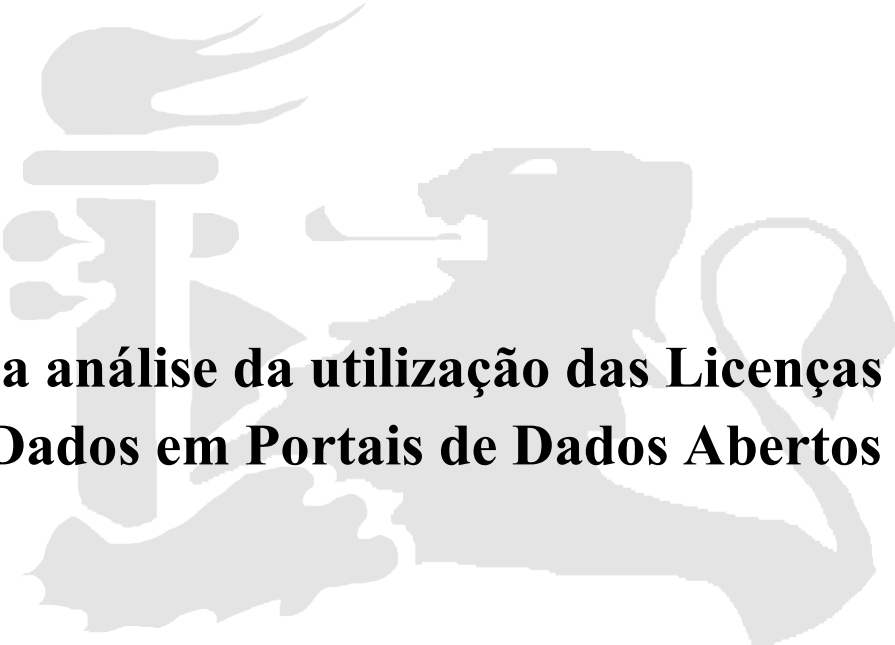




UNIVERSIDADE FEDERAL DE PERNAMBUCO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
CENTRO DE INFORMÁTICA
2017.1



Uma análise da utilização das Licenças de Dados em Portais de Dados Abertos

Aluna: Amanda Trigueiro de Araújo (ata3@cin.ufpe.br)

Orientadora: Bernadette Farias Lóscio (bfl@cin.ufpe.br)

Recife, 14 de julho de 2017



UNIVERSIDADE FEDERAL DE PERNAMBUCO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
CENTRO DE INFORMÁTICA
2017.1

Amanda Trigueiro de Araújo

Uma análise da utilização das Licenças de Dados em Portais de Dados Abertos

*Trabalho apresentado à disciplina de Trabalho de Graduação em
Sistemas de Informação do Centro de Informática da
Universidade Federal de Pernambuco como requisito parcial
para obtenção do grau de Bacharel em Sistemas de Informação.*

Orientadora: Prof^a. Dra. Bernadette Farias Lóscio

Recife, 14 de julho de 2017

Dedico este trabalho a Deus, minha família e
amigos que de alguma forma contribuíram
para realização deste trabalho.

Agradecimentos

Agradeço primeiramente a Deus que me concedeu capacidade e a Nossa Senhora que me concedeu paciência para chegar até aqui.

A toda minha família: pai, mãe, irmã, avós e padrinhos, que desde minha infância se sacrificaram para oferecer uma educação de qualidade pra mim. Hoje, vejo que foi um caminho árduo, porém gratificante. Agradeço em especial ao meu avô Armando, cujo nome carrego em sua homenagem e que hoje é um anjo no Céu, que sempre me disse que a única coisa que ninguém nunca pode tirar de nós é o conhecimento e que quanto mais a gente tiver, melhor. Ele sempre esteve certo.

A professora Bernadette e seu doutorando Lairson, que não me deixaram faltar apoio e orientação no decorrer deste trabalho.

Por fim, agradeço imensamente a Henrique e todos os meus amigos que de alguma forma contribuíram pela minha caminhada até aqui.

Este trabalho é a certeza do fim de um ciclo para o começo de outro.

Resumo

Utilizar a Web para a publicação de dados está se tornando uma prática cada vez mais estabilizada, comum, aceita e utilizada atualmente, seja por seus provedores, seja pelos seus consumidores. Dados são abertos quando qualquer pessoa pode livremente usá-los e reutilizá-los sem qualquer restrição legal, tecnológica ou social, estando sujeito a, no máximo, creditar a sua autoria e compartilhar pela mesma licença. As discussões sobre os usos das licenças e qual deve ser utilizada em cada conjunto de dados publicado ainda é incipiente, porém, é um ponto de extrema importância para fomentar a publicação de dados abertos. De uma maneira geral, espera-se, com este trabalho, apresentar os tipos de licenças existentes, facilitando e incentivando a utilização das mesmas no fornecimento de dados, apresentar processos sobre a escolha e publicação e observar como, atualmente, está a utilização das licenças em alguns portais de dados abertos.

Palavras-Chave: Licença de dados, Licenças Abertas, Dados Abertos

Abstract

Using the Web for data publishing is becoming an increasingly stable, commonplace, accepted and currently used practice, both by its providers and by its consumers. Data are opened when anyone freely use them and reuse them without any legal, technological or social restriction, being subject to at most credit its authorship and share by the same license. Discussions on the use of licenses and which should be used in each published dataset are still incipient, but it is an extremely important point to encourage the publication of open data. In general, it is expected that this work will present the types of existing licenses, facilitating and encouraging their use in data provision, presenting processes on the choice and publication, and observing how the use of licenses is currently in some open data portals.

Keywords: Data License, Open License, Open Data

Lista de Figuras

Figura 1. Código que consulta os Datasets existentes em um repositório e a saída retornando os atributos solicitados no código.	21
Figura 2. Gráfico de licença por dataset do Portal Brasileiro de Dados Abertos (valores em porcentagem).....	23
Figura 3. Gráfico de licença por dataset do Portal de Dados Abertos de Recife (valores em porcentagem).....	24
Figura 4. Gráfico de licença por dataset do Portal de Dados Abertos do Rio de Janeiro (valores em porcentagem).....	25
Figura 5. Gráfico de licença por dataset do Portal de Dados Abertos do Reino Unido (valores em porcentagem).....	27

Lista de Tabelas

Tabela 1. Comparativo entre licenças.....	17
Tabela 2. Licenças por dataset do Portal Brasileiro de Dados Abertos.	22
Tabela 3. Licenças por dataset do Portal de Dados Abertos de Recife.....	24
Tabela 4. Licenças por dataset do Portal de Dados Abertos do Rio de Janeiro.....	25
Tabela 5. Licenças por dataset do Portal de Dados Abertos do Reino Unido.	26

Sumário

1. Introdução	10
1.1 Motivação.....	10
1.2 Objetivos e Contribuições.....	12
1.3 Estrutura do Documento.....	12
2. Licença de Dados	13
2.1 Conceito Geral.....	13
2.2 Níveis de Licenças	14
2.3 Levantamento das Licenças Existentes.....	15
2.3.1 Creative Commons	15
2.3.2 Open Data Commons	16
2.3.3 Principais diferenças entre as licenças Creative Commons e Open Data Commons.....	17
3. Metodologia	19
3.1 Trabalhos Relacionados	19
3.2 Critérios de Análise	19
3.3 Seleção dos Portais	20
3.4 Extração dos Dados	20
4. Análise dos Portais de Dados Abertos.....	22
4.1 Portal Brasileiro de Dados Abertos	22
4.2 Portal de Dados Abertos do Recife	24
4.3 Portal de Dados Abertos do Rio de Janeiro	25
4.4 Portal de Dados Abertos do Reino Unido	26
5. Conclusão e Trabalhos Futuros.....	28
Referências	31

1. Introdução

Neste capítulo será feita uma breve introdução do trabalho, mostrando qual a motivação para o seu desenvolvimento, os objetivos e contribuições, além de uma descrição de como o documento está estruturado.

