtre sua busca" section with three main categories: Secretaria, Orgão, and Tipo de Conteúdo. Under Tipo de Conteúdo, there's a list of content types with their respective counts: LICITAÇÕES (334,037), DESPACHO (175,658), SERVIDORES (130,143), EDITAIS (88,777), CÂMARA MUNICIPAL (8,808), CONCURSOS (6,329), and TRIBUNAL DE CONTAS (5,343). Below this, there's a "Data de Publicação" section with input fields for "Data Inicial" and "Data Final", a "Filtrar" button, and a list of date ranges with counts: Menos de 1 semana (3,240), Menos de 1 mês (13,763), Menos de 1 ano (154,276), Menos de 5 anos (749,737, selected), and Mais de 5 anos (645). On the right, there's a "Sua busca por:" section showing "Data de Publicação" and "Menos de 5 anos". Below this, there's an "Exportar Página:" section with options: JSON, XML, Atom, RSS, and ?. The main content area shows three search results, each with a "Favorito" button. Result 1: Data: 11/04/2015, Assunto: Subprefeitura Penha - GABINETE DO SUBPREFEITO - LICITAÇÕES, Texto: ATA DA LICITAÇÃO Pregão Eletrônico N° 3/2015 - Processo 2015-0.056.212-9 Objeto: Aquisição de Concreto Pré Misturado a Seco. Result 2: Data: 11/04/2015, Assunto: SAÚDE - GABINETE DO SECRETÁRIO - LICITAÇÕES, Texto: PROCESSO: 2014-0.007.189-1 DIVISÃO TÉCNICA DE SUPRIMENTOS, SMS-3 AUTORIZAÇÃO DE ABERTURA DE PROCEDIMENTO LICITATÓRIO DESPACHO DO DIRETOR I – Em face da competência delegada pela Portaria nº 890/2013-SMS.G, AUTORIZO a abertura de procedimento licitatório objetivando a AQUISIÇÃO DE PROJETO MULTIMÍDIA, a pedido de ESCOLA MUNICIPAL DA SAÚDE - EMS, com fundamento na Lei M. Result 3: Data: 11/04/2015, Assunto: CULTURA - EDUCAÇÃO - LICITAÇÕES, Texto: DEPARTAMENTO DE ALIMENTAÇÃO ESCOLAR PA. At the bottom, there's a banner that says "Este site faz parte de um projeto experimental. Por favor, clique aqui e responda nosso questionário!" with a close button.

Fonte: autoria própria

Os CSVs são também compactados e disponibilizados no site para o download integral da base, atendendo a outro requisito inicial. São disponibilizadas *hashes* relativas aos arquivos da base, permitindo que a integridade dos mesmos seja verificada após download. Essa foi uma recomendação feita por um participante do primeiro evento. A [figura 6](#) exemplifica uma tela do DL atual.

Sobre o reuso do DL em outras iniciativas de dados abertos, é possível afirmar que tanto o *Solr* como o *BlackLight* são ferramentas genéricas para a indexação, filtragem e exibição de informação textual. Logo essa parte do processo de publicação do DL poderia

ser reaproveitada praticamente por completo em outras iniciativas de abertura de dados textuais. Um exemplo disso é o fato do código do DL ter sido usado como base para a construção de outra ferramenta, a [Pedido Respondido](#)⁵, que publica os pedidos de informação respondidos pela prefeitura de São Paulo. Por outro lado, geralmente será necessário um esforço de adaptação dos arquivos que se deseja publicar para um formato indexável pelo *Solr*. No caso do DOM de São Paulo o formato original era um conjunto de arquivos TXT, que foram então convertidos através de um *script* para um conjunto de arquivos CSV indexáveis pelo *Solr*. No caso de outra prefeitura que desejasse utilizar o DL para publicar seu DO, muito provavelmente o formato original dos arquivos seria diferente, requerendo alterações no *script* de conversão.

5.1.1 Estatísticas de Acesso

Para coletar estatísticas de acesso ao DL foi utilizado o programa *Piwik*⁶. Este é um software livre que utiliza *cookies* para identificar visitantes humanos. Dessa forma acessos automáticos, como a indexação por máquinas de busca, não são contabilizados. Optou-se pelo Piwik por ser software livre, respeitar usuários que não desejam ser rastreados⁷ e não colocar os dados sob o domínio de uma terceira organização, uma vez que eles ficam no servidor local. Sendo assim, foi instalada uma instância do *Piwik* no mesmo servidor hospedeiro do DL.

A partir dos dados coletados foram construídos os gráficos a seguir. Para a correta compreensão dos mesmos, alguns termos precisam ser esclarecidos:

Visitantes únicos: Visitantes que acessaram o site em um determinado período de tempo. Por *acessar o site* se considera abrir qualquer página do mesmo. Busca-se atestar a unicidade de um visitante através do *cookie* colocado no navegador, e que o identificaria unicamente.

Visitantes de retorno: São visitantes que já acessaram o site anteriormente, há mais de 30 minutos atrás. Caso o primeiro e único acesso de um visitante tenha sido há menos de 30 minutos atrás, esse acesso ainda será contabilizado como uma única

⁵ <http://pedidorespondido.info>

⁶ <https://piwik.org>

⁷ <http://donottrack.us>

visita contínua, não caracterizando um *retorno* ao site. Nesse caso será um visitante único, mas não de retorno.

Taxa de saída: Porcentagem dos visitantes que saíram após abrirem uma única página do site. Uma alta taxa de saída pode indicar que o visitante não se interessou pelo site, ou não conseguiu visualizá-lo corretamente. Mas também pode indicar que a informação que buscava já se encontrava na primeira página que abriu, não requerendo mais interações.

Pesquisas: Buscas realizadas através da interface do site.

Ações: Interações entre o visitante e o site, geralmente o número de páginas que o visitante abriu em uma mesma visita. Para que componham a mesma visita, as ações não podem estar distanciadas a mais de 30 minutos entre si.

A [figura 7](#) mostra o número de visitantes por dia. Pode-se notar primeiramente uma oscilação de alta frequência ao longo do período analisado. Essa oscilação, como pode ser visto na [figura 8](#), é devida à variação no número de visitantes ao longo de uma mesma semana. Sendo que os finais de semana apresentam o menor número de visitas.

Na [figura 9](#) pode-se observar novamente os dois picos descritos anteriormente. Porém, como nesse gráfico cada ponto representa uma semana, a oscilação de maior frequência foi eliminada. É possível ver quedas no número de visitas no fim de dezembro e na penúltima semana de fevereiro, ambas provavelmente relacionadas aos feriados. A curva revela também um crescente número de visitantes ao longo do tempo, que atualmente, ao todo, já ultrapassou 100 mil visitantes únicos.

Já sobre os dispositivos utilizados para acessar o site, como pode ser visto na [figura 10](#), a grande maioria dos acessos é feita através de *desktops* ou *notebooks*, seguido de longe, por *smartphones* e *tablets*.

Com respeito à forma pela qual os visitantes ingressam no site, a [figura 11](#) evidencia a predominância de sistemas de buscas, como *Google* e *Yahoo*. Esse grupo é seguido pela entrada direta (quando, por exemplo, um usuário acessa o site digitando seu endereço na barra do navegador, copiando-o de um e-mail, ou a partir dos favoritos de seu navegador). A situação menos frequente foi acesso via outros sites, como notícias sobre o sistema.

Porém, quando avaliado sob a ótica do número de ações por visita, a lógica se inverte, como pode ser visto na [figura 12](#).

Figura 7 – Número de visitantes únicos por dia. A oscilação de maior frequência é devido à redução no número de visitas nos fins de semana.

Fonte: autoria própria

Figura 8 – Média de visitantes únicos por dia da semana.

Fonte: autoria própria

Figura 9 – Número total de visitantes e pesquisas por semana. Contam-se aí as pesquisas feitas pela interface interna do site.

Fonte: autoria própria

Figura 10 – Comparação por dispositivo da quantidade total de visitantes e número de visitantes que saíram após visualizar apenas uma página.

Fonte: autoria própria

Figura 11 – Quantidade total de visitantes por tipo de referência de entrada.

Fonte: autoria própria

Figura 12 – Média de ações por visita por tipo de referência de entrada.

Fonte: autoria própria

Esses dados corroboram com a hipótese de que visitantes provenientes de notícias tenderiam a explorar mais o site. Já aqueles que simplesmente o encontraram em uma busca na Internet, teriam um contato mais rápido com o mesmo.

Como dito anteriormente, nos gráficos não estão inclusas visitas automáticas feitas, por exemplo, por motores de busca. Porém, estas começaram a ocorrer antes mesmo do lançamento do sistema.

Sobre os termos buscados por esses visitantes, na maioria das vezes, são nomes de pessoas, empresas, ou identificadores oficiais para eles (CPF e CNPJ). Esses são termos raramente encontrados fora de um DO, fazendo com que o DL esteja entre os poucos resultados retornados para buscas do tipo.

5.1.2 Questionário Online

Buscando permitir uma via permanente de retorno por parte dos usuários do sistema, nele foi disponibilizado um questionário online. Este conteve perguntas sobre a acessibilidade do sistema, funcionalidades, forma de uso da ferramenta e espaços para comentários gerais. Além disso, caso o respondente desejasse, ele também poderia prover um meio para contatá-lo, de forma que seus possíveis questionamentos fossem respondidos.

Ressalta-se que várias pessoas forneceram contatos e foram contatadas posteriormente, pela equipe do projeto, sobre suas dúvidas e sugestões. Sendo assim, o questionário adotou um carácter pouco comum para esse tipo de instrumento, servindo não apenas para coleta de dados, como também para iniciação de diálogo.

Para a viabilização do questionário online foi utilizado o programa *Limesurvey*⁸. Este é um software livre bastante flexível e próprio para a disponibilização de questionários online. Ele foi adotado por permitir um maior controle dos dados e evitar colocá-los nos servidores de terceiros como ocorre, por exemplo, com o *Google Forms*. Sendo assim, foi instalada uma instância do *Limesurvey* no mesmo servidor hospedeiro do DL.

Foram obtidos 105 questionários completos (quando o usuário prossegue até o fim do questionário e o submete) e 87 questionários incompletos (quando o questionário é abandonado antes da submissão). Para os gráficos foram contabilizadas apenas os questionários completos (54% do total). Nenhuma pergunta era obrigatória de ser respondida.

⁸ <https://www.limesurvey.org>

Figura 13 – Número de questionários online respondidos ao longo do tempo. As cores diferenciam entre os questionários que foram preenchidos até o fim, e aqueles que, por algum motivo, não foram completados. Para as análises foram utilizados apenas os questionários completados.

Fonte: autoria própria

Na [figura 13](#) pode-se observar novamente os dois picos de acesso referentes ao lançamento do site, no fim de outubro, e ao e-mail de divulgação interna da prefeitura, no meio de dezembro. Comparando-se com a [figura 7](#) (página 81), percebe-se que a grande maioria dos visitantes não respondeu ao questionário, apesar dele estar anunciado em todas as páginas do DL. O momento de maior porcentagem de respostas completas por visita foi no segundo pico, atingindo cerca de 5%.

Um indício de que a linguagem utilizada nos enunciados não foi completamente adequada.

Das 7 pessoas que atestaram dificuldades com o DL (cerca de 7% do total), como visto na [figura 14](#), 4 afirmaram não ter localizado algum elemento da interface, 2 tiveram problema com o certificado de segurança e 1 não especificou o motivo. Dessas 7, apenas 2 (2%) afirmaram não ter tido problemas com o site oficial da IO. Comparando-se com a [figura 15](#), 39 pessoas (37%) tiveram dificuldades com o site da IO, e dessas, 33 não tiveram dificuldade com o DL (31% do total). Entre os problemas apontados com o site da IO está

Figura 14 – Respostas para a pergunta: “**Você teve dificuldade para utilizar o site do Diário Oficial Experimental?**”

Fonte: autoria própria

Figura 15 – Respostas para a pergunta: “**Você já teve dificuldade para utilizar o site do Diário Oficial da Imprensa Oficial?**”

Fonte: autoria própria

a dificuldade para utilizar a busca, procurar termos dentro dos PDFs, copiar texto dos PDFs, imprimir apenas uma parte dos textos, e fonte pequena dificultando a leitura. Por fim, conforme a [figura 16](#), 65 respondentes (62%) ficaram satisfeitos com o DL, contra 13 (12%) insatisfeitos. Tais dados corroborariam com a hipótese do DL ser de mais fácil utilização do que o site da IO.