1.1 Motivação

Ao final da década de 1980, mais precisamente em 1989, o cientista americano Tim Berners-Lee criou um sistema que nasceu com a finalidade de ligar as universidades americanas entre si para que os trabalhos e pesquisas acadêmicas fossem utilizados mutuamente em um ambiente de contribuição dos lados envolvidos [1]. Essa invenção, nada mais foi, do que o que conhecemos hoje como WWW - *World Wide Web*.

A partir daí, surgiram os navegadores (Mosaic, NetScape e Internet Explorer são alguns exemplos), que eram utilizados para que os usuários visualizassem as páginas Web e pudessem navegar entre as mesmas, mas não existia a possibilidade de contribuir com conteúdo nas respectivas páginas. A interação era, portanto, limitada à visualização passiva de conteúdo.

A velocidade de navegação e a consequente exigência de dinamismo do usuário, fez com que novas tendências surgissem na Web por volta dos anos 2000, com o conceito de Web 2.0 - termo usado para designar uma segunda geração de comunidades e serviços oferecidos na internet, que tem como objetivo tornar o ambiente online mais dinâmico e fazer com que os usuários colaborem para a organização de conteúdo [2]. A base desta geração é estar conectado a algo ou alguém - sem se preocupar se este alguém está em um país distante ou se é seu vizinho. Neste contexto, os usuários começaram a gerar conteúdo de maneira independente por meio de sites, blogs ou redes sociais, como as pioneiras *Orkut* e *Twitter*, por exemplo [1].

Com tudo isso, a Web deixou de ser um lugar em que a informação já vinha pronta, para ser um meio de divulgação livre de conteúdos. Consequentemente, o volume de dados disponíveis na Web passou a crescer exponencialmente, principalmente dos dados provenientes das redes sociais. Diante desta realidade, podemos afirmar que a facilidade de compartilhamento de dados na Web tem impulsionado a publicação de Dados Abertos e ajudado, cada vez mais, a motivar e consolidar a Web como plataforma de compartilhamento de dados.

A *Open Knowledge Foundation* (OKF) define, em resumo, que dados são abertos quando qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los, sem qualquer restrição legal, tecnológica ou social, estando sujeito a, no máximo, creditar a sua autoria e compartilhar pela mesma licença [3]. Entretanto, apenas disponibilizar estes dados não é suficiente. Ainda existem diversos problemas a serem resolvidos e a falta de padrões para a prática de publicação e consumo torna o processo mais complexo.

O W3C (*World Wide Web Consortium*) - consórcio internacional no qual organizações filiadas, uma equipe em tempo integral e o público trabalham juntos para desenvolver padrões para a Web [4] - publicou recentemente uma recomendação intitulada *Data on the Web Best Practices* (DWBP), que tem como objetivo fornecer melhores práticas para a publicação e uso de dados na Web [5]. Assim, o DWBP pode ser utilizado por publicadores e consumidores de dados de forma a ajudá-los a superar os diferentes desafios enfrentados ao publicar e consumir dados na Web. Em outras palavras, por meio das suas 35 boas práticas, o DWBP elenca os principais desafios enfrentados no contexto de publicação e consumo de dados, como formatos de dados, acesso aos dados, licença dos dados, identificadores de dados e metadados. As boas práticas visam promover uma série de benefícios como a reutilização de dados e uma maior confiança nos dados, além de possibilitar um entendimento comum entre quem utiliza e quem publica os dados.

Dentre as 35 boas práticas citadas no documento, neste trabalho estamos interessados na boa prática que diz respeito às licenças de dados. Especificamente, a boa prática 4 que diz: fornecer informações sobre a licença de dados [5]. Uma licença é uma informação extremamente útil para ser anexada aos dados na Web, uma vez que, de acordo com o tipo de licença adotada pelo publicador, pode haver eventuais restrições no compartilhamento e reutilização de dados. Além disso, a respectiva boa prática enfatiza a importância de que tanto os seres humanos sejam capazes de compreender a licença de dados, quanto os agentes de softwares também possam detectá-la automaticamente [5].

No entanto, as discussões sobre os usos das licenças e qual deve ser utilizada em cada conjunto de dados publicado ainda é incipiente, porém, é um ponto de extrema importância para fomentar a publicação de dados abertos. Ao aplicar uma determinada licença, o autor está explicitamente expressando que direitos quer manter, ou abdicar, sobre a propriedade do seu trabalho e as condições em que os consumidores podem utilizar esse trabalho [6]. Para um projeto de dados abertos, a escolha da licença deverá ter em conta o potencial de exploração dos dados e o interesse do público-alvo. Neste contexto, esperamos contribuir

com a apresentação de um documento onde será possível encontrar padrões de licenças de forma eficiente e clara, por meio de um estudo detalhado sobre os tipos existentes e suas definições. Além disso, também será feita uma análise sobre quais as licenças estão sendo mais utilizadas atualmente em quatro Portais de Dados Abertos.

1.2 Objetivos e Contribuições

O objetivo deste trabalho é apresentar um estudo sobre os tipos de licenças de dados existentes, suas definições e descrições, quais as mais utilizadas e as principais diferenças entre elas. Também serão realizadas análises de alguns Portais de Dados Abertos para verificar se os dados disponibilizados estão sendo oferecidos com licenças. A partir das coletas, serão apresentadas as licenças que estão sendo mais utilizadas nestes portais.

Assim, como principais contribuições deste trabalho, podemos destacar:

- Estudo detalhado sobre os tipos de licença de dados.
- Apoio na disseminação da importância da utilização da licença de dados ao disponibilizar dados na Web.
- Mostrar, por meio de análises de Portais de Dados Abertos, quais licenças são mais utilizadas atualmente.
- Verificar se a licença de Domínio Público, onde o provedor de dados abdica de todos os direitos sobre sua criação, é a mais utilizada.

1.3 Estrutura do Documento

O restante deste documento está organizado da seguinte forma: O Capítulo 2 aborda a contextualização deste trabalho, apresentando o conceito geral de licença de dados, no contexto de Dados Abertos, quais os tipos de licenças existentes e principais diferenças entre elas. O Capítulo 3 refere-se à metodologia utilizada para a coleta dos dados nos portais de dados abertos, quanto aos critérios de análise, seleção dos portais e extração dos dados. O Capítulo 4, por sua vez, descreve as análises realizadas nos Portais de Dados Abertos, que irão mostrar quais as licenças que estão sendo mais utilizadas pelos provedores de dados. Por fim, o Capítulo 5 expõe a conclusão deste trabalho, além de sugestões para projetos futuros.

2. Licenças de Dados

Neste capítulo serão apresentados conceitos essenciais para o entendimento deste trabalho. Na seção 2.1 será apresentado o conceito geral de licença de dados, bem como os princípios para suas publicações. Na seção 2.2 serão apresentados os diferentes níveis de licenças existentes, no universo de Dados Abertos. Por fim, a seção 2.3 discorre sobre as mais populares licenças abertas existentes e as principais diferenças entre elas.

2.1 Conceito Geral

Quando alguém cria algo, um trabalho original, como uma arte para marketing ou uma fotografia, por exemplo, o seu autor possui automaticamente os direitos sobre o seu trabalho, podendo determinar de que forma este trabalho pode ser ou não utilizado por outros. Ao aplicar uma determinada licença, o autor está expressando explicitamente quais direitos quer manter ou abdicar sobre a propriedade do seu trabalho e as condições em que outros utilizadores podem manipulá-los. Dependendo dos direitos de propriedade que o autor está disposto a abdicar, são várias as opções de licenças disponíveis.

Na maioria das jurisdições, existem direitos de propriedade intelectual em dados que impedem que terceiros usem, reutilizem e redistribuam dados sem permissão explícita. Mesmo em locais onde a existência de direitos é incerta, é importante aplicar uma licença simplesmente por razões de clareza. Assim, ao disponibilizar dados, deve-se atribuir uma licença - e se desejar que os dados estejam abertos, isso é ainda mais importante [7].

No universo dos dados abertos, a licença serve para expressar claramente que o autor abdica de direitos de propriedade originais para dar a outros utilizadores a possibilidade de reutilizar, modificar e partilhar o seu trabalho. Além disso, a licença dos dados é essencial para garantir aos consumidores a clareza na utilização e exploração da informação disponibilizada [6]. Por fim, deverá respeitar um dos princípios-chave da abertura de dados: garantir uma utilização sem quaisquer restrições. Como a abertura é a essência de um projeto de dados abertos, sua licença também deve ter um perfil aberto. Ou seja, sendo desejado disponibilizar dados abertos, será sempre necessário aplicar uma licença aberta.

A escolha da licença deverá ter em conta o potencial de exploração dos dados e os interesses do público-alvo. Havendo várias opções, a escolha da licença deverá estar em linha com a estratégia de abertura de dados da organização e orientada pelas metas a alcançar, em

termos dos objetivos de utilização dos dados e do perfil dos utilizadores a serem conquistados.

Selecionada a licença mais adequada à estratégia e objetivos da iniciativa de dados abertos, a licença deve ser aplicada de forma clara, sem deixar dúvidas ao consumidor que quiser explorar e reutilizar os dados.

Segundo [6], é aconselhável adotar alguns procedimentos na publicação da licença, conforme abaixo:

- **Identificação visual:** O tipo de licença deve estar perfeitamente visível em todo portal ou página onde são disponibilizados os dados abertos, por meio de logos identificativos da licença ou com referência expressa. De preferência, a identificação da licença deve ter hiperligação para conteúdo descritivo dos termos da licença.
- **Página de termos e condições:** Os termos e condições da licença aplicada aos dados devem estar descritos, de forma clara e concisa, numa área dedicada. Se a disponibilização dos dados abertos estiver integrada ao web site da própria, a licença dos dados deve ser bem destacada de forma a não se confundir com termos e condições gerais do site/portal.
- **Clareza e simplicidade:** A descrição dos termos e condições deve ser o mais claro e simples possível, de forma a não confundir o consumidor.

2.2 Níveis de Licenças

Dentre as diversas categorias existentes, vamos apresentar as licenças de dados abertos em 4 níveis popularmente já consolidados: *public domain*, atribuição *by*, atribuição *by-nc* e atribuição *by-share-alike* [8].

A licença *Public Domain* é a licença que deve ser utilizada para quem pretende renunciar de todos os direitos sobre o seu trabalho, em todo o mundo [9]. É a que dá mais garantias de sucesso à utilização dos dados, pois, ao abdicar de todos os direitos de propriedade, ela dá mais garantias aos consumidores que querem explorar os dados para transformá-los em aplicações e serviços.

Nos termos da licença *atribuição by*, a utilização da obra é livre, e os consumidores podem fazer dela uso comercial ou criar obras derivadas a partir da obra original. É exigido apenas que seja dado o devido crédito ao seu autor [10].

De acordo com a licença *atribuição by-non-commercial (by-nc)*, o autor permite uma utilização ampla da sua obra, porém limitada pela impossibilidade de se obter através dessa utilização uma vantagem comercial [11]. É também essencial que seja dado o devido crédito ao autor da obra original. Exigir uma licença do tipo “atribuição” (que é utilizada no Portal Nacional de Dados Abertos – Dados.Gov, por exemplo), pode não ser tão atrativa para os utilizadores comerciais, porque obriga a fazer hiperligação (link) para o conteúdo original. Por outro lado, pode ser uma ótima opção para ajudar a promover a iniciativa *open data*.

Já quando um autor opta pela utilização da licença *atribuição by-share-alike*, ou simplesmente *by-sa*, ele exige que lhe seja dado crédito pela criação da sua obra e também que as obras que descenderem desta sejam licenciadas nos mesmos termos em que o foi a sua própria obra [12]. Esta licença é muitas vezes comparada com as licenças de software livre. É uma opção que também pode afastar os utilizadores comerciais e pessoas que querem fazer lucro com a utilização dos dados.

2.3 Levantamento das Licenças Existentes

Dentre as licenças abertas existentes, as mais populares são as *Creative Commons* e a *Open Data Commons*. Neste tópico, vamos explorar um pouco sobre cada uma delas e, ao final, expor as diferenças entre as mesmas.

2.3.1 Creative Commons

O *Creative Commons* (CC) é uma organização mundial sem fins lucrativos que permite o compartilhamento e a reutilização da criatividade e do conhecimento por meio do fornecimento de ferramentas gratuitas. O CC possui afiliados ao redor do mundo que ajudam a garantir que as licenças funcionem internacionalmente. As licenças e ferramentas *Creative Commons* foram projetadas especificamente para trabalhar com a web, o que facilita a busca, descoberta e o uso dos conteúdos oferecidos sob seus termos [13].

As licenças CC podem ser aplicadas a qualquer tipo de obra, incluindo: recursos educativos, músicas, fotografias, bases de dados, informações governamentais e do setor

público, entre outros. As únicas categorias de obras para as quais o CC não recomenda o uso de suas licenças são software e hardware de computadores. Também não se deve aplicar as licenças *Creative Commons* às obras que não são mais protegidas por direitos autorais ou que já estejam em domínio público. Em vez disso, é recomendado que as obras em domínio público mundial sejam sinalizadas com a Marca de Domínio Público.

As licenças *Creative Commons* permitem que um material seja compartilhado e reutilizado em termos flexíveis e juridicamente seguros. O projeto *Creative Commons* oferece um conjunto central de seis licenças de direito autoral. Por não existir uma “licença *Creative Commons*” única, é importante identificar qual das seis licenças está sendo aplicada ao material, qual das seis licenças foi aplicada ao material que pretende ser utilizado e, em ambos os casos, qual a sua versão específica [13].