A [figura 17](#) demonstra uma clara preponderância de acessos via *desktop*. Tal fato reafirma o atestado na [figura 18](#), de que o uso do site é feito principalmente no ambiente

Figura 16 – Respostas para a pergunta: “**Você ficou satisfeito com o sistema de busca por artigos do Diário Oficial Experimental?**”

Fonte: autoria própria

Figura 17 – Respostas para a pergunta: “**Que dispositivos você utilizou para acessar o site?**”

Fonte: autoria própria

de trabalho. Já a [figura 19](#) mostra que 10% dos respondentes dessa pergunta afirmaram ter baixado a base de dados. Apesar de ser uma minoria, ainda aparenta ser um número expressivo, dado a especificidade dessa funcionalidade.

Na [figura 20](#) pode-se ver a distribuição dos respondentes de acordo com a frequência de acesso, apresentando uma concentração maior entre visitantes mais assíduos. A [figura 21](#) indica que a maioria dos visitantes conheceu o site a partir de um e-mail, provavelmente a divulgação interna da prefeitura. Esse grupo é seguido por usuários provenientes de notícias.

Figura 18 – Respostas para a pergunta: “Qual seu interesse no Diário Oficial?”

Fonte: autoria própria

Figura 19 – Respostas para a pergunta: “Você realizou o download integral da base de dados dos Diários Oficiais fornecidos?”

Fonte: autoria própria

Figura 20 – Respostas para a pergunta: “**Com que frequência você consulta o Diário Oficial?**”

Fonte: autoria própria

Figura 21 – Respostas para a pergunta: “**Como você ficou sabendo do site experimental do diário oficial?**”

Fonte: autoria própria

Tendo em vista que a maioria dos visitantes do DL o acessam a partir de buscadores, é razoável afirmar que uma porcentagem menor deste último grupo respondeu o questionário em comparação, por exemplo, aos que conheceram o site por e-mail.

Por fim, como visualizado na [figura 22](#), 29 pessoas (28%) afirmaram sentir falta de alguma funcionalidade no DL. Das funcionalidades requisitadas, diversas já estavam

Figura 22 – Respostas para a pergunta: “**Você sentiu falta de alguma funcionalidade no site do Diário Oficial Experimental? Ou haveria alguma funcionalidade que você gostaria de sugerir?**”

Fonte: autoria própria

disponíveis no sistema, possivelmente indicando um desenho da interface não ótimo. Das funcionalidades requeridas não implementadas pode-se apontar:

- Listas para decretos, portarias e demais atos frequentemente consultados.
- Informação sobre em que página do DOM o artigo foi publicado.
- Buscas via Sparql⁹.
- Possibilidade de dispor os títulos dos artigos em linha temporal.
- Envio automático de e-mail quando publicadas determinadas palavras chave.
- Visualização semelhante ao jornal impresso, com todos os artigos em um único documento.
- Acesso via RSS ou Atom.
- Mais metadados sobre o conteúdo do artigo e o publicante.

Dessas funcionalidades, a única que foi implementada até o momento atual foi o acesso via RSS ou Atom, pois já estavam inclusas no *BlackLight*, bastando ativá-las.

⁹ <https://en.wikipedia.org/wiki/SPARQL>

5.1.3 Críticas e Limitações

Aqui são apresentadas algumas críticas recebidas e limitações relativas à ferramenta. Para conferir maior privacidade e segurança no acesso ao DL e ao questionário online, a página foi disponibilizada inicialmente como HTTPS. Como o projeto todo foi feito de forma não onerosa, utilizou-se um certificado auto assinado, evitando gastos com a compra do mesmo. Porém, certificados auto assinados não são reconhecidos pela maioria dos navegadores, resultando em avisos de segurança por parte dos últimos.

Geralmente tais avisos podem ser ignorados escolhendo a opção correspondente no navegador, mas a mera aparição deles já foi o bastante para assustar alguns visitantes. Temendo se tratar de algum vírus, deixaram de acessar o site e entraram em contato para avisar sobre o problema. Tendo em vista a situação, a equipe decidiu alterar o acesso padrão de HTTPS para HTTP. É interessante notar que, com a alteração, o grau de segurança do sistema caiu, porém, na percepção de parte dos usuários o resultado foi o contrário, uma vez que os navegadores deixaram de apresentar os avisos *assustadores*.

Algumas pessoas não descobriram como realizar a filtragem dos dados por data, mesmo a funcionalidade estando presente entre os filtros. Para os usos atuais do DOM, a filtragem por data é um ponto chave. Já para o *BlackLight*, ela é apenas mais uma opção de filtro, não tendo um destaque especial na interface. Esse é um ponto que provavelmente deveria ter sido remodelado na interface do DL.

Outros criticaram a exibição por artigo. Uma pessoa chegou inclusive a afirmar que o DOM é um jornal, e como tal, deveria ser exibido em documento único e contínuo. Essa crítica evidencia os diversos usos desse tipo de dado. Se algumas pessoas estão interessadas em apenas um artigo, querendo sanar uma dúvida específica, outras podem querer ler o DOM, ou folheá-lo em sua totalidade. Para o segundo caso, os PDFs da IO poderiam ser uma melhor opção. Porém, a necessidade de baixar um PDF por página também parece não ótimo. Uma solução seria gerar um documento, possivelmente um PDF automaticamente diagramado, contendo todos os artigos do dia.

Uma funcionalidade ausente no DL, mas requisitada e presente em sistemas semelhantes, é a referência à publicação oficial. Cada artigo teria um link que apontaria para o PDF da IO onde o primeiro foi publicado. Tal funcionalidade é importante pois apenas o segundo tem valor legal. Esse recurso está implementado na maioria dos sites que extraem os DOs. Porém, implementá-lo no DL não parece ser trivial, uma vez que os

dados usados não foram diagramados e não possuem qualquer referência à página onde o artigo foi publicado oficialmente. Já para os sistemas que extraem os textos dos PDFs, como o fazem posteriormente à diagramação e diretamente do PDF oficial, conseguem o número da página com maior facilidade.

Outra funcionalidade desejada, mas não implementada, foi a apresentação de resumos dos artigos em linguagem acessível à população. Apesar de ser possível, até certo ponto, gerá-los automaticamente, não se sabe se a qualidade obtida seria compatível com a expectativa dos usuários. Vale-se notar que em alguns DOs, como no caso do *Federal Register* discutido anteriormente, a escrita desses resumos é feita manualmente pelo governo.

Foi sugerida também a capacidade de um usuário receber avisos por e-mail quando uma determinada palavra-chave, de sua escolha, aparecesse em um artigo novo. Serviço semelhante ao oferecido gratuitamente pelo DO da União e comercialmente pela IO. Tal funcionalidade não estava presente no *BlackLight*, requerendo desenvolvimento próprio. Porém, uma vez descoberta a disponibilização via RSS e Atom, já inclusas no sistema e capazes de substituir o envio por e-mail, preferiu-se esperar para avaliar se seria realmente necessário implementar a última.

Um ponto questionado foi a filtragem de resultados. Segundo a configuração feita no *Solr*, por padrão são retornados resultados que contenham pelo menos um dos termos procurados, de forma semelhante à maioria dos sites de busca. Porém, alguns usuários demonstraram preferência para que se retornasse apenas resultados com todos os termos procurados.

Outra crítica foi quanto ao endereço do site (`devcolab.each.usp.br/do`), difícil de ser lembrado ou escrito. O endereço foi consequência de onde o sistema está hospedado: na máquina *devcolab* do Colab-USP, relacionado à EACH, que por fim é uma escola da USP. O `/do` define a aplicação dentro da máquina *devcolab*, uma vez que há diversas outras aplicações rodando na mesma máquina virtual, algumas nem mesmo relacionadas à esse projeto. Se tratando de um projeto experimental que possivelmente venha a ser internalizado pela prefeitura, adquirindo um endereço governamental, optou-se por não alterar o endereço atual.

Apesar do código do sistema e dos *scripts* relacionados estarem abertos, e de algumas funcionalidades e correções terem sido requisitadas, nenhum desenvolvedor externo à equipe e à prefeitura, contribuiu no desenvolvimento do sistema. Uma crítica relacionada surgiu de

um desenvolvedor independente que participou do evento de lançamento do DL. Segundo ele, sem uma forma clara de recompensa (seja financeira, seja com reconhecimento) dificilmente outros desenvolvedores se envolveriam no projeto. A crítica parece válida, uma vez que as únicas contribuições diretas de código vieram como parte deste trabalho de mestrado, de um trabalho de conclusão de curso e por um desenvolvedor da PRODAM.

Outro ponto que merece nota é o fato do DL não ter um processo rígido que garanta a fidedignidade das informações publicadas por ele. Esse problema tenta ser minimizado com um aviso na página principal da ferramenta, dizendo que as informações disponíveis ali não possuem valor legal. Porém, não se pode descartar a possibilidade de que haja uma falha na publicação dos dados no DL que leve usuários a obter informação errônea. Essa etapa de garantia da validade das informações publicadas costuma não receber muita atenção no desenvolvimento de protótipos e experimentos de dados abertos, mas pode ser um obstáculo significativo para a internalização da ferramenta pela administração pública.

5.1.4 Considerações

Nesta subseção são feitas algumas considerações gerais sobre a ferramenta desenvolvida. Inicialmente pode-se dizer que, apesar de ter levado vários meses e requerer a ajuda de desenvolvedores internos à administração pública, o processo de publicação do DOM em formato aberto foi totalmente automatizado, ocorrendo diariamente.

A disponibilização conseguiu atender a parte significativa dos critérios do modelo de [Berners-Lee \(2010\)](#), publicando dados em formatos não proprietários, processáveis por máquinas, sob licenças livres e acessíveis por URIs únicos, equivalendo a 4 estrelas.

Como discutido em maior profundidade anteriormente, o número dos respondentes do questionário online que afirmaram ter dificuldade com o DL foi cerca de um quarto dos que afirmaram ter dificuldades com o DOM publicado pela IO. Indicando que a ferramenta utilizada nesse projeto realmente consegue apresentar vantagens de usabilidade quando comparada com a da IO.

Em média são publicados aproximadamente 500 novos artigos por dia e mais de 100 mil artigos por ano. As buscas, apesar de envolverem uma base com mais de uma dezena de gigabytes, conseguem ser realizadas em frações de segundo, não prejudicando a usabilidade da ferramenta.

A disponibilização dos artigos de forma que pudessem ser indexados por motores de busca não passou despercebida: a indexação ocorreu em larga escala, mesmo antes do DL ser oficialmente lançado. E de todo os acessos recebidos, cerca de 75% é proveniente desses sites de busca. Atualmente uma busca no *Google* pelos termos *diário livre* (mesmo sem aspas) retorna como primeiro resultado o site do DL (e é o segundo resultado no *Yahoo-Bing*), mostrando que, mesmo sem grandes esforços nesse sentido, apenas disponibilizando a informação, a ferramenta conseguiu uma boa colocação entre os motores de busca.

O número de visitas, mesmo depois de mais de quatro meses do último esforço de divulgação, tem apresentado crescimento contínuo, atingindo atualmente a marca de 300 visitantes por dia.

Funcionalidades voltadas ao usuário comum já existente na plataforma convencional oficial tais como filtragem por data e ferramenta de busca textual também são oferecidas no Diário Livre. Além dessas, o Diário Livre torna mais simples e conveniente a marcação e cópia de textos e permite a filtragem por categorias e o acesso completo aos arquivos *brutos* facilitando que pesquisas possam ser feitas sobre a massa de dados de uma série de mais de dez anos de publicação, características que não são observadas na versão convencional. Uma das funcionalidades oferecidas, a de notificação automática de termos, permite que qualquer pessoa possa receber de forma simples e gratuita um serviço que atualmente é oferecido mediante pagamento à IO.

Mas as características que não possuem paralelo na versão oficial estão relacionadas ao potencial de reuso promovido pelo consumo automatizado desses dados. Com o acesso aos dados através de API, espera-se que o DL facilite ou viabilize soluções de mineração de dados e novos aplicativos de interesse social e econômico na cidade de São Paulo.

Por fim, ressalta-se que todas as tecnologias utilizadas para a abertura desses dados estão sob licenças livres, permitindo o reuso e apropriação das mesmas por outros grupos e, possivelmente, sendo usadas na abertura de outros DOs.