Todas as licenças exigem que os usuários deem a atribuição (BY) ao criador quando o seu material é usado e compartilhado. Alguns licenciadores escolhem a licença BY, que exige a atribuição ao criador, como a única condição para a reutilização do material. As outras cinco licenças combinam Atribuição (BY) com um ou mais dos três elementos de licença adicionais, que citamos no tópico 2.2: *Non-Commercial* (NC), que proíbe o uso comercial do material; *No-Derivs* (ND), que proíbe o compartilhamento de alterações do material; e *share-alike* (SA), que exige que as obras derivadas do material sejam lançadas sob a mesma licença.

2.3.2 Open Data Commons

O *Open Data Commons* fornece um conjunto de licenças de dados abertos que permite que organizações e indivíduos tornem seus dados legalmente abertos de forma gratuita, para qualquer pessoa usar, reutilizar e redistribuir. As licenças do *Open Data Commons*, desenvolvidas em 2009, foram as primeiras licenças abertas projetadas especificamente para dados. Elas foram amplamente adotadas e utilizadas por usuários, incluindo alguns dos maiores e bem sucedidos projetos de dados abertos no mundo, como *Open Street Map* e *Farm Subsidy* [14].

Além disso, as licenças do *Open Data Commons* estimularam a mudança por parte de outros nesta área - por exemplo, inspirando o *Creative Commons* para atualizar suas licenças para cobrir dados (na versão *Creative Commons 4.0* lançada em 2013).

As licenças possuem as mesmas características das licenças dos *Creative Commons*, mudando apenas a nomenclatura, como segue abaixo [15]:

- **Public Domain Dedication and License (PDDL):** aplicável aos dados e a bancos de dados. O licenciante renuncia de todos os direitos, ou seja, coloca-os em domínio público.
- **Open Data Commons Attribution License:** aplicável aos dados e aos bancos de dados, mas com exigência apenas de atribuição de autoria.
- **Open Database License (ODbL):** pode ser utilizada em esquemas de bancos de dados, arquitetura da informação e na forma de organização dos dados. Exige atribuição de autoria e compartilhamento pela mesma licença.
- **Database Content License (DbCL):** mesmas exigências que a ODbL, mas aplicável ao conteúdo dos bancos de dados, ou seja, os dados em si;

A fim de facilitar o entendimento, a tabela abaixo faz um comparativo entre o nível de licença e a respectiva licença da Creative Commons e do Open Data Commons.

Tabela 1. Comparativo entre licenças.

Nível da Licença	Licença Creative Commons	Licença Open Data Commons
Todos os direitos abdicados	CC0	PDDL
Atribuição	CC-by	ODC-by
Atribuição – Partilha	CC-by-as	ODbL

2.3.3 Principais diferenças entre as licenças Creative Commons e as Open Data Commons

Antes de citarmos as principais diferenças entre as duas licenças mencionadas nos tópicos anteriores, é importante apresentar o conceito de direitos *sui generis* de banco de dados: esses direitos conferem aos fabricantes de banco de dados o poder de proibir a extração e reutilização de uma parte substancial de uma base de dados e são concedidos aos fabricantes de banco de dados que fazem um investimento substancial de tempo e recursos para criar o banco de dados [13].

A Licença *Open Database* e a Licença *Open Data Commons Attribution* são licenças projetadas especificamente para uso em bancos de dados e não em outros tipos de material. Há muitas diferenças entre essas licenças e licenças *Creative Commons*, mas o mais importante é estar ciente de relacionar o escopo e funcionamento da licença.

As licenças de *Open Data Commons* aplicam-se apenas aos direitos *sui generis* de banco de dados e quaisquer direitos de autor na estrutura de banco de dados, porém não se aplicam aos conteúdos individuais do banco de dados. A versão mais recente das licenças *Creative Commons*, por outro lado, aplica-se aos direitos de banco de dados *sui generis* e todos os direitos de autor e direitos conexos na estrutura de banco de dados, bem como o conteúdo.

Outra diferença importante é que as licenças de *Open Data Commons* podem criar obrigações contratuais mesmo em jurisdições onde os direitos de banco de dados não existiriam de outra forma e, portanto, a permissão de licença não seria necessária. A *Creative Commons* redigiu suas licenças para garantir que elas nunca imponham obrigações onde a permissão não seja obrigatória para a utilização o material licenciado [13].

3. Metodologia

O objetivo deste trabalho é realizar uma análise em alguns portais para identificar quais licenças estão sendo mais utilizadas atualmente.

A presente análise foi realizada em três etapas bem definidas: (i) definição dos critérios de análise, (ii) seleção dos portais de dados abertos, (iii) extração de dados e análise. Os resultados da análise serão apresentados no capítulo 4.

3.1 Trabalhos Relacionados

Existem alguns trabalhos que fazem um levantamento das licenças mais importantes, citadas no capítulo 2. Dentre eles, podemos citar o Guia de Dados Abertos, publicado pela Agência para a Modernização Administrativa [6]. Este documento é, como o próprio título infere, um guia para publicação de dados abertos: apresenta o conceito de dados abertos e seu potencial, como é possível abrir esses dados e expõe também sobre a reutilização dos mesmos. Neste último ponto, existem alguns sub-tópicos que falam sobre as licenças, processos de escolha e publicação.

Outro trabalho que também apresenta um estudo sobre as licenças é o *Licensing Open Data: A Practical Guide*, produzido por Naomi Korn e Charles Oppenheim [25]. Este foi desenvolvido para organizações que estão considerando os problemas associados ao licenciamento na abertura de dados e/ou querendo entender os termos em que eles podem usar dados que foram licenciados por terceiros. Fornece uma visão geral das várias questões legais que pode surgir no contexto de licenciamento de dados abertos, bem como diferentes tipos de licenças que estão disponíveis.

Entretanto, nenhum destes trabalhos apresenta uma análise das licenças que já estão sendo utilizadas nos Portais de Dados Abertos, apenas expõe as mais populares. Desta forma, queremos identificar quais licenças estão sendo mais utilizadas aos portais, de acordo com a metodologia que será apresentada neste capítulo.

3.2 Critérios de Análise

O que se pretende analisar após as coletas dos dados é a licença aplicada para cada conjunto de dados do respectivo portal de dados abertos. Assim, poderemos verificar quais licenças apresentadas no capítulo 2 estão sendo mais utilizadas atualmente.

3.3 Seleção dos Portais

A seleção das plataformas para análise foi feita com objetivo de explorar os dados nacionais federais e de duas grandes cidades do Brasil. Com isso, foi escolhido o Portal de Dados Abertos do Governo Federal Brasileiro para representar os dados no âmbito de dados federais. Para representar os dados a nível regional, foram escolhidos o Portal de Dados Abertos do Recife e o Portal de Dados Abertos do Rio de Janeiro. Já a nível internacional, foi selecionado o Portal de Dados Abertos do Reino Unido, pelo fato da iniciativa de dados abertos desta nação ser uma das mais consolidadas no mundo e é considerado por especialistas um exemplo a ser seguido por outros países [16].

3.4 Extração dos Dados

Nesta fase, buscou-se obter as informações das plataformas para analisá-los de acordo com os critérios previamente definidos.