5.2 Interesses e Desafios

Tendo em vista que um dos objetivos propostos para esse trabalho se trata de explorar os interesses e obstáculos relacionados à implantação de um sistema de informação para a publicação de DGA, esta seção busca trazer luz a essa questão sob a ótica dos

resultados deste projeto. Parte das afirmações feitas nesta seção referenciam a entrevista feita com gestores que pode ser encontrada em maior detalhe no [Apêndice D](#).

Um dos fatos que motivou a criação da parceria descrita neste trabalho foi quando os gestores da COPI-CGM descobriram que a maior parte dos acessos ao portal de transparência municipal vinha dos próprios funcionários da prefeitura. Os acessos eram feitos para verificar se o pagamento do salário do mês já havia sido realizado. Apesar de ser comum o uso de tais portais por funcionários da própria administração pública, o fato dessa ser a principal fonte de tráfego ao portal serviu de alerta sobre sua possível subutilização. Tal fato aponta a necessidade da análise de acessos como parte da avaliação de um portal de transparência.

Outro ponto importante para a criação da parceria foi o potencial alto custo de desenvolvimento dos sistemas de disponibilização caso fosse contratada a PRODAM para tal. Fato que fez a COPI-CGM procurar uma outra via: a criação de protótipos de sistemas através de uma parceria não onerosa. É provável que essa abordagem, além de ter poupado a administração pública de gastos maiores, também tenha reduzido o tempo necessário para o desenvolvimento da ferramenta, ao suprimir intermediários e aproximar desenvolvedores e gestores públicos.

Entre as intenções, por parte dos gestores da COPI-CGM, com a abertura dos dados, está o atendimento às leis e a redução de pedidos de acesso à informação, os mesmos aspectos que também motivaram os gestores entrevistados por [Halonen \(2012\)](#). Além disso, a COPI-CGM demonstrou uma preocupação especial em realizar um trabalho de convencimento dos demais gestores quanto à publicação de dados abertos. Isso provavelmente se deve também ao *modus operandi* da própria COPI-CGM, que, por não caber a ela o papel de punir, atua principalmente pelo diálogo e convencimento. Uma postura mais impositiva na demanda de dados, segundo os próprios gestores, pode levar o demandado a adotar uma atitude não cooperativa, criando empecilhos e, possivelmente, inviabilizando o processo.

[Halonen \(2012\)](#) cita alguns casos de abertura de dados onde desenvolvedores foram chamados para contribuir sobre como gostariam de consumir dados públicos e como o respectivo portal deveria ser construído. Esse autor ressalta que havia muita confiança entre governo e desenvolvedores, o que teria permitido o processo. Através de conversas com membros de um projeto semelhante na Índia, foi afirmado que, no caso deles, a confiança precisou ser construída ao longo de quatro anos antes que uma parceria fosse finalmente firmada.

De forma análoga, porém mais breve, o diálogo entre o Colab-USP e a prefeitura municipal começou cerca de um ano antes que o convite de parceria fosse feito. Dessa forma, pode-se apontar a confiança como um fator importante para o estabelecimento de parcerias para a abertura de dados governamentais. Essa exigência dialoga com o fato de, muitas vezes, uma iniciativa de DGA colocar o gestor em uma posição de vulnerabilidade, seja revelando dados inconvenientes, seja frustrando expectativas internas ou externas ao governo (HALONEN, 2012).

Como dito anteriormente, esse projeto almejava inicialmente acompanhar a abertura de quatro bases de dados. Porém, a única base que caminhou em tempo hábil para viabilizar sua análise neste trabalho de mestrado foi a do DOM. Pode-se dizer que isso se deu principalmente porque os dados do DOM já estavam em formato digital e disponíveis para a COPI-CGM devido a um acordo com a SEMPLA.

Halonen (2012) afirma que funcionários públicos reclamavam sobre se o custo da coleta de dados compensa a publicação dos mesmos (no caso se tratava de um maior detalhamento sobre as despesas públicas). A abertura da base do DOM não requereu grande esforço extra de coleta por parte dos funcionários, uma vez que os dados já estavam digitalizados. Além disso, a parte que requereu esforço extra (quando um funcionário ainda tinha que repassar os arquivos diariamente, pois o processo não estava totalmente automatizado), teve diversos atrasos, o que indica esse elemento (o esforço dispendido por parte dos funcionários) como um ponto importante na questão.

No mesmo sentido, como os DOs já estavam publicados de alguma forma na Internet (em PDF), a publicação em outro formato possivelmente não causou grandes atritos internos sob o argumento do sigilo. Avalia-se então que, para um município interessado em começar uma iniciativa de dados abertos, o DO pode ser uma primeira base interessante, permitindo aos gestores um contato inicial com o tema, e sem provocar grandes resistências. Porém, o impacto real da abertura de tais dados ainda é uma questão em aberto.

Ainda sobre os temores quanto à disponibilização dos dados, gestores demonstraram a preocupação de que dados mais detalhados sobre pequenos gastos (no caso da base de automóveis) fossem mal interpretados pela mídia e usados de forma sensacionalista, enquanto os grandes desvios de dinheiro não seriam denunciados. Preocupações semelhantes às encontradas por Halonen (2012). Um gestor também apontou que, quando um órgão público expõe seus dados, uma série de defeitos, antes escondidos, ficam evidentes, caberia

à população e à mídia ter maturidade para lidar com a questão de forma construtiva, e não com acusações que poderiam desincentivar futuras aberturas.

Outro desafio observado foi a dificuldade de articulação entre secretarias da mesma prefeitura. Quando a parceria foi iniciada, a articulação para a abertura do DOM já estava feita. Porém, para as outras bases, como a de imóveis, isso ainda não estava dado. Por diversos meses a prioridade da secretaria responsável por essa última base era a conclusão do Plano Diretor Estratégico Municipal. Sendo assim, o não alinhamento de cronogramas e interesses entre o grupo encarregado da iniciativa de DGA e a secretaria responsável pelos dados é mais um fator que pode inviabilizar a iniciativa.

Alguns gestores também demonstraram a crença de que a publicação de dados em formato PDF garantiria a não adulteração dos mesmos. Dessa forma, dados que às vezes estão disponíveis internamente em tabelas, para publicação, são oferecidos apenas em PDF. Aparentemente, para esses gestores, o PDF transmite uma sensação de segurança e autenticidade necessárias aos documentos oficiais.

Em entrevista, gestores alegaram também que a própria administração pública faz uso intenso do DO, e que uma maior facilidade no reúso desses dados favoreceria bastante a primeira. Ponto reforçado pelo fato do DO, em alguns casos, ser o único lugar que consolida e centraliza uma série de informações, uma vez que elas precisam estar publicadas nele para serem oficializadas.

Porém, ainda segundo os gestores, muitos dos sistemas contratados ou desenvolvidos não foram pensados com a transparência em vista, dificultando hoje a abertura dos dados. Segundo eles existe também um grande atrito cultural (funcionários apegados à cultura do sigilo e não da transparência) e as pessoas demonstram resistência à mudança dos sistemas utilizados. Muitos funcionários estão acostumados a tratar as informações públicas como sigilosas. O fato do DO já ser publicado, eliminaria esse argumento contrário, facilitando essa iniciativa. O próprio fato da LAI ser recente, indicaria o fato dessa mudança cultural ser recente, levando anos para se consolidar. Essa consolidação passaria inclusive pela lenta substituição dos funcionários atuais, conforme forem se aposentando, por novos já cientes da importância da disponibilização das informações públicas.

Sobre os desafios técnicos, pode-se citar a desorganização e falta de padronização da base, questão que ainda é um entrave no reúso dos mesmos. A necessidade de uma parte do processo ser automatizado internamente à prefeitura, e a ausência de desenvolvedores próprios, chegou a gerar atrasos na atualização dos dados, mas foi superada ao fim. O

tamanho da base, apesar de significativo, não chegou a ser um grave problema, uma vez que o software livre empregado conseguiu realizar as buscas em tempo satisfatório.

Vale ressaltar que a infraestrutura disponível para esse projeto (máquina virtual da USP), apesar de gratuita nesse caso, possuía recursos consideravelmente acima das opções mais baratas em uma hospedagem convencional. Um grupo que tentasse replicar uma iniciativa do tipo, mas sem acesso a uma infraestrutura como essa, e sem recursos para contratar uma equivalente, poderia ter seu projeto inviabilizado.

Ainda sobre a questão técnica, a Imprensa Nacional passou esse ano a disponibilizar o download dos PDFs agrupados por seção, não só apenas por página, como antes. Porém, esse recurso só está disponível das 18h às 0h, segundo a notícia, para evitar a sobrecarga dos servidores ([IMPRENSA NACIONAL, 2015](#)). O DL permite o download não de uma seção por vez, mas de 10 anos de dados a qualquer momento e não apresentou problema de sobrecarga no servidor. Porém, vale-se notar que muito provavelmente atende a um público muito menor (um município) enquanto o DOU se refere ao país. Uma alternativa para a publicação do DOU, caso tenham receio de sobrecarga dos servidores, seria justamente a utilização do protocolo *Torrent*, que no caso deste projeto não se mostrou necessário.

De certa forma, pode-se dizer, mas não comprovar, que um dos resultados importantes desse trabalho foi contribuir com uma mudança cultural relacionada aos dados abertos. Acredita-se nisso pois, muitos dos funcionários que receberam o e-mail de divulgação da prefeitura, ou dos ouvintes da entrevista na rádio, estavam tendo ali seu primeiro contato com o termo *dado aberto*, passando assim a ter uma ideia um pouco mais concreta do que ele pode significar, materializada no DL.

Tendo em vista os pontos apresentados nessa seção, a seguir são listados os principais deles. Primeiramente os que podem ser considerados interesses que motivaram a administração pública em questão a desenvolver iniciativas de dados abertos:

- Favorecer a apropriação do portal de transparência pela população.
- Atendimento às leis relacionadas.
- Redução de pedidos de acesso à informação.
- Fortalecimento da cultura de abertura entre os próprios gestores.
- Permitir um melhor aproveitamento dos dados pela própria administração.

E a seguir, os pontos que podem ser considerados obstáculos para essa ou outras iniciativas relacionadas, desenvolvidas pela mesma administração pública:

- Sistemas atuais não terem sido pensados com transparência em vista.
- Potencial alto custo do desenvolvimento de sistemas.
- Potencial alto custo da coleta de dados.
- Falta de padrão nos dados, especialmente ao longo dos anos.
- Insegurança com relação à garantia de confiabilidade e segurança dos dados.
- Insegurança sobre se expor ao abrir dados (como o uso dos mesmos de forma sensacionalista pela mídia).
- Dificuldade de articulação entre secretarias, uma vez que nem sempre a coleta, armazenamento e publicação será feita pela mesma secretaria.
- Cultura de sigilo e desconhecimento dos DGA.
- Resistência à mudança dos sistemas utilizados.

5.2.1 Questionário do Primeiro Evento

Relacionado aos temores e expectativas envolvendo os DGA, durante o primeiro evento, ao longo das quatro partes, foi distribuído um questionário aos presentes, sendo que 37 foram devolvidos respondidos. O questionário pode ser visto no [Apêndice A](#). Este questionário buscou conhecer melhor o público do evento, assim como suas expectativas sobre dados abertos e os desafios que os cercam. Por fim, no questionário, havia um espaço para comentários e avaliação do próprio evento. Maiores detalhes sobre esse primeiro evento podem ser encontrados no [Apêndice B](#).

O público presente foi majoritariamente masculino (cerca de 73%) e, segundo a [figura 23](#), com pelo menos curso superior. Pode-se dizer assim que não foi um recorte representativo da população como um todo, mas sim de um grupo específico. Além disso, é importante ressaltar que apesar do evento ter sido aberto para qualquer pessoa, a divulgação foi feita também para uma lista de e-mails de pessoas que já participavam de eventos sobre governo aberto, o que pode ter influenciado o viés da amostra.

A partir das respostas sobre profissão e cargo fornecidas por cada participante foi inferido a qual grupo ele pertencia. Os dados estão representados na [figura 24](#) onde se vê uma predominância na presença de gestores públicos no evento. É importante ressaltar também que nem todos os participantes responderam ao questionário.

A grande maioria dos respondentes, segundo a [figura 25](#), avaliou que o projeto em questão traria benefícios significativos para a sociedade. De acordo com a [figura 26](#), esses

Figura 23 – Respostas para a pergunta: “**Escolaridade**”

Fonte: autoria própria

Figura 24 – Grupo a qual pertencia cada participante do evento. Os grupos foram inferidos a partir das respostas sobre profissão e cargo fornecidas por cada participante.

Fonte: autoria própria

Figura 25 – Respostas para a pergunta: “**Você acha que esse projeto trará benefícios para a sociedade como um todo?**”

Fonte: autoria própria

benefícios afetariam diversos grupos da sociedade, principalmente OSCs, gestores públicos, cidadãos em geral e movimentos sociais. Quando perguntados sobre quais grupos deveriam ser convidados para um evento desse tipo, [figura 27](#), os participantes deram prioridade para cidadãos, gestores públicos e OSCs.

A grande maioria dos presentes também considerou, segundo a [figura 28](#), que a abertura dos dados governamentais é algo *muitíssimo* importante. E, apesar de uma quantidade significativa dos participantes considerar a abertura dos dados como algo dispendioso, [figura 29](#), a grande maioria ainda afirma que o custo compensa o benefício, [figura 30](#).

Sobre os grupos interessados sobre esses dados, a [figura 31](#) mostra que as OSCs ganharam destaque nas respostas. Quando perguntados sobre se a abertura dos dados aumentaria a confiança da população nos governos, segundo a [figura 32](#), a grande maioria afirmou que sim.

Quanto ao papel da mídia, segundo a [figura 33](#), as opiniões ficaram divididas, mas houve uma predominância em afirmar que às vezes ela usa os dados de forma irresponsável. De forma semelhante, na [figura 34](#), a maioria afirmou que os gestores públicos podem ter medo da abertura de certos dados.

Figura 26 – Respostas para a pergunta: “**Quem você acha que se beneficia com um projeto desses?**”

Fonte: autoria própria

Figura 27 – Respostas para a pergunta: “**Quem você acha que deveria ser convidado para um evento desse tipo?**”

Fonte: autoria própria

Figura 28 – Respostas para a pergunta: “**Você acha que a abertura de dados governamentais é importante?**”

Fonte: autoria própria

Figura 29 – Respostas para a pergunta: “**Você julga que a abertura dos dados públicos é dispendiosa para o estado?**”

Fonte: autoria própria

Figura 30 – Respostas para a pergunta: “**Você acha que os benefícios da abertura de dados públicos compensam o esforço dispendido?**”

Fonte: autoria própria

Figura 31 – Respostas para a pergunta: “**Quais grupos da sociedade você acha que atualmente se interessam por esses dados?**”

Fonte: autoria própria

Figura 32 – Respostas para a pergunta: “Você acha que a abertura dos dados aumenta a confiança da população nos governos?”

Fonte: autoria própria

Figura 33 – Respostas para a pergunta: “Você acha que a mídia usa dados governamentais para denunciar o governo de forma irresponsável?”

Fonte: autoria própria

Figura 34 – Respostas para a pergunta: “**Você acha que gestores públicos podem ter medo da abertura de certos dados governamentais?**”

Fonte: autoria própria

Sobre a legislação relacionada aos DGA, como pode ser visto na [figura 35](#), uma parcela significativa a considerou de qualidade mediana, havendo algumas respostas a mais para o aspecto negativo do que positivo da mesma. Segundo a [figura 36](#), na opinião dos presentes, o prefeito é um apoiador de iniciativas de abertura de dados.

Na opinião da maioria dos presentes, segundo a [figura 37](#), participar de um evento do tipo expõe pouco ou quase nada um gestor público. Segundo a [figura 38](#), as opiniões ficaram divididas sobre se o posicionamento de um gestor perante uma iniciativa de dados abertos pode colocá-lo em situação vulnerável. Porém, houve uma predominância em afirmar que não, ou apenas quando se posicionar contrariamente. Por fim, de acordo com a [figura 39](#), a maioria dos presentes afirmou se sentir satisfeito com o evento.

Pelos dados pode-se ver que o evento consistia maioritariamente de pessoas favoráveis aos DGA, fato talvez influenciado pela divulgação do mesmo entre já frequentadores de eventos sobre governo aberto. Quanto aos grupos mais relacionados com DGA, os mais citados foram OSCs, cidadãos, gestores e movimentos sociais. Do lado oposto ficaram os partidos políticos e empresas.

Uma pessoa escreveu no questionário que desconhecia o significado do termo *hacker*. Ela pode não ter sido a única, uma vez que o grupo *hacker* teve praticamente o mesmo

Figura 35 – Respostas para a pergunta: “**Você acha que a legislação atual sobre dados abertos é adequada?**”

Fonte: autoria própria

Figura 36 – Respostas para a pergunta: “**Você acha que o prefeito de São Paulo apoia iniciativas de abertura de dados?**”

Fonte: autoria própria

Figura 37 – Respostas para a pergunta: “Você acha que a presença de gestores em um evento público sobre abertura de dados governamentais pode colocá-los em situação vulnerável?”

Fonte: autoria própria

Figura 38 – Respostas para a pergunta: “Você acha que se posicionar contra ou a favor da abertura de dados pode colocar um(a) gestor(a) em situação vulnerável?”

Fonte: autoria própria

Figura 39 – Respostas para a pergunta: “Esse evento atendeu às suas expectativas?”

Fonte: autoria própria

número de citações em cada pergunta, possivelmente sendo assinalado apenas por aqueles que o conheciam. O termo deveria ter sido explicitado no questionário.

5.3 Participação

Nessa seção os resultados são avaliados sob a perspectiva do objetivo, inicialmente proposto, de contribuir para que o processo de abertura fosse participativo. Como colocado por Rocha (2009) e Alfano (2011), também foi confirmada uma certa dificuldade para permitir uma participação mais ampla da sociedade ao longo do processo. Essa dificuldade se deu por diversos motivos, entre eles o esforço necessário para organizar eventos, sistematizar contribuições e responder de forma satisfatória a elas. Se a equipe da prefeitura já era pequena e não dispunha de muito tempo para dedicar à organização de eventos, o mesmo valia para a equipe de desenvolvimento e as demandas técnicas levantadas.

Havia a intenção do lançamento de uma ferramenta online para colher contribuições sobre todos os projetos relacionados a essa parceria. Tal ferramenta ficou sob a responsabilidade de um estagiário contratado pela prefeitura unicamente para o projeto. Porém, após alguns meses de dificuldade na instalação da ferramenta de participação, o estagiário foi chamado para trabalhar em uma empresa, deixando a ferramenta não totalmente funcional e que, por fim, acabou não sendo utilizada.

As contribuições online ficaram restritas então apenas ao DL, através do questionário disponível no site. Vale a ressalva de que esse questionário possuía espaço para comentários livres, que foi usado por muitos dos respondentes. Buscou-se, sempre que possível, responder aos comentários quando estes forneciam um e-mail para contato.

Também pode-se dizer que havia um certo receio por parte dos gestores em realizar eventos mais amplos para a discussão dos projetos. Em parte, receio da presença de pessoas que *desvirtuassem* o foco do evento. Fato que de certa forma ocorreu nos dois eventos com, por exemplo, denúncias de corrupção. Estas, apesar de serem importantes, ficaram fora de contexto, já que os eventos não tratavam especificamente do tema e elas não envolviam nenhum gestores presentes. Por outro lado, essas participações podem ser compreendidas quando levada em consideração a dificuldade que um cidadão sente ao se tentar fazer ouvir.

Tendo em vista as limitações expostas acima e fazendo-se a ressalva de que as decisões finais permaneceram centralizadas na equipe inicial, pode-se afirmar mesmo assim que as contribuições externas recebidas foram positivas. O primeiro evento, detalhado no [Apêndice B](#), contribuiu na formulação de diversos requisitos do sistema, revelou alguns medos internos, e serviu também para fortalecer a cultura dos DGA entre os próprios gestores. O segundo evento, detalhado no [Apêndice C](#), auxiliou na divulgação do projeto, assim com no balanço do mesmo e levantamento de algumas questões importantes sobre sua replicação e apropriação. Sendo assim, apesar das expectativas iniciais do projeto em permitir uma participação mais ampla do que o ocorrido, o saldo total continua sendo positivo.

5.4 Apropriações Externas

Buscando avaliar como está se dando a apropriação do projeto por pessoas externas a ele, nessa seção este tema é debatido em maior profundidade, começando por sua repercussão.

Uma vez lançado, o DL foi divulgado no site principal da prefeitura¹⁰, na página da CGM¹¹ e no site da USP¹². Cada uma dessas notícias foi repostada pelo menos uma

¹⁰ <http://www.capital.sp.gov.br/portal/noticia/4906>

¹¹ http://www.prefeitura.sp.gov.br/cidade/secretarias/controladoria_geral/noticias/?p=183620

¹² <http://www5.usp.br/68247/plataforma-criada-na-each-facilita-uso-de-informacoes-do-diario-oficial>

vez por sites não relacionados ao projeto^{13,14,15,16}. Além disso, houve uma entrevista na rádio CBN¹⁷ e mais uma notícia publicada¹⁸. A COPI-CGM também divulgou o projeto através de um e-mail interno, enviado para todos os funcionários da prefeitura (mais de 100 mil), e publicando um anúncio no próprio DOM¹⁹ (17/12/2014, página 25). Os efeitos da divulgação no número de acessos ao site são discutidos na [subseção 5.1.1](#).

O fato dos dados estarem disponíveis agora em formatos mais fáceis de serem processados por máquinas já possibilitou algumas iniciativas. Um dos desenvolvedores empregados pela prefeitura fez um *script* para extrair algumas informações utilizadas diariamente por alguns órgãos²⁰. Entre elas está a lista de gestores exonerados ou empossados, ou então referências a um determinado órgão.

O trabalho recebeu elogios e segundo um gestor, chegou a economizar 30 minutos diários da funcionária de um desses órgãos, encarregada de marcar (com caneta grifa texto) as informações relativas àquele órgão em uma versão impressa do DOM. Imagina-se que a economia total de tempo tenha sido maior. Porém mudanças no formato do texto dos artigos causaram falhas nas expressões regulares utilizadas na extração das informações. Apesar de terem sido feitas correções, sem um padrão na estruturação do texto dos artigos não há garantia de que um problema semelhante não volte a ocorrer. Essa experiência ilustra o potencial de tais dados e, ao mesmo tempo, os obstáculos para sua utilização.

De forma semelhante um entusiasta do projeto utilizou expressões regulares para extrair dos artigos as portarias e decretos publicados. Estuda-se o uso dessas últimas informações para cruzamento com outras bases, mais especificamente, orçamentárias.

Pode-se supor assim que o DL reduziu a barreira de acesso para o processamento em larga escala dos dados, uma vez que a extração integral da base, ou acesso através de API, foi facilitada. Essa hipótese foi defendida também por um gestor e um empresário

¹³ <http://nossasaopaulo.org.br/noticias/diario-oficial-da-cidade-de-sao-paulo-ganha-nova-versao-line-para-consultar-leis>

¹⁴ <http://igovsp.net/sp/controladoria-e-colab-usp-lancam-plataforma-colaborativa-para-aprimorar-as-consultas-ao-diario-oficial-de-sp/>

¹⁵ <http://www.pesquisamundi.org/2014/11/plataforma-facilita-uso-de-informacoes.html>

¹⁶ <http://each.uspnet.usp.br/site/conteudo-imprensa-noticia.php?noticia=1866>

¹⁷ <http://cbn.globoradio.globo.com/programas/cbn-sao-paulo/2014/10/28/PLATAFORMA-COLABORATIVA-VAI-APRIMORAR-A-CONSULTA-E-A-REUTILIZACAO-DAS-INFORMACOES-DO-D.htm>

¹⁸ <http://www.guiadascidadesdigitais.com.br/site/pagina/dirio-oficial-em-formato-aberto-permite-big-data-popular>

¹⁹ <http://www.docidadesp.imprensaoficial.com.br/NavegaEdicao.aspx?ClipID=3G8T3LUMU9RI0eDEJ3EI6PCFBV6>

²⁰ <https://github.com/pnspinelli/BuscaDiarioLivre>

participantes do primeiro evento. Uma pergunta que pode ser levantada é se o projeto trouxe benefícios para grupos com menor familiaridade tecnológica.