Os dados foram extraídos utilizando o OurCrawler [17], um código em Python que utiliza o CKAN – a maior plataforma para portal de dados em software livre do mundo. O CKAN é uma solução completa e pronta para usar que torna os dados acessíveis e utilizáveis, provendo ferramentas para simplificar a publicação, o compartilhamento, o encontro e a utilização dos dados (incluindo o armazenamento de dados). Além do portal brasileiro, ele também é usado nos principais portais de outros países, como Canadá, Reino Unido, Estados Unidos, Irlanda, Austrália, Argentina e Alemanha [18].

```
def getDatasets(portalUrl, datasetDataFile):
    ckanVersion = getCKANVersion(portalUrl);
    results = getDatasetList(portalUrl, ckanVersion)
    datasetID = 0
    for datasetName in results:
        datasetID += 1;
        if (datasetID < 221) :
            continue

        datasetMetaData = getDatasetMetadata(portalUrl, datasetName, ckanVersion)

        if (datasetMetaData == -1 or getAttributeValue(datasetMetaData, 'num_resor
            with codecs.open('skipList.dat', 'a+', encoding='utf-8') as arq:
                arq.write(datasetName + "\n")
            continue;

        resources = datasetMetaData['resources'];

        datasetFolder = portalDatasetsFolders + "/" + str(datasetID).zfill(3);
        if not os.path.exists(datasetFolder):
            os.mkdir(datasetFolder)

        for resource in resources:
            text = "";
            text += tab(portalId);
            text += tab(portalName);
            text += tab(portalUrl);
            text += tab(datasetID);
            text += tab(getAttributeValue(datasetMetaData, 'name'));
            text += tab(getAttributeValue(datasetMetaData, 'title'));
            text += tab(getAttributeValue(datasetMetaData, 'license_title'));

            datasetUrl = resource['url'];
            fileName = datasetFolder + "/" + unicode(normalizeFileName(resource
            fileName = removeAccents(fileName)

            #Como nao vamos baixar o dataset, escrevi comentei a funcao de down
            #downloadFile(datasetUrl, fileName)
            #size = os.path.getsize(fileName)
            print '#'+ datasetUrl + '#'+
            try :
                #size = os.path.getsize(fileName)
                #downloadFile(datasetUrl, fileName)
                #size = os.path.getsize(fileName)
```

URL	ID	Nome	Título	Licença
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2012-11.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	7730	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2012-12.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	4556	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-01.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	4389	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-02.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	4556	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-03.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	8222	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-04.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	6732	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-05.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	9622	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-06.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	12520	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-07.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	7715	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-08.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons	10799	Licença A
http://dadosabertos.rio.rj.gov.br/api1746/apresentacao/csv/historico/1746_tipo242_2013-09.csv#	11	berta para Bases de Dados (00bL) do Open Data Commons		Licença A

Figura 1. Código que consulta os Datasets existentes em um repositório e a saída retornando os atributos solicitados no código.

Na figura 1, podemos ver o código do OurCrawler acessando o portal de Dados do Rio de Janeiro e retornando os atributos que foram solicitados (portalId, portalName, portalUrl, datasetID, name, title e license_title) em um arquivo de extensão .dat.

Os dados retornados no arquivo .dat foram extraídos para o Microsoft Excel para que se fosse possível gerar gráficos, a fim de facilitar a visualização da utilização das licenças. Os resultados serão apresentados a seguir, no capítulo 4.

4. Análise de Portais de Dados Abertos

Neste capítulo serão apresentados os resultados das análises feitas em cada um dos seguintes portais: Portal de Dados Abertos do Governo Federal Brasileiro, Portal de Dados Abertos do Recife, Portal de Dados Abertos do Rio de Janeiro e Portal de Dados Abertos do Reino Unido.

4.1 Portal Brasileiro de Dados Abertos

O Portal Brasileiro de Dados Abertos é a ferramenta disponibilizada pelo governo federal para que todos possam encontrar e utilizar os dados e as informações públicas. O portal preza pela simplicidade e organização para que o usuário possa encontrar facilmente os dados e informações que precisa. Também tem o objetivo de promover a interlocução entre atores da sociedade e com o governo para pensar a melhor utilização dos dados em prol de uma sociedade melhor [19].

Tabela 2. Licenças por dataset do Portal Brasileiro de Dados Abertos.

Rótulos de Linha	Contagem de DatasetID
Open Data Commons Open Database License (ODbL)	6478
License not specified	5685
Creative Commons Attribution Share-Alike	1760
Other (Public Domain)	1028
Creative Commons Attribution	955
Empty	570
Creative Commons Non-Commercial (Any)	471
Other (Open)	200
GNU Free Documentation License	12
Public Domain Dedication and License (PDDL)	4
Other (Attribution)	2
Total Geral	17165

Na coleta realizada, foram obtidos 17.165 datasets, com a seguinte distribuição de licença por portal, conforme tabela 2.

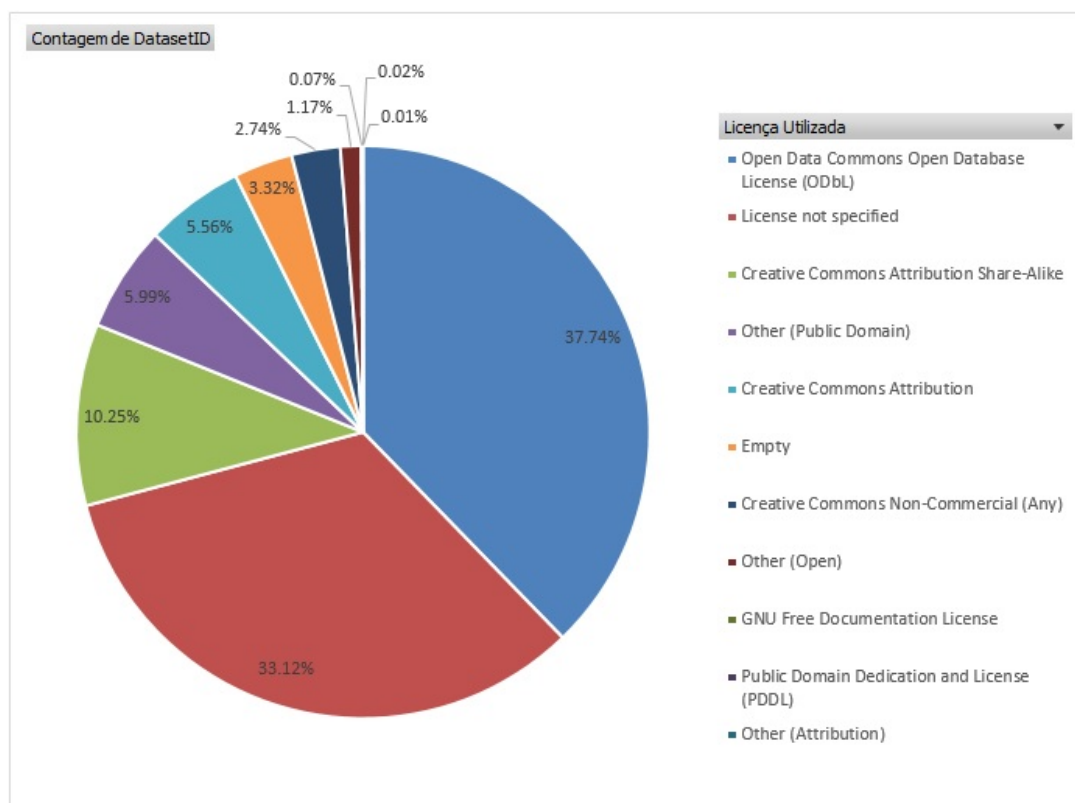


Figura 2. Gráfico de licença por dataset do Portal Brasileiro de Dados Abertos (valores em porcentagem).