No primeiro evento, uma representante de movimento social questionou sobre como essa iniciativa beneficiaria as camadas mais pobres da sociedade, em suas palavras: “*o que sobra para o povão?*”. Em sua resposta o gestor argumentou que, apesar da disponibilização atual não ser ideal para todos os públicos, uma vez que os dados estão em formato aberto seria mais fácil para, por exemplo, uma *startup* os reestruturar e disponibilizar algo mais útil para outros segmentos da sociedade. Já na opinião da senhora que levantou a questão, o conteúdo do DOM deveria ser resumido em temas e distribuído nos núcleos de assistência social da cidade, pois seriam nesses lugares que a população mais vulnerável poderia se informar sobre os atos mais recentes da política municipal e se mobilizar ao redor das pautas de seu interesse, tarefa que dificilmente seria cumprida por uma *startup*. Implementar algo do tipo talvez envolva questões políticas e financeiras bastante complexas, porém vai ao encontro do que defende [Gurstein \(2011\)](#) sobre a necessidade de se colocar os dados abertos a serviço dos grupos marginalizados.

Esse questionamento revela um debate ainda não resolvido na área de DGA: até que ponto vai a responsabilidade do governo na disponibilização dos dados? Algumas linhas defendem que, uma vez com os dados disponíveis, intermediários assumirão o papel de reformatá-los para os diversos públicos. Já outros acreditam que não se pode contar com o surgimento de intermediários e, especialmente, que estes irão atender às camadas mais marginalizadas e de menor poder aquisitivo. Se o *mercado* muitas vezes já não as atendem, intermediários regidos pelas mesmas regras provavelmente também não o fariam. E nesse caso, como argumenta [Gurstein \(2011\)](#), a abertura dos dados poderia favorecer aos já favorecidos, e não o contrário.

Por fim, apesar do prazo imposto ao encerramento deste documento, o DL continua online, possivelmente produzindo novos resultados importantes. Uma questão relevante e ainda em aberto é quanto à internalização do sistema pela prefeitura, uma vez que ele continua hospedado no servidor do Colab-USP. Tendo em vista que o sistema foi desenvolvido de forma experimental, a internalização do mesmo não está garantida, o que pode comprometer sua existência a longo prazo.

6 Considerações Finais

Chegando ao fim desse documento, diversas considerações podem ser feitas sobre este trabalho como um todo. Seguindo a metodologia proposta, foi firmada uma parceria com um órgão municipal responsável pela abertura de dados, e através dela foi possível acompanhar e auxiliar o processo. Este envolveu diversas reuniões entre a equipe e, em conjunto com um primeiro evento e com a bibliografia relacionada, possibilitou o levantamento dos requisitos para a ferramenta publicadora dos dados. Esta foi então, dentro da mesma parceria, projetada, implementada e lançada em um segundo evento.

Além disso, como um todo, avalia-se que o processo cumpriu várias das recomendações levantadas na literatura sobre a abertura de dados governamentais. E a seguir é discutido o cumprimento dos objetivos iniciais propostos.

Especificamente sobre a ferramenta produzida, apesar dela não satisfazer todos os usos possíveis para os dados, pode-se dizer que ela atendeu aos principais requisitos levantados, recebendo uma boa avaliação pelos usuários e pela administração pública, e obtendo um número crescente de acessos.

Sobre os desafios e interesses relacionados à iniciativa, estes foram explorados através de eventos, entrevista, questionário e conversas com os gestores envolvidos. Como os principais interesses pode-se elencar a busca por uma maior transparência e o cumprimento de leis relacionadas. Os desafios, por sua vez, são provocados principalmente pelo fato da preocupação com os DGA ser relativamente recente, requerendo diversas mudanças tecnológicas e culturais.

Do ponto de vista da participação, buscou-se levá-la em consideração em todas as etapas do processo. Apesar das limitações impostas, que talvez levaram a um nível de participação inferior ao inicialmente almejado, considera-se que o resultado final nesse aspecto foi satisfatório. Foram realizados dois eventos, sendo que o primeiro contribuiu com requisitos presentes na ferramenta final, e o segundo auxiliou em um balanço geral do processo. Além disso, o questionário online também possibilitou mais uma via de contato, que foi utilizada por mais de uma centena de usuários.

O sistema teve alguma repercussão na mídia, o que envolveu uma entrevista de rádio. A publicação dos dados também já permitiu algumas tentativas de reuso dos mesmos, tanto por parte da própria administração pública, como por pessoas externas a ela. Essas tentativas já demonstraram o potencial da base na obtenção de informações

não sistematizadas em qualquer outro lugar, porém, elas também revelaram os desafios no reúso desses dados.

Dessa forma, consideram-se atingidos os principais objetivos desse projeto. Ele porém, traz novos questionamentos e possibilidades para pesquisas futuras. Entre estas está a possibilidade de publicação da base como dados ligados, permitindo um novo conjunto de usos para estes. Outra seria a relação com projetos de *Open Contracting*¹, que buscam a publicação de contratos de forma padronizada, permitindo a análise dos mesmos em escala global. Por fim, outra questão em aberto diz respeito ao impacto que tais dados podem vir a produzir, seja na própria administração pública, seja para os diversos grupos que compõem a sociedade: *hackers*, jornalistas, acadêmicos, OSCs, movimentos sociais, pessoas com pouca familiaridade tecnológica e outros.

¹ <http://www.open-contracting.org>

Referências²

- ABRAMO, C. W. *Gutenberg em bits - Breve panorama dos Diários Oficiais brasileiros*. Transparência Brasil, 2007. Disponível em: <<http://transparencia.org.br/docs/DOs.pdf>>. Acesso em: 24 abr. 2015. Citado 2 vezes nas páginas 25 e 33.
- ALBANO, C. S. *Dados governamentais abertos: proposta de um modelo de produção e utilização de informações sob a ótica conceitual da cadeia de valor*. Tese (Doutorado) — Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, São Paulo, 2014. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/12/12139/tde-03062014-170642/>>. Acesso em: 2014-07-31. Citado 2 vezes nas páginas 26 e 56.
- ALFANO, G. Adapting bureaucracy to the Internet. The case of Venice Local Government. *Information Polity*, IOS Press, v. 16, n. 1, p. 5–22, 2011. Citado 2 vezes nas páginas 27 e 109.
- BASKERVILLE, R. L. Investigating information systems with action research. *Communications of the AIS*, Association for information Systems, v. 2, n. 3es, p. 4, 1999. Citado 3 vezes nas páginas 43, 44 e 45.
- BENJAMIN, S.; BHUVANESWARI, R.; RAJAN, P. Bhoomi: ‘E-governance’, or, an anti-politics machine necessary to globalize bangalore? *CASUM-m Working Paper*, 2007. Citado 7 vezes nas páginas 26, 27, 28, 36, 37, 38 e 40.
- BERNERS-LEE, T. 2009. Disponível em: <http://www.ted.com/talks/tim_berniers_lee_the_year_open_data_went_worldwide>. Acesso em: 23 nov. 2014. Citado na página 39.
- BERNERS-LEE, T. *Is your Linked Open Data 5 Star?* 2010. Disponível em: <<http://www.w3.org/DesignIssues/LinkedData.html>>. Acesso em: 17 set. 2014. Citado 4 vezes nas páginas 25, 34, 76 e 93.
- BRANDAO, S. N. et al. Open government knowledge base. In: *ICDS 2013, The Seventh International Conference on Digital Society*. [s.n.], 2013. p. 13–19. Disponível em: <http://www.thinkmind.org/index.php?view=article&articleid=icds_2013_1_30_10168>. Acesso em: 26 abr. 2015. Citado 2 vezes nas páginas 27 e 32.
- BRASIL. Decreto nº 4.520. 16 de dezembro de 2002. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/2002/D4520.htm>. Citado 2 vezes nas páginas 31 e 55.
- BRASIL. Lei nº 11.419. 19 de dezembro de 2006. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2006/Lei/L11419.htm>. Citado na página 31.
- BRASIL. Decreto nº 6.932. 11 de agosto de 2009. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2009/Decreto/D6932.htm>. Citado na página 56.
- BRASIL. Lei complementar nº 131. 27 de maio de 2009. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/LCP/Lcp131.htm>. Citado na página 56.

² De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

BRASIL. Lei nº 12.527 (lai). 18 de novembro de 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm>. Citado 2 vezes nas páginas 57 e 137.

CHIGNARD, S. *A brief history of Open Data*. 2013. Blog. Disponível em: <<http://www.paristechreview.com/2013/03/29/brief-history-open-data/>>. Acesso em: 20 nov. 2014. Citado na página 33.

CIFUENTES-SILVA, F.; SIFAQUI, C.; LABRA-GAYO, J. E. Towards an architecture and adoption process for linked data technologies in open government contexts: A case study for the library of congress of chile. In: *Proceedings of the 7th International Conference on Semantic Systems*. New York, NY, USA: ACM, 2011. (I-Semantics '11), p. 79–86. ISBN 978-1-4503-0621-8. Disponível em: <<http://doi.acm.org/10.1145/2063518.2063529>>. Citado 2 vezes nas páginas 27 e 32.

COLQUHOUN, A. The cost of freedom of information. *Constitution Unit, University*, 2010. Disponível em: <<http://www.ucl.ac.uk/constitution-unit/research/foi/countries/cost-of-foi.pdf>>. Citado na página 35.

CONTROLADORIA GERAL DO MUNICÍPIO. *Apresentação*. 2013. Disponível em: <http://www.prefeitura.sp.gov.br/cidade/secretarias/controladoria_geral/a_cgm/index.php?p=142874>. Acesso em: 5 ago. 2014. Citado na página 50.

COORDENADORIA DE PROMOÇÃO DA INTEGRIDADE. *Atribuições*. 2013. Disponível em: <http://www.prefeitura.sp.gov.br/cidade/secretarias/controladoria_geral/coordenadoria_de_promocao_da_integridade/a_copi/index.php>. Acesso em: 5 ago. 2014. Citado na página 50.

COORDENADORIA DE PROMOÇÃO DA INTEGRIDADE. *Legislação*. 2013. Disponível em: <http://www.prefeitura.sp.gov.br/cidade/secretarias/controladoria_geral/coordenadoria_de_promocao_da_integridade/legislacao/index.php?p=153320>. Acesso em: 5 ago. 2014. Citado na página 57.

CREDENDIO, J. E. *Após ameaça, 'xerife' de Haddad usa carro blindado e tem seguranças*. 2013. Notícia. Disponível em: <<http://www1.folha.uol.com.br/cotidiano/2013/11/1366128-apos-ameaca-xerife-de-haddad-usa-blindado.shtml>>. Acesso em: 5 ago. 2014. Citado na página 137.

DAVIES, T. *Evaluating the Autumn Statement Open Data Measures*. 2011. Blog. Disponível em: <<http://www.timdavies.org.uk/2011/12/02/3090>>. Acesso em: 5 ago. 2014. Citado na página 36.

DAVIES, T.; FRANK, M. 'There's no such thing as raw data': exploring the socio-technical life of a government dataset. In: ACM. *Proceedings of the 5th Annual ACM Web Science Conference*. [S.l.], 2013. p. 75–78. Citado 2 vezes nas páginas 26 e 65.

DAVIES, T.; PERINI, F.; ALONSO, J. Researching the emerging impacts of open data. *World Wide Web Foundation*, v. 20, 2013. Citado na página 25.

DAVIES, T. G. Open data policies and practice: An international comparison. *Available at SSRN 2492520*, 2014. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2492520>. Acesso em: 25 abr. 2015. Citado na página 32.

DENIS, J.; GOËTA, S. Exploration, Extraction and 'Rawification'. The Shaping of Transparency in the Back Rooms of Open Data. In: *After The Reveal. Open Questions on Closed Systems - Neil Postman Graduate Conference*. New York, USA: [s.n.], 2014. Disponível em: <<http://halshs.archives-ouvertes.fr/halshs-00954302>>. Citado 2 vezes nas páginas 26 e 139.

DUDLEY, L. R. Federal register 2.0: Public participation in the twenty-first century. *Legis. & Pol'y Brief*, HeinOnline, v. 3, p. vii, 2011. Citado na página 33.

DUPLAIN, R.; BALSER, D. S.; RADZIWILL, N. M. *Build great web search applications quickly with Solr and Blacklight*. 2010. 774011-774011-12 p. Disponível em: <<http://dx.doi.org/10.1117/12.857899>>. Citado na página 71.

FREIRE, P. *Pedagogia da autonomia: saberes necessários à prática educativa*. [S.l.]: Paz e Terra, 2007. (Coleção Leitura). ISBN 9788577530151. Citado na página 9.

GRAINGER, T.; POTTER, T.; SEELEY, Y. *Solr in action*. [S.l.]: Manning, 2014. Citado na página 69.

GRIMMELIKHUIJSEN, S. A good man but a bad wizard. about the limits and future of transparency of democratic governments. *Information Polity: The International Journal of Government & Democracy in the Information Age*, v. 17, n. 3/4, p. 293 – 302, 2012. ISSN 15701255. Disponível em: <<http://search.ebscohost.com/login.aspx?direct=true&db=afh&AN=84341711&lang=pt-br&site=ehost-live>>. Citado 2 vezes nas páginas 36 e 40.

GURSTEIN, M. Open data: Empowering the empowered or effective data use for everyone? *First Monday*, v. 16, n. 2, 2011. ISSN 13960466. Disponível em: <<http://firstmonday.org/ojs/index.php/fm/article/view/3316/2764>>. Citado 7 vezes nas páginas 25, 26, 28, 39, 40, 112 e 140.

HALONEN, A. Being open about data: Analysis of the uk open data policies and applicability of open data. *London: The Finnish Institute in London*, 2012. Citado 8 vezes nas páginas 25, 28, 35, 36, 39, 40, 95 e 96.

HASWANI, M. O discurso obscuro das leis. *Comunicação pública*, v. 5508, p. 53, 2012. Disponível em: <<http://www3.eca.usp.br/sites/default/files/form/ata/pos/ppgcom/ciencias%20da%20comunicacao.%20e-books.%20%20comunica%C3%A7%C3%A3o%20publica.pdf#page=54>>. Citado na página 31.

IMPrensa NACIONAL. 2011. Disponível em: <<http://imprensa2.in.gov.br/inbuscatotal/siscore.do?metodo=orientacoes>>. Acesso em: 5 ago. 2014. Citado na página 58.

IMPrensa NACIONAL. 2015. Disponível em: <<http://portal.in.gov.br/noticias/IMPrensa%20NACIONAL%20LIBERA%20PDF%20COMPLETO%20DO%20DOU>>. Acesso em: 11 mai. 2015. Citado na página 98.

IMPrensa OFICIAL. 2010. Disponível em: <<http://www.docidadesp.imprensaoficial.com.br>>. Acesso em: 5 ago. 2014. Citado na página 57.

IMPrensa OFICIAL. 2011. Disponível em: <http://www.imprensaoficial.com.br/PortalIO/DO/BoletimDOAssinatura_16_1.aspx>. Acesso em: 5 ago. 2014. Citado na página 58.

IMPrensa Oficial. *Manual de Conversão para PDF Envio de Arquivos ao Diário Oficial*. [S.l.], 2011. Disponível em: <<https://pubnet.imprensaoficial.com.br/pubnetii/manuais/ManualGeracaoPDF.pdf>>. Acesso em: 5 ago. 2014. Citado na página 59.

IMPrensa Oficial. 2014. Diário Oficial. Disponível em: <<http://www.docidadesp.imprensaoficial.com.br/CertificadorWrapper.aspx?Ticket=20292379>>. Acesso em: 5 ago. 2014. Citado na página 62.

IMPrensa Oficial. 2014. Diário Oficial. Disponível em: <<http://www.docidadesp.imprensaoficial.com.br/CertificadorWrapper.aspx?Ticket=20294315>>. Acesso em: 5 ago. 2014. Citado na página 67.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, Taylor & Francis, v. 29, n. 4, p. 258–268, 2012. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/10580530.2012.716740>>. Citado 3 vezes nas páginas 25, 26 e 35.

JÄRVINEN, P. Action research is similar to design science. *Quality & Quantity*, Kluwer Academic Publishers, v. 41, n. 1, p. 37–54, 2007. ISSN 0033-5177. Disponível em: <<http://dx.doi.org/10.1007/s11135-005-5427-1>>. Citado na página 43.

KIM, P. H. Action research approach on mobile learning design for the underserved. *Educational Technology Research and Development*, Springer, v. 57, n. 3, p. 415–435, 2009. Citado na página 45.

KUČERA, J.; CHLAPEK, D.; MYNARZ, J. Czech CKAN repository as case study in public sector data cataloging. *Systémová Integrace*, v. 19, n. 2, p. 95–107, 2012. Disponível em: <http://www.cssi.cz/cssi/system/files/all/SI_2012_02_09_Kucera_Chlappek_Mynarz.pdf>. Citado na página 68.

LACERDA, D. P. et al. Design Science Research: método de pesquisa para a engenharia de produção. *Gestão & Produção*, scielo, v. 20, p. 741 – 761, 00 2013. ISSN 0104-530X. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-530X2013000400001&nrm=iso>. Citado 2 vezes nas páginas 43 e 44.

LINDERS, D. From e-government to we-government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly*, Elsevier, v. 29, n. 4, p. 446–454, 2012. Citado na página 33.

OLIVEIRA, C. M. Lislaine Krupek Braz de. Anais do WTRANS13 - Workshop de Transparência em Sistemas, Brasília, Brasil, Setembro 29, 2013. In: LEAL MARÍLIA GUTERRES FERREIRA, J. C. S. d. P. L. André Luiz de C. (Ed.). [S.l.: s.n.], 2013. ISBN ISSN: 2178-6097. Citado na página 26.

OPEN KNOWLEDGE FOUNDATION. *Open Data Handbook Documentation*. [S.l.], 2012. Disponível em: <<http://opendatahandbook.org/pdf/OpenDataHandbook.pdf>>. Acesso em: 14 set. 2014. Citado 2 vezes nas páginas 25 e 40.

POLLOCK, R. *Open Data: a means to an end, not an end in itself*. 2011. Open Knowledge. Disponível em: <<http://blog.okfn.org/2011/09/15/open-data-a-means-to-an-end-not-an-end-in-itself/>>. Acesso em: 22 jul. 2014. Citado na página 35.

PRODAM. *Organização da empresa*. 2013. Disponível em: <<http://www.prefeitura.sp.gov.br/cidade/secretarias/planejamento/prodam/empresa/prodam>>. Acesso em: 5 ago. 2014. Citado na página 51.

RADAR OFICIAL. 2012. Disponível em: <<http://www.radaroficial.com.br>>. Acesso em: 5 ago. 2014. Citado 2 vezes nas páginas 57 e 130.

REID, C. Advancing women's social justice agendas: A feminist action research framework. *International journal of qualitative methods*, v. 3, n. 3, 2004. Citado 3 vezes nas páginas 43, 45 e 47.

RICHARDS, R. C. 2010. Disponível em: <<http://legalinformatics.wordpress.com/2010/07/26/federal-register-2-0-now-available>>. Acesso em: 25 mai. 2015. Citado 2 vezes nas páginas 27 e 32.

ROCHA, R. A gestão descentralizada e participativa das políticas públicas no Brasil. *Revista Pós Ciências Sociais*, v. 6, n. 11, 2009. Citado na página 109.

ROSENBAUM, M. *Why Tony Blair thinks he was an idiot*. 2010. Blog. Disponível em: <https://web.archive.org/web/20130503225511/http://www.bbc.co.uk/blogs/opensecrets/2010/09/why_tony_blair_thinks_he_was_a.html>. Acesso em: 5 ago. 2014. Citado na página 36.

SADLER, E. B. Project blacklight: a next generation library catalog at a first generation university. *Library Hi Tech*, Emerald Group Publishing Limited, v. 27, n. 1, p. 57–67, 2009. Citado na página 71.

SANTANA M., T. *Uma Proposta de Publicação de Dados do Orçamento Público na Web*. Dissertação (Mestrado) — Universidade de São Paulo, 2014. Citado na página 56.

SEMPLA. *Diário Oficial da Cidade de São Paulo: manual de instruções*. [S.l.], 2010. Disponível em: <http://www.prefeitura.sp.gov.br/cidade/secretarias/upload/chamadas/manual_de_instrucoes_do_diario_oficial_2010_1306171567.pdf>. Acesso em: 26 jul. 2014. Citado 3 vezes nas páginas 60, 61 e 62.

SMILEY, D.; PUGH, E. *Solr 1.4 Enterprise Search Server*. [S.l.]: Packt Publishing Ltd, 2009. Citado na página 69.

SODRÉ, N. W. *História da imprensa no Brasil*. [S.l.]: Mauad Editora Ltda, 1998. Citado na página 31.

SUNLIGHT FOUNDATION. 2014. Disponível em: <<http://sunlightfoundation.com/opendataguidelines>>. Acesso em: 6 mar. 2015. Citado 2 vezes nas páginas 25 e 40.

SÃO PAULO. Decreto nº 53.623. 12 de dezembro de 2012. Disponível em: <http://www3.prefeitura.sp.gov.br/cadlem/secretarias/negocios_juridicos/cadlem/integra.asp?alt=13122012D%20536230000>. Citado na página 57.

SÃO PAULO. Decreto nº 54.779. 22 de janeiro de 2014. Disponível em: <<http://camaramunicipalsp.qaplaweb.com.br/iah/fulltext/decretos/D54779.pdf>>. Citado na página 57.

THE 8 PRINCIPLES OF OPEN GOVERNMENT DATA. 2007. Disponível em: <<http://opengovdata.org>>. Acesso em: 5 ago. 2014. Citado 2 vezes nas páginas 25 e 33.

THE GAZETTE. 2014. Disponível em: <<https://www.thegazette.co.uk/history>>. Acesso em: 7 abr. 2015. Citado na página 31.

UBALDI, B. Open government data: Towards empirical analysis of open government data initiatives. OECD Publishing, 2013. Citado 2 vezes nas páginas 25 e 35.

UPLEXIS. 2011. Disponível em: <<http://www.uplexis.com>>. Acesso em: 5 ago. 2014. Citado 2 vezes nas páginas 57 e 59.

VAISHNAVI, V.; KUECHLER, W. *Design Science Research in Information Systems*. 2004. Disponível em: <<http://www.desrist.org/design-research-in-information-systems>>. Acesso em: 15 nov. 2015. Citado 2 vezes nas páginas 43 e 45.

WORLD BANK'S OPEN GOVERNMENT DATA WORKING GROUP. *Open Data Readiness Assessment*. 2013. Citado 2 vezes nas páginas 25 e 40.

XAVIER, J. F. 2012. Disponível em: <<http://fci.uib.es/Servicios/libros/veracruz/xavier2/A-Publicacao-Como-Condicao-de-Eficacia.cid221837>>. Acesso em: 7 abr. 2015. Citado na página 31.

Apêndice A – Questionário Impresso

Nas sete páginas seguintes está uma réplica do questionário distribuído no primeiro evento, e discutido ao longo deste trabalho.

Questionário

Olá! Somos de um grupo de pesquisa da EACH-USP, o Colab, e estamos desenvolvendo uma pesquisa participativa sobre processos de abertura de dados governamentais.

Para tal, gostaríamos de pedir a sua colaboração respondendo as perguntas abaixo e adicionando qualquer outro comentário que julgar pertinente.

Quando achar cabível, sinta-se livre para assinalar mais de uma alternativa ou adicionar novas alternativas.

Sobre você

Nome:

Profissão/Atividade:

Cargo:

E-mail:

Telefone:

Gênero:

☐

Feminino

☐

Masculino

☐

Outro

Escolaridade:

☐

Não-Formal

☐

Fundamental

☐

Médio

☐

Superior

☐

Pós-Graduação

Se superior, qual curso?

Este Projeto

O que te motivou a estar aqui?

O que você sabe sobre o assunto desse encontro? O que mais você gostaria de saber sobre ele?

Você acha que esse projeto trará benefícios para a sociedade como um todo? Por quê?

☐ Pouquíssimo ☐ Pouco ☐ Médio ☐ Muito ☐ MUITÍSSIMO

Quem você acha que se beneficia com um projeto desses? Por quê?

- ☐ Gestores Públicos
- ☐ Movimentos Sociais
- ☐ Organizações da Sociedade Civil
- ☐ Partidos Políticos
- ☐ Empresas
- ☐ Hackers
- ☐ Academia
- ☐ Jornalistas
- ☐ Cidadãos em geral
- ☐ Outros

Quem você acha que deveria ser convidado para um evento desse tipo?

- ☐ Gestores Públicos
- ☐ Movimentos Sociais
- ☐ Organizações da Sociedade Civil
- ☐ Partidos Políticos
- ☐ Empresas
- ☐ Hackers
- ☐ Academia
- ☐ Jornalistas
- ☐ Cidadãos em geral
- ☐ Outros

Você tem conhecimento sobre outros processos de abertura de dados? Qual sua opinião sobre eles?