Analisando os mesmo números da tabela, mas em gráfico com os valores em porcentagem na Figura 2, podemos perceber que com 37,74%, é predominante a utilização da licença Open Data Commons Open Database License (ODbL), mas que também existe um número muito alto (33,12%) de utilização de licenças não especificadas. As demais licenças aparecem em menor proporção.

Sendo assim, uma ação relevante para os administradores deste portal, seria realizar um trabalho de consciência entre os provedores de dados para que, de alguma forma, essas licenças não especificadas que estão sendo amplamente utilizadas sejam substituídas por licenças conhecidas para todos os consumidores de dados ou que seja reduzida consideravelmente a utilização de licenças não especificadas, pois isso pode trazer consequências, como a diminuição da utilização de dados deste portal, pela falta de uma licença reconhecida.

4.2 Portal de Dados Abertos do Recife

O Portal de Dados Abertos da Prefeitura da Cidade do Recife, desenvolvido pela EMPREL - Empresa Municipal de Informática - , tem o objetivo de disponibilizar de forma pública e fácil o acesso e a busca de dados governamentais gerados por secretarias e órgãos da gestão municipal. A publicação dos dados em formato aberto permite que qualquer um desenvolva aplicações ou visualizações, buscando facilitar a análise dos dados, promovendo a melhoria de serviços por meio da inovação e da criatividade, e contribuindo para uma maior participação da sociedade junto ao governo municipal [20].

Tabela 3. Licenças por dataset do Portal de Dados Abertos de Recife.

Rótulos de Linha	Contagem de DatasetID
Open Data Commons Open Database License (ODbL)	629
Total Geral	629

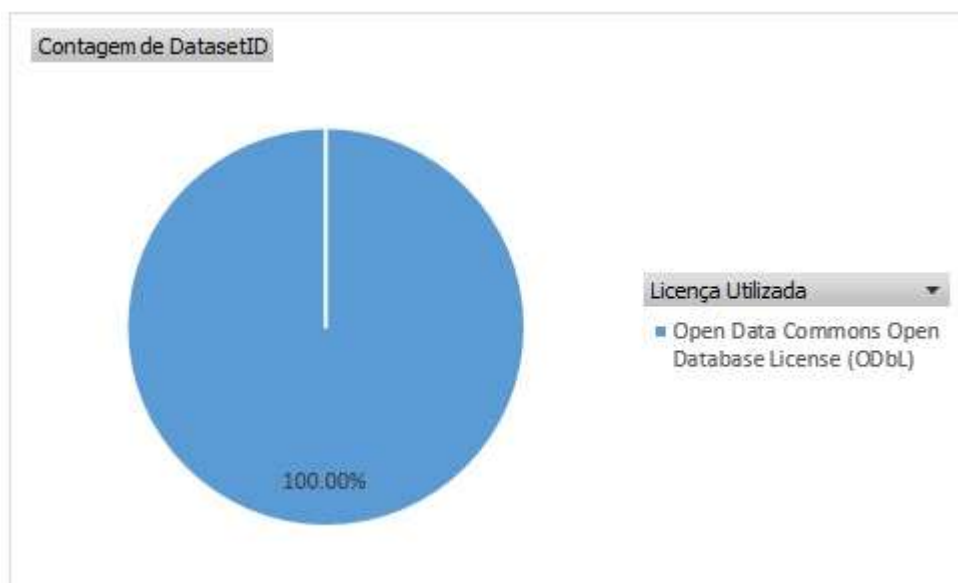


Figura 3. Gráfico de licença por dataset do Portal de Dados Abertos de Recife (valores em porcentagem).

Analisando os números da coleta na Figura 3, percebemos que em 100% dos casos é utilizada a licença Open Data Commons Open Database License (ODbL). Isto é uma evidência de que já deve existir uma padronização no portal para que todos os conjuntos de

dados disponibilizados nele estejam com a licença ODbL.

4.3. Portal de Dados Abertos do Rio de Janeiro

O Portal de Dados Abertos da Prefeitura do Rio de Janeiro é a ferramenta disponibilizada pelo governo municipal para que os desenvolvedores de aplicativos possam encontrar e utilizar os dados abertos governamentais, bem como outros da cidade do Rio de Janeiro. O portal disponibiliza o acesso às informações de saúde, educação, transporte, entre outras [21].

Tabela 4. Licenças por dataset do Portal de Dados Abertos do Rio de Janeiro.

Rótulos de Linha	Contagem de DatasetID
Open Data Commons Open Database License (ODbL)	11986
Empty	63
GNU Free Documentation License	12
Total Geral	12061

Na análise deste portal, obtivemos um número de 12.061 conjuntos de dados, conforme tabela 4. A partir dela, podemos concluir que a disponibilização de dados no Portal do Rio de Janeiro possui uma boa assiduidade. Se compararmos com o Portal de Dados Abertos do Recife, temos um acréscimo de mais de 1800% no número de datasets.

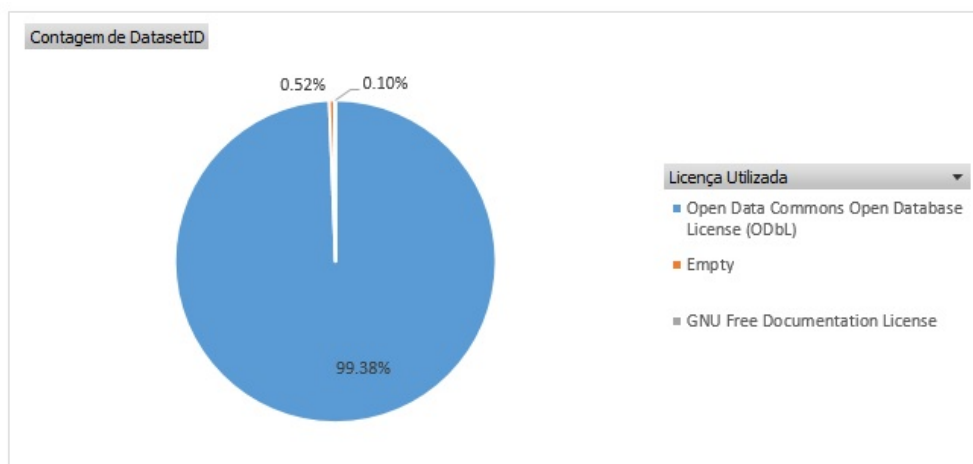


Figura 4. Gráfico de licença por dataset do Portal de Dados Abertos do Rio de Janeiro (valores em porcentagem).

Assim como no Portal de Dados Abertos da Prefeitura do Recife, percebemos que é predominante a utilização da licença Open Data Commons Open Database License (ODbL), com um número de 99,38% de aplicação desta licença aos datasets deste portal, de acordo com a Figura 4. Em um número bem inferior, existe também a utilização da licença *GNU Free Documentation* e uma quantidade de datasets sem licença aplicada, mas a predominância da ODbL é o que merece destaque, de fato. Isto é uma evidência de que já deve existir uma padronização no portal para que todos os conjuntos de dados disponibilizados nele estejam com a licença ODbL, da mesma maneira que acontece no Portal de Dados Abertos da Prefeitura do Recife.