Abertura de Dados Governamentais

Você acha que a abertura de dados governamentais é importante? Por quê?

- ☐ Pouquíssimo ☐ Pouco ☐ Médio ☐ Muito ☐ MUITÍSSIMO

Você julga que a abertura dos dados públicos é dispendiosa para o estado? Por quê?

- ☐ Pouquíssimo ☐ Pouco ☐ Médio ☐ Muito ☐ MUITÍSSIMO
-
-

Você acha que os benefícios da abertura de dados públicos compensam o esforço dispendido?

☐ Pouquíssimo ☐ Pouco ☐ Médio ☐ Muito ☐ MUITÍSSIMO

Como a abertura de dados governamentais te afeta?

Quais grupos da sociedade você acha que atualmente se interessam por esses dados?

- ☐ Gestores Públicos
- ☐ Movimentos Sociais
- ☐ Organizações da Sociedade Civil
- ☐ Partidos Políticos
- ☐ Empresas
- ☐ Hackers
- ☐ Academia
- ☐ Jornalistas
- ☐ Cidadãos em geral
- ☐ Outros

Em que tipos de dados você acha que a sociedade tem mais interesse?

Que tipos de dados você mais gostaria de ver abertos?

Você acha que a abertura dos dados aumenta a confiança da população nos governos?
Por quê?

☐ Pouquíssimo ☐ Pouco ☐ Médio ☐ Muito ☐ MUITÍSSIMO

Você acha que a mídia usa dados governamentais para denunciar o governo de forma irresponsável?

☐ Não ☐ Raramente ☐ Às vezes ☐ Frequentemente

Você acha que gestores públicos podem ter medo da abertura de certos dados governamentais?

☐ Não ☐ Raramente ☐ Às vezes ☐ Frequentemente

Você acha que a legislação atual sobre dados abertos é adequada?

☐ Pouquíssimo ☐ Pouco ☐ Médio ☐ Muito ☐ MUITÍSSIMO

Você acha que o prefeito apoia iniciativas de abertura de dados?

☐ Pouquíssimo ☐ Pouco ☐ Médio ☐ Muito ☐ MUITÍSSIMO

Você acha que a presença de gestores em um evento público sobre abertura de dados governamentais pode colocá-los em situação vulnerável? Por quê?

☐ Não ☐ Pouco ☐ Médio ☐ Muito ☐ Certamente

Você acha que se posicionar contra ou à favor da abertura de dados pode colocar um(a) gestor(a) em situação vulnerável? Por quê?

- ☐ Sim em ambos os casos
- ☐ Não para ambos os casos
- ☐ Apenas quando se posicionar contra
- ☐ Apenas quando se posicionar à favor

Que grupos você acha que se incomodariam mais com a abertura de dados?

Há alguma iniciativa de dados abertos por parte da prefeitura que você considera importante?

Você vê alguma preocupação legal quanto à abertura de dados governamentais? Qual?

Balanço

Esse evento atendeu às suas expectativas? Por quê?

☐ Não

☐ Pouco

☐ Médio

☐ Muito

O quê você achou desse questionário? Gostaria de sugerir mais perguntas? Retirar perguntas? Modificá-las?

Críticas, dúvidas, sugestões ou questões existenciais?

Apêndice B – Primeiro Evento

A ideia inicial era de que esse primeiro evento fosse o lançamento da parceria e tivesse o máximo de participação possível. Porém, com os atrasos na formalização da primeira, por volta do meio do mês de maio sugeriu-se deixar o lançamento mais para frente, em agosto, realizando um primeiro evento em julho, e com outro enfoque.

Na deliberação sobre os moldes do primeiro evento, os gestores defenderam um evento de menor porte, com menos participação externa, mas incentivando a participação de outros gestores, responsáveis pela bases envolvidas na parceria. A ideia, segundo os primeiros, era que um evento com ampla participação popular poderia intimidar os gestores convidados. Dessa forma, após algumas divergências, ficou definido um evento para cerca de 40 pessoas e durante o dia, para que os gestores pudessem ir em horário de trabalho. Porém, observado o limite de pessoas, imposto inclusive pelo tamanho do espaço, qualquer pessoa poderia participar.

O formato consistiria de quatro partes de 1h30 cada onde seriam discutidas as quatro bases a serem abertas pela parceria. Cada parte teria um convidado da prefeitura e outro externo compondo a mesa. A mesa exporia inicialmente sobre o assunto e então abriria para contribuições de qualquer presente. Ao longo do evento seria sistematizado, na projeção visível por todos, um documento sintetizando as contribuições.

A maior parte das tarefas relacionadas ao evento (como agendamento do espaço, divulgação, convites, compra de petiscos) ficou a cargo da equipe da prefeitura. A divulgação foi feita principalmente por redes sociais e através de e-mail, enviado para cerca de 350 pessoas, uma semana antes do evento. Os endereços de e-mail foram obtidos a partir das listas de presença de outros eventos da prefeitura relacionados a governo aberto. O convite pedia para que os participantes confirmassem presença através de um formulário online, ajudando a estimar o número de pessoas que estaria presente. O número de inscritos não atingiu os limites estipulados.

Nos convites, o objetivo do evento ficou anunciado como “reunir pessoas com experiência ou interesse no tema, para sugerir conteúdos e funcionalidades para os projetos das seguintes plataformas: Portal da Transparência, Sistema de Convênios, Diário Oficial Aberto e Sistema dos Imóveis Cedidos”. O texto também apresentava mais detalhes sobre cada uma das bases.

Como o foco deste trabalho é a abertura do DOM, apenas a primeira parte do evento será narrada em maior detalhe. O evento começou às 9h40 (10 minutos de atraso e cerca de 30 pessoas presentes) com apresentação da COPI-CGM, Colab-USP e da parceria entre ambos. Às 9h50 começa a explanação da mesa, havendo dois representantes da SEMPLA e um advogado empresário como pessoa externa à prefeitura.

A representante da SEMPLA falou sobre o porte do DOM (cerca de 90 páginas por dia, 30 milhões ao todo) e o uso de certificação digital. Disse que dados abertos são importantes mas demonstrou preocupação sobre como lidar com a segurança nesse caso.

Já o outro membro da mesa é dono de uma empresa que vende serviços relacionados a DOs ([RADAR OFICIAL, 2012](#)). Disse que extraem dados de 550 diários no Brasil todo e os disponibilizam, como texto, na Internet. Segundo ele, a barreira de acesso aos dados é na captação dos DOs, e faz com que apenas pessoas com mais recursos consigam utilizá-los em larga escala. Porém, para o usuário comum, o PDF atenderia às necessidades. Ele recomendou também que o projeto buscasse realizar algo mais simples primeiro, e não tentar implementar um sistema perfeito. Afirmou também que a exibição atual realiza uma desestruturação do formato original, o que impediria a indústria da inovação. Como requisitos básicos recomendou que o texto não fosse colocado em colunas, pois isso dificulta o processamento automático, e que houvesse mais metadados sobre os artigos. Ressaltou que para o diário em formato aberto não há necessidade de certificação digital.

Iniciado o momento para contribuições da plateia, uma cidadã demonstrou preocupação sobre quem conseguiria acessar o novo sistema desenvolvido, recomendando que houvesse capacitação para os interessados. Um dos gestores respondeu demonstrando a crença de que abrindo os dados seria possível a uma *startup* criar algo que facilitasse o acesso das pessoas às informações.

O segundo representante da SEMPLA afirmou terem preocupação com segurança dos dados, mas reconheceu que para alguns grupos isso não é tão crítico. Afirmou também que na Europa os DGA movimentam um mercado milionário. Por fim, usando como exemplo a impressão de mapas que saíram no DOM com algumas distorções e geraram críticas, demonstrou a preocupação de que com DGA os problemas poderiam ser ainda maiores. Um cidadão ativista se apressou em corrigir a última afirmação, dizendo que DGA atuariam justamente no sentido contrário, evitando distorções decorrentes da formatação dos dados para impressão.

Um cidadão engenheiro afirmou que gostaria de receber avisos quando o nome dele fosse publicado no DOM. E sugeriu também a produção de resumos para as notícias, facilitando a leitura.

Uma jornalista pediu que a interface de busca possuisse informações sobre os dados que estão ali dispostos, e como interpretá-los. Além de que fossem disponibilizadas tabelas simples para que pudessem ser importadas em outros programas depois.

Já uma representante da academia pediu mais metadados, *URIs* e maior padronização dos dados. Um cidadão ativista afirmou que PDF não é um bom formato para publicação de dados da Internet, e que a questão de segurança poderia ser facilmente resolvida utilizando códigos *hashes*. Às 11h15 interrompe-se essa parte do evento para um breve intervalo seguido da segunda parte, relativa ao sistema de imóveis.

Durante o intervalo, conversando com uma cidadã, ela sugeriu um esforço maior por parte da administração pública na divulgação das informações contidas no DOM. Isso ocorreria, por exemplo, entregando uma versão impressa, repartida por temas como saúde e educação, a assistentes sociais, subprefeituras, conselhos populares. A ideia é que esses grupos estariam mais próximos da comunidade e conseguiriam mobilizar melhor a população sobre os temas relevantes.

Ao longo do dia foi distribuído um questionário aos presentes. Os resultados do mesmo estão discutidos na [subseção 5.2.1](#).

Apêndice C – Segundo Evento

Aqui consta a narração do segundo evento. Este foi gravado e encontra-se disponível online¹. O evento foi marcado para o fim da tarde, mas começou com certo atraso devido a uma reunião entre os gestores participantes. Foi um dia de chuva o que não deve ter colaborado com a participação, que ficou em torno de 50 pessoas.

O evento foi dividido em duas etapas principais, a primeira tratava da parceria em si, e a segunda do lançamento do DL. Cada etapa possuía uma mesa própria para a explanação do tema, seguida de abertura para perguntas dos presentes.

Nesse sentido, o evento começou com uma mesa composta de dois membros da administração pública e dois membros da academia, todos os quatro envolvidos na parceria. Eles apresentaram a COPI-CGM, o Colab-USP, e suas perspectivas sobre participação, transparência e controle social. Depois de cerca de meia hora apresentando os projetos envolvidos na parceria, abriu-se para perguntas.

As perguntas incluíram tópicos como aumento da democratização da informação (especialmente para os grupos com menor poder aquisitivo), combate à corrupção, e replicação da iniciativa em outras cidades do interior de São Paulo. Nem todas as perguntas tiveram respostas claras dentro do escopo do evento, mas o caminho apontado pela mesa foi no sentido de publicar dados para aumentar a transparência. Mais especificamente sobre a replicabilidade da iniciativa, foi apontado que o projeto e o código desenvolvido são abertos também para facilitar a replicação.

Após cerca de uma hora do início do evento, foi iniciada a segunda etapa. A mesa foi constituída dos dois desenvolvedores do Colab-USP que desenvolveram o código da ferramenta, a professora orientadora do trabalho, dois gestores envolvidos na parceria, e o gestor guardião da base que foi aberta, o DOM. A mesa apresentou o projeto DL tanto de um ponto de vista mais geral como também mais técnico, e abriu para perguntas.

As perguntas giraram ao redor de como fomentar o uso da ferramenta por desenvolvedores, e como garantir a sua manutenção a longo prazo. Sobre a manutenção, a garantia viria da apropriação da ferramenta pela população. Uma vez que, no caso de descontinuação, surgiriam manifestações por parte da sociedade.

Já sobre o envolvimento com mais desenvolvedores, a crítica levantada por um desenvolvedor independente foi no seguinte sentido: no projeto atuaram pessoas que tinham

¹ <https://www.youtube.com/watch?v=EltJWSP0ZDA>

algum interesse concreto envolvido, seja um trabalho de pesquisa (para os acadêmicos), ou a necessidade de mostrar serviço perante a sociedade (para os gestores); porém, como fica a questão de desenvolvedores externos? Como garantir que serão recompensados por seus esforços? As respostas dadas foram na linha de editais públicos, ou de divulgação do trabalho desses desenvolvedores no próprio site do DL.

O gestor também deixou claro que talvez a maior beneficiada com a abertura da base do DOM é justamente a própria administração pública, uma vez que os funcionários fazem uso diário das informações publicadas ali.

Apêndice D – Entrevista

Tendo em vista o papel central da SEMPLA na gestão do DOM, realizou-se uma entrevista com os gestores deste órgão envolvidos diretamente com a publicação. Na entrevista participaram dois gestores, um responsável pela parte de TIC da SEMPLA, E a responsável pela publicação do DOM em si.

Segundo eles, a responsabilidade da SEMPLA perante o DOM é formalizar a relação institucional (ex.: assinar os contratos com a empresa publicante), normatizar a publicação, e resolver situações atípicas (como oficializar o Plano Diretor Estratégico, que envolvia a publicação de um conjunto de mapas).

Ao longo dos anos a tiragem do DOM tem sido reduzida. Foram 6100 exemplares diários no ano 2000, caindo para 5100, e então 1100, tiragem que permanece até hoje. Devido aos custos fixos, não compensaria reduzir ainda mais esse número, restando apenas o corte completo da versão impressa. Contrariando as expectativas da SEMPLA, poucos se manifestaram contra a redução no número de exemplares e, quando se manifestam, são aconselhados a usar a versão online.

O contrato com a IO surge do fato da inexistência de uma gráfica própria da prefeitura. Apontaram também que o sistema de publicação sempre foi descentralizado, cada órgão sendo responsável por publicar suas próprias matérias, e com a adoção do Pubnet, mais recentemente, essa característica se manteve.

Quanto a dados abertos, os gestores se mostraram favoráveis, inclusive afirmaram que recebem alguns pedidos nesse sentido, geralmente de dados relacionados a publicações mais polêmicas no DOM (como no caso do Plano Diretor Estratégico).

Apêndice E – Detalhamento das Outras Bases

Aqui são descritos alguns detalhes relativos às outras bases de dados que compunham a mesma iniciativa de abertura relatada neste projeto. Essas outras bases, por terem apresentado um desenvolvimento menos promissor ou mais vagaroso, acabaram não sendo analisadas em maior profundidade neste trabalho.

E.1 Frota Municipal

Uma OSC, bastante atuante na área de controle social, realizou um pedido de informação sobre os veículos oficiais utilizados pelos funcionários da prefeitura. Gestores internos à CGM perceberam então que eles próprios não tinham controle sobre o uso dos veículos e decidiram desenvolver um sistema para coletar os dados. Esse, inicialmente, controlaria as viagens apenas de funcionários da própria CGM. Os dados seriam então publicados no portal de transparência da prefeitura já existente.

Como os funcionários da CGM atuam, entre outras coisas, no combate à corrupção, tendo sido inclusive ameaçados de morte ([CREDENDIO, 2013](#)), logo surgiu a preocupação de que a divulgação de tais dados poderia colocar em risco a vida desses funcionários. O histórico das viagens poderia ser usado para traçar padrões e prever onde um funcionário estaria em um determinado dia.

Para remediar tal situação, como já é previsto pela LAI ([BRASIL, 2011](#)), algumas viagens poderiam ser marcadas como sigilosas, não sendo divulgadas no portal. A coleta dos dados se daria na seguinte forma:

1. O requerente da viagem (funcionário da prefeitura) e o motorista preencheriam um formulário cada com dados sobre a viagem.
2. O requerente assinalaria também se a viagem é sigilosa ou não e, em caso afirmativo, a justificativa para tal, que deve ser uma das estabelecidas pela lei.
3. Os dados dos formulários seriam então transcritos para um sistema eletrônico que armazenaria todas as informações.

Caso a viagem tivesse sido marcada como sigilosa, os seguintes campos seriam omitidos nos dados finais publicados: motivo, origem, destino e utilizador.

O Colab-USP, discutindo a questão, planejou então que a base com todas as informações (sigilosas e não sigilosas) deveria ficar em um servidor diferente daquele que realmente disponibilizasse as informações à população. Dessa forma, em caso de invasão do servidor exposto à população, a chance de vazamento das informações sigilosas seria menor. Porém, após discussões com gestores mais altos na hierarquia, decidiu-se que a abertura dessa base não era prioridade no momento por uma série de motivos:

- Consideraram uma informação trabalhosa de coletar, já que os funcionários teriam que preencher mais um formulário.
- As informações poderiam não ser confiáveis, já que os formulários poderiam não ser preenchidos corretamente.
- Os gastos com os veículos representam uma fatia pequena do orçamento, não justificando o esforço de coleta dos dados.
- A publicação dos dados poderia levar ao uso pela mídia em denúncias sensacionalistas, não contribuindo para a discussão de políticas públicas, e focando em gastos pequenos enquanto o problema estaria nos grandes gastos.

A coleta e abertura dessa base ficou então suspensa por tempo indeterminado.

E.2 Imóveis

A prefeitura possui um arquivo com informações em papel sobre os imóveis de sua propriedade. Porém, o fato dessa base não estar em formato digital dificulta sua consulta e controle tanto pela própria prefeitura como pela população. Sendo assim, a destinação desses imóveis pode ficar comprometida.

Segundo conversas com gestores envolvidos, a linguagem usada nesses documentos não digitalizados é específica da área, requerendo um profissional especializado para obter as informações desejadas. O formato dos documentos mudou diversas vezes ao longo do tempo, não estando padronizados.

Tendo em vista a complexidade e o esforço necessários para converter a base inteira ao formato digital, planejou-se, pelo menos em um primeiro momento, transcrever apenas parte das informações encontradas nos documentos. A definição de quais serão essas informações ainda é uma questão em aberto.

Para começar a estudar possibilidades de digitalização dessa base, os gestores da COPI-CGM iniciaram um diálogo com o departamento responsável pela mesma. Porém, o departamento foi mudado de secretaria, indo da SEMPLA, com a qual a COPI-CGM já possuía certa proximidade, para a Secretaria Municipal de Desenvolvimento Urbano (SMDU). Apesar da mudança apresentar justificativas interessantes, a SMDU estava com seus esforços totalmente voltados ao desenvolvimento do novo Plano Diretor Estratégico do município de São Paulo, o que dificultou a realização de reuniões para discutir a digitalização da base de imóveis.

Paralelamente, começou-se a trabalhar com outra base também relativa a imóveis da prefeitura, a base de imóveis cedidos. Essa base, ao contrário da anterior, já encontra-se em formato digital, porém, não facilmente processável por máquina. Uma vez que um imóvel está sob a posse da prefeitura, ela pode destiná-lo para algum fim próprio, deixá-lo temporariamente sem destinação, ou cedê-lo para algum grupo ou pessoa temporariamente ou permanentemente. Os imóveis que estão atualmente cedidos constituem a base de imóveis cedidos. A cessão de um imóvel pode ser feita com ou sem contrapartida.

Um exemplo de cessão sem contrapartida poderia ser o uso de um imóvel municipal para abrigar uma escola estadual, ou talvez uma creche conveniada com a prefeitura. Já uma cessão com contrapartida poderia ser a destinação de um imóvel para uso privado, mediante pagamento de aluguel.

Quando há um pedido de destinação de um imóvel, é criado um processo que será avaliado por várias secretarias da prefeitura. Elas podem aprovar aquele pedido ou requerer outro uso para o mesmo imóvel. Esse processo pode levar alguns anos, sendo que, mesmo ao fim do processo, uma secretaria pode pedir para revê-lo.

A base de imóveis cedidos usada pela funcionária responsável é constituída de um conjunto de arquivos DOC. Apesar dessa não ser uma estrutura convencional do ponto de vista computacional, ela atende às necessidades da funcionária que a utiliza, uma vez que não há alterações concorrentes (apenas uma usuária altera a base) e as atualizações não são frequentes. Este fato corrobora com as afirmações de [Denis e Goëta \(2014\)](#), segundo as quais, uma base de dados, apesar de não estar em formato adequado para disponibilização, pode estar em formato apropriado para o uso que é dado dentro da administração pública.

Como os dados não estão organizados de forma que possam ser facilmente extraídos automaticamente, eles estão sendo extraídos manualmente por membros da COPI-CGM e colocados em uma planilha.

Com o intuito de aumentar as possibilidades posteriores de uso desses dados, pretende-se realizar a geolocalização dos mesmos, acrescentando-se a longitude e a latitude de cada imóvel à base. Como a base é muito numerosa, pretende-se realizar um processo automático e corrigir possíveis falhas manualmente.

Para a disponibilização final, planeja-se, além de busca textual nos dados dos imóveis, representá-los sobre o mapa da cidade de São Paulo, permitindo que os usuários realizem uma busca por região geográfica.

Essa representação geográfica dos dados, segundo um gestor, vai ao encontro da nova orientação que está sendo tomada pela prefeitura, na qual as decisões estão sendo descentralizadas, passando a ser pensadas no nível das subprefeituras.

Das bases em processo de abertura pela prefeitura, esta é a que talvez mais se assemelhe ao caso Bhoomi, uma vez que também lida com a questão da propriedade de terra. Nas palavras de uma gestora envolvida, “Isso é um vespeiro!”, se referindo a como está feita a atual cessão de imóveis municipais. Outro gestor afirmou que um relatório sobre o tema colocou as escolas de samba e os clubes de futebol como os que mais ocupam áreas públicas, porém, segundo o mesmo gestor, o relatório omitiu as áreas que são usadas para comércio e habitação, sendo que estas seriam muito maiores. Logo, há fortes evidências de que essa questão envolva interesses de grupos de grande poder econômico e político.

Um gestor e um acadêmico denunciaram ainda que esses grupos poderosos usam da atual desorganização da máquina pública para conseguir controlá-la de acordo com seus próprios interesses. A desorganização atual em que se encontra a base de imóveis da prefeitura seria um ambiente propício para tal prática. Além disso, tentativas de organizá-la poderiam gerar oposição por parte desses grupos.

Segundo um gestor, se a disponibilização de tal base possibilitaria submeter os dados ao escrutínio público, por outro lado, os grupos de maior poder econômico poderiam ter também melhores condições tecnológicas para se apropriar mais rapidamente da base. Sendo assim, é incerto se a abertura teria impactos sociais positivos ou negativos.

Segundo [Gurstein \(2011\)](#), para buscar o impacto positivo, a abertura deveria ser feita dando suporte para que os grupos menos favorecidos possam se apropriar dos dados e advogar seus próprios interesses.

E.3 Convênios

A prefeitura constantemente firma convênios com Organizações da Sociedade Civil (OSCs). Esses convênios podem visar diversas finalidades, como a manutenção de uma creche ou o atendimento à população em situação de rua, e geralmente envolvem repasse de dinheiro público. Porém, não existe um sistema interno na prefeitura que unifique o controle desses convênios e, além disso, não há uma norma que padronize como esse controle deve ser feito nas diversas secretarias municipais. Como consequência, cada secretaria acaba desenvolvendo seu próprio sistema de controle de convênios.

Visando aumentar o controle sobre o tema, a gestão atual da prefeitura está debatendo a criação de um decreto que sirva de marco regulatório municipal para o controle de convênios, normatizando e centralizando a prestação de contas. Por outro lado, há uma preocupação por parte das entidades conveniadas para que o controle não seja demasiadamente rígido, engessando-as e impossibilitando a prestação de serviço.

Após esse processo de definição do decreto, será implementado um sistema para coletar os dados e viabilizar o controle, tanto por parte da prefeitura como pela população. A ideia nesse caso é fazer com que o próprio sistema que coletará os dados primários seja responsável pela transparência dos mesmos, disponibilizando-os na web. Estando a preocupação com a transparência inclusa no próprio sistema usado internamente pela prefeitura, acredita-se que os dados finais disponibilizados serão mais fidedignos e atuais.

O novo sistema mapearia as transferências de verba para as entidades conveniadas, controlando tanto a coleta dos dados como a execução do serviço. Como dito anteriormente, o próprio sistema faria a disponibilização dos dados ao cidadão, para fins de controle social. Como o sistema de disponibilização estaria embutido no sistema interno de controle, os dados provavelmente seriam mais confiáveis. O sistema, de certa forma, seria semelhante ao [SICONV¹](https://www.convenios.gov.br), do governo federal. Porém, vale notar que foram levantadas, em um dos eventos realizados, diversas críticas ao SICONV, pontos que devem ser levados em consideração no desenvolvimento do sistema municipal.

Por ser um sistema que envolve repasse de dinheiro público e que seria usado por centenas de usuários, o Colab-USP não aceitou a responsabilidade do desenvolvimento do mesmo, porém, concordou em contribuir com as especificações técnicas, que serão então utilizadas no posterior desenvolvimento, possivelmente pela própria PRODAM.

¹ <https://www.convenios.gov.br>