4.4. Portal de Dados Abertos do Reino Unido

O governo do Reino Unido liberou os dados públicos para se tornar mais transparente e promover a inovação. Alguns desses dados estavam disponíveis antes, mas agora este portal os reúne em um site pesquisável. Fazer com que esses dados sejam facilmente disponíveis significa que será mais fácil para as pessoas tomar decisões e sugestões sobre políticas governamentais com base em informações detalhadas [22].

Tabela 5. Licenças por dataset do Portal de Dados Abertos do Reino Unido.

Rótulos de Linha	Contagem de DatasetID
UK Open Government Licence (OGL)	15714
Empty	1506
Creative Commons Attribution	28
Other (Not Open)	23
License Not Specified	18
Open Data Commons Public Domain Dedication and License (PDDL)	5
Total Geral	17294

Como já apresentado neste trabalho, o Portal de Dados Abertos do Reino Unido é um exemplo de portal a ser seguido pelos demais países e a iniciativa de dados abertos nesta nação já é extremamente consolidada. Prova disso é que, durante as coletas realizadas, este portal foi o que retornou um maior número de datasets se comparado aos anteriores.

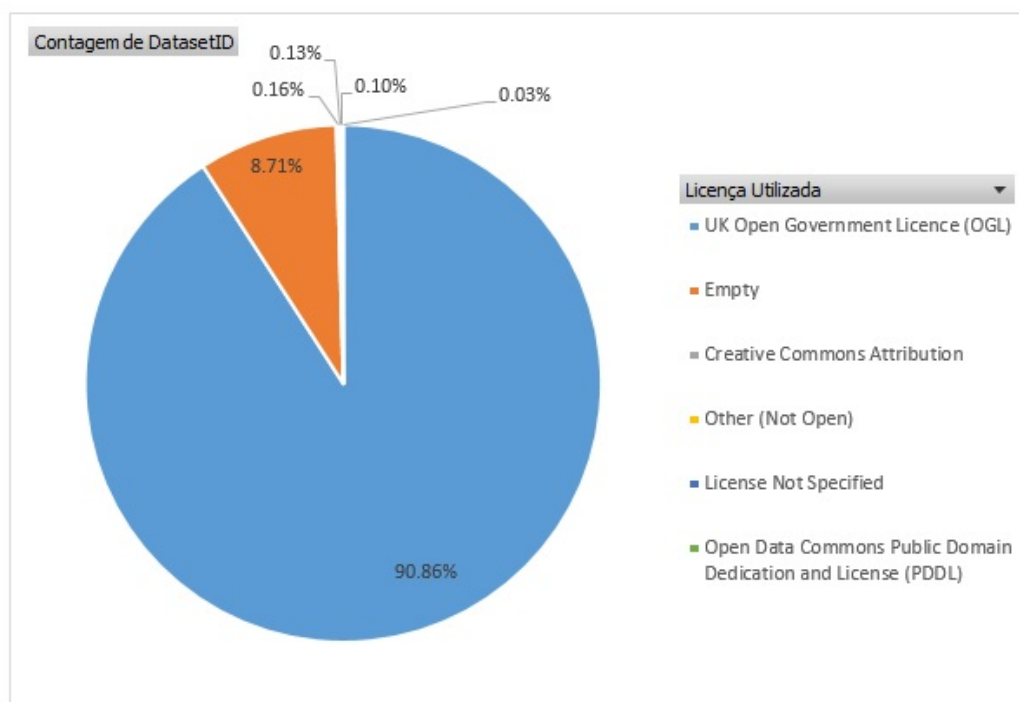


Figura 5. Gráfico de licença por dataset do Portal de Dados Abertos do Reino Unido (valores em porcentagem).

Neste portal, dentre as licenças utilizadas, temos uma diferença em relação aos demais: é utilizada uma licença aberta governamental, a *UK Open Government Licence*. Ela foi lançada para facilitar a reutilização do governo e outras informações do setor público [23].

Muitos países acabam definindo licenças próprias para uso padronizadas na abertura de dados governamentais, como os Estados Unidos e o Canadá, por exemplo, além do Reino Unido simplesmente para adicionar alguma marca ou para incluir referências à legislação local. Isto pode ser razoável, mas é desnecessário, pois existe o perigo de que a comunidade de dados abertos acabe em pior situação com dezenas de licenças ligeiramente diferentes com base em versões ligeiramente diferentes de algumas licenças básicas, como as apresentadas no capítulo 2 deste trabalho [24].

5. Conclusão e Trabalhos Futuros

Este trabalho dividiu-se essencialmente em três partes: um estudo sobre as licenças abertas para dados abertos; definição de uma metodologia para realizar análises nos Portais de Dados Abertos; e as coletas e análises dos dados obtidos, nos portais previamente definidos na fase anterior (metodologia).

No estudo realizado, foram destacadas as licenças viabilizadas pela *Creative Commons* e *Open Data Commons*. Chegou-se à conclusão de que as licenças disponíveis nessas plataformas possuem características bem específicas, porém semelhantes para cada caso, mudando apenas a nomenclatura da licença para cada plataforma a qual ela pertence. Como exemplo, podemos citar a licença que permite livre utilização dos dados, ou seja, onde o provedor dos dados renuncia totalmente de todos os direitos que ele poderia vir a ter sobre seus dados: para o *Creative Commons*, ela é chamada de CC0. Já para o *Open Data Commons*, a licença para o mesmo nível de utilização é chamada de *Public Domain Dedication and Licence* (PDDL).

Na metodologia, foram definidos quais critérios seriam coletados para análises, quais portais foram selecionados e de que forma se daria essa coleta. Dessa forma, o trabalho está totalmente passível de reprodução, pois possui, de forma genérica, um passo a passo, aplicável para quaisquer outros casos de pesquisa relacionados a este tema.

Por fim, pudemos perceber com as análises dos dados dos quatro portais, em um total de 47.149 datasets, que alguns deles estão extremamente padronizados com o tipo de licença que será aplicada aos conjuntos de dados daquele respectivo portal, como o Portal de Dados Abertos do Recife e do Rio de Janeiro, que utilizam quase que exclusivamente uma única licença para seus conjuntos de dados – a *Open Database Licence*, disponibilizada pelo *Open Data Commons*. Esta mesma licença também foi predominante para o Portal Brasileiro de Dados Abertos, com 37,74% de ocorrências, porém também existe um número muito alto de licenças não especificadas (33,12%) nos datasets deste portal. Até então, a grande maioria das licenças retornadas das consultadas aos portais tinham sido previamente estudadas no Capítulo 2 deste trabalho, pois são, de fato, as mais popularizadas.

Porém, ao analisarmos os dados do Portal de Dados Abertos do Reino Unido, nos deparamos com uma licença governamental, disponibilizada pelo próprio Reino Unido. Deste ponto, podemos concluir que as licenças já padronizadas apresentadas no Capítulo 2 estão sendo bifurcadas, seja para localizar a licença para se referir à legislação local ou ainda para

adicionar gestão de marcas (*branding*), como logotipos e nomes de cidades. Tudo isso está adicionando despesas indiretas em grande parte desnecessárias para os consumidores de dados que precisam entender se alguma dessas alterações é relevante, muitas vezes revisando o texto da licença. A proliferação de licenças adiciona complexidade aos reutilizadores, especialmente se as licenças forem incompatíveis. Como já citado neste trabalho, existe o perigo de que a comunidade de dados abertos possa acabar em uma situação com diversas licenças ligeiramente diferentes, com base em versões ligeiramente diferentes de algumas licenças básicas.

Outro ponto demonstrado com este trabalho foi que, mesmo a licença de Domínio Público sendo a que garante maior segurança aos seus consumidores devido ao fato do autor abdicar de todos seus direitos, ela é muito pouco utilizada pelos portais. Podemos inferir então que os provedores de dados estão poucos dispostos a cederem todos os direitos sobre seus dados. A grande maioria espera que no mínimo, seja creditada a autoria pelo seu trabalho.

Com a realização deste trabalho, algumas linhas de ação futuras podem ser traçadas. Dentre elas, podemos destacar as seguintes:

- **Ampliação do número dos portais analisados** – Ampliar a quantidade de portais a serem analisados, que não foram coletados durante a realização deste trabalho. Um exemplo de ação com este foco pode ser a análise de outros portais brasileiros, para tentar afirmar com mais precisão a padronização do tipo da licença utilizada nos portais de dados abertos dos estados e cidades espalhados pelo Brasil. Além disso, pode-se também analisar outros portais de dados abertos internacionais, como o dos Estados Unidos e do Canadá, que também possuem licenças governamentais. Com este estudo, também seria possível afirmar pontualmente em quais portais as licenças governamentais são mais utilizadas.
- **Comparativo entre a licença aplicada ao conjunto de dados e o número de downloads desse conjunto específico** – Pode-se realizar também uma análise comparativa entre a licença aplicada a cada um dos conjuntos de dados de um portal específico e o número de downloads desses conjuntos de dados, para que possa ser analisado qual o impacto que a disponibilização da licença (ou a

falta dela) causa no número de downloads de cada dataset.

- **Padronizar as licenças governamentais** – Como percebemos, não existe uma padronização para as licenças governamentais, ou seja, cada governo ou nação pode lançar a sua. Uma possível melhoria seria criar um tipo genérico de licença governamental, compatível com as licenças já difundidas como as apresentadas neste trabalho, deixando apenas alguns atributos modificáveis, onde cada país poderia definir e setar sua particularidade específica, evitando assim que tenhamos uma licença para cada país no futuro, o que dificultaria bastante a utilização dos dados pelos consumidores.

Em suma, este trabalho contribuiu para a disseminação da importância da utilização das licenças como boa prática para publicação de dados na Web. Mesmo as iniciativas de dados abertos sendo recentes, elas têm a sua devida importância nas diversas áreas em que são aplicadas e mesmo já estando bem desenvolvidas, é preciso que se sigam padrões para que os dados possam ser utilizados da maneira mais fácil e correta possível.

Referências

- [1] TecMundo. **A World Wide Web completa 20 anos, conheça como ela surgiu.** Disponível em: <<https://www.tecmundo.com.br/historia/1778-a-world-wide-web-completa-20-anos-conheca-como-ela-surgiu.htm>>. Acesso em 04 de maio de 2017.
- [2] Significados. **Web 2.0.** Disponível em: <<https://www.significados.com.br/web-2-0/>>. Acesso em: 24 de junho de 2017.
- [3] Open Knowledge Foundation. **What is Open?** Disponível em: <<https://okfn.org/opendata>>. Acesso em: 09 de abril de 2017.
- [4] W3C. **Sobre.** Disponível em: <<http://www.w3c.br/Sobre/>>. Acesso em: 15 de março de 2017.
- [5] W3C. Introduction. **Data on the Web Best Practices.** Disponível em: <<https://www.w3.org/TR/dwbp/>>. Acesso em: 09 de abril de 2017.
- [6] Agência para a Modernização Administrativa. **Guia Dados Abertos.** Disponível em: <http://www.dados.gov.pt/media/46120/guia_dados_abertos_ama.pdf>. Acesso em: 14 de abril de 2017.
- [7] Open Data Handbook. **How to Open up Data** - <<http://opendatahandbook.org/guide/en/how-to-open-up-data/>>. Acesso em: 28 de junho de 2017.
- [8] Open Definition. **Conformant Licenses.** Disponível em: <<http://opendefinition.org/licenses/>>. Acesso em: 06 de junho de 2017.
- [9] Creative Commons. **Public Domain.** Disponível em: <https://wiki.creativecommons.org/wiki/Public_domain>. Acesso em: 06 de junho de 2017.
- [10] Creative Commons. **Attribution 4.0 International - CC BY.** Disponível em: <<https://creativecommons.org/licenses/by/4.0/>>. Acesso em: 06 de junho de 2017.
- [11] Creative Commons. **Attribution-NonCommercial.** Disponível em: <<https://creativecommons.org/licenses/by-nc/2.0/>>. Acesso em: 06 de junho de 2017.
- [12] Creative Commons. **Attribution-ShareAlike.** Disponível em: <<https://creativecommons.org/licenses/by-sa/2.0/>>. Acesso em: 06 de junho de 2017.
- [13] Creative Commons. **O que é Creative Commons (CC)?** Disponível em: <<https://br.creativecommons.org/faq/>>. Acesso em: 15 de junho de 2017.
- [14] Open Knowledge. **Open Data Commons.** Disponível em: <<https://okfn.org/about/our-impact/legaltools/>>. Acesso em: 15 de junho de 2017.

- [15] GitHub. **Licenças de Dados Abertos.** Disponível em: <<https://github.com/dadosgovbr/course-opendata-license>>. Acesso em: 15 de junho de 2017.
- [16] Estadão. **EUA e Reino Unido: exemplos de sites de dados abertos.** Disponível em: <<http://politica.estadao.com.br/blogs/publicos/dados-abertos-nos-eua-e-no-reino-unido/>>. Acesso em: 02 de julho de 2017.
- [17] GitHub. **OurCrawler.** Disponível em: <<https://github.com/lerao/OurCrawler>>. Acesso em: 15 de junho de 2017.
- [18] Portal Brasileiro de Dados Abertos. **Sobre.** Disponível em: <<http://dados.gov.br/about>>. Acesso em: 02 de julho de 2017.
- [19] Portal Brasileiro de Dados Abertos. **Sobre o portal.** Disponível em: <<http://dados.gov.br/pagina/sobre>>. Acesso em: 22 de junho de 2017.
- [20] Portal de Dados Abertos da Prefeitura do Recife. **Sobre.** Disponível em: <<http://dados.recife.pe.gov.br/about>>. Acesso em: 22 de junho de 2017.
- [21] Portal de Dados Abertos da Prefeitura do Rio de Janeiro. **Sobre.** Disponível em: <<http://data.rio/>>. Acesso em: 22 de junho de 2017.
- [22] Portal de Dados Abertos do Reino Unido. **About.** Disponível em: <<https://data.gov.uk/about>>. Acesso em: 02 de julho de 2017.
- [23] Wikipedia. **Open Government Licence.** Disponível em: <https://en.wikipedia.org/wiki/Open_Government_Licence>. Acesso em: 06 de julho de 2017.
- [24] Open Data Institute. **The Proliferation of Open Government Licences.** Disponível em: <<https://theodi.org/blog/the-proliferation-of-open-government-licences>>. Acesso em: 06 de julho de 2017.
- [25] Naomi Korn and Professor Charles Oppenheim. **Licensing Open Data: A Practical Guide.** Disponível em: <http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf> Acesso em: 20 de julho de 2017.