**Paulo Henrique Cardoso**

# Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação

**Dissertação de mestrado**
**Setembro de 2019**

Universidade Federal do Rio de Janeiro - UFRJ
Escola de Comunicação - ECO
Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT

# Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação do Instituto Brasileiro de Informação em Ciência e Tecnologia e da Escola de Comunicação da Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Ciência da Informação.

Orientador: Prof. Dr. Marcelo Fornazin

RIO DE JANEIRO
2019

Paulo Henrique Cardoso
Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação/ Paulo Henrique Cardoso. – RIO DE JANEIRO, 2019-
110p. ; 30 cm.

Orientador: Prof. Dr. Marcelo Fornazin

Dissertação (Mestrado) – Universidade Federal do Rio de Janeiro – UFRJ
Programa de Pós-Graduação em Ciência da Informação, 2019.

1. Ciência de Dados. 2. Ciência da Informação. 3. Dados Governamentais Abertos. I. Prof. Dr. Marcelo Fornazin. II. Universidade Federal do Rio de Janeiro. III. Programa de Pós-Graduação em Ciência da Informação. IV. Ciência de Dados aplicada a Dados Governamentais Abertos sob a ótica da Ciência da Informação

Paulo Henrique Cardoso

# Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação do Instituto Brasileiro de Informação em Ciência e Tecnologia e da Escola de Comunicação da Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Ciência da Informação.

Trabalho aprovado em 12 de setembro de 2019

---

**Prof. Dr. Marcelo Fornazin(Orientador)**
PPGCI-IBICT/UFRJ

---

**Prof. Dr. Fabio Gouveia**
PPGCI-IBICT/UFRJ

---

**Prof. Dr. José Viterbo Filho**
PPGC/UFF

---

**Prof. Dr. Ricardo Medeiros Pimenta**
PPGCI-IBICT/UFRJ

---

**Prof. Dr. Gabriel Marcuzzo do Canto Cavalheiro**
PPGAd/UFF

*Dedico este trabalho a minha família, que me deu todo o suporte necessário para a elaboração deste trabalho, em especial aos meus filhos, Dante e Hugo. Também dedico aos docentes, discentes e técnicos do PPGCI UFRJ/IBICT que direta ou indiretamente me ajudaram e incentivaram.*

# Agradecimentos

Agradeço principalmente a minha esposa, Helen Pedroso, por durante o período do curso do Mestrado em Ciência da Informação ter me dado todo o suporte e incentivo para que eu completasse essa etapa da minha trajetória acadêmica. Além dela, agradeço ao Professor Dr. Marcelo Fornazin pela ajuda, aconselhamentos e paciência durante a execução da pesquisa e escrita da dissertação.

Por fim, o último agradecimento é voltado ao Programa de Pós-Graduação em Ciência da Informação – UFRJ/IBICT por prover todos os recursos materiais e pessoais para a prática do ensino e pesquisa em Ciência da Informação. Além do programa, agradeço os discentes, docentes e técnicos que atuam no PPGCI.

CARDOSO, Paulo Henrique. **Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação**. Orientador: Marcelo Fornazin. 2019. Dissertação (Mestrado em Ciência da Informação) - Escola de Comunicação, Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019.

# Resumo

Buscando entender as influências e contribuições da Ciência da Informação no campo recém-criado da Ciência de Dados, são apresentados os fundamentos teóricos da Ciência da Informação. São abordados aspectos como o Regime de Informação, que engloba temas legais como Acesso à Informação e legislações relacionadas e temas socioculturais como a Transparência e o Movimento Aberto desde a cultura até os Dados Governamentais Abertos. Como suporte aos fundamentos teóricos também são apresentados alguns dos aspectos técnicos relacionados, como o Ciclo de Vida dos Dados, Metodologia CRISP-DM, algumas das linguagens de programação utilizadas em Ciência de Dados e formato de arquivos. Como produto final deste trabalho são apresentadas duas análises de dados governamentais abertos, que ancoradas nos fundamentos teóricos, demonstram a aplicação das técnicas.

**Palavras-chave**: Ciência de Dados. Ciência da Informação. Dados Governamentais Abertos.

CARDOSO, Paulo Henrique. **Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação**. Orientador: Marcelo Fornazin. 2019. Dissertação (Mestrado em Ciência da Informação) - Escolade Comunicação, Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019.

# Abstract

Seeking to understand the influences of Information Science in the newly created field of Data Science, is presented the theoretical foundations of Information Science as the Information Regime that encompasses legal issues such as Access to Information and related legislations and sociocultural issues such as Transparency and Open Movement from culture to Open Government Data. In support of the theoretical foundations are also presented some of the related technical aspects such as Data Lifecycle, CRISP-DM Methodology, some of the programming languages used in Data Science and file format. As a final product of this work are presented two analyzes of open government data, which anchored in the theoretical foundations demonstrate the application of the techniques.

**Keywords**: Data Science. Information Science. Open Government Data.

# Lista de Figuras

# Lista de Quadros

# Lista de abreviaturas e siglas

| | |
|---|---|
| API | Application Programming Interface |
| BCI | Biblioteconomia e Ciência da Informação |
| BI | Business Intelligence |
| CDO | Chief Data Officer |
| CVD | Ciclo de Vida dos Dados |
| CVDGA | Ciclo de Vida dos Dados Governamentais Abertos |
| CD | Ciência de Dados |
| CI | Ciência da Informação |
| CAN | Civic Analytics Network |
| CRISP-DM | CRoss Industry Standard Process for Data Mining |
| TIC | Tecnologias da Informação e Comunicação |
| CRM | Customer Relationship Management |
| DA | Dados Abertos |
| DGA | Dados Governamentais Abertos |
| DBD | Decisão Baseada em Dados |
| DUDH | Declaração Universal dos Direitos Humanos |
| DOU | Diário Oficial da União |
| E-Gov | Governo Eletrônico |
| FOI | Freedom of Information |
| GA | Governo Aberto |
| ISP | Informações do Setor Público |
| LAI | Lei de Acesso à Informação |
| OEA | Organização dos Estados Americanos |

| | |
|---|---|
| OGP | Open Government Partnership |
| RI | Regime de Informação |
| SCM | Supply Chain Management |
| XML | eXtensible Markup Language |

# Sumário

# 1 Introdução

Os avanços das Tecnologias da Informação e Comunicação (TICs) têm conduzido a humanidade para uma sociedade informacional, revolucionando os valores sociais e culturais, além de promover reflexos na economia e política (CASTELLS, 2011).

Os governos estão, cada vez mais, utilizando dados em todos os aspectos de seu funcionamento. Com essa utilização, abre-se a oportunidade para aplicar as técnicas relacionadas à Ciência de Dados. Técnicas essas que podem ser utilizadas em diversas fases do processo de geração da informação, como: extração, limpeza, tratamento, persistência, interpretação e apresentação de *insights* a partir de dados estruturados ou não estruturados e que podem ser dados abertos ou não (MATHEUS; JANSSEN; MAHESHWARI, 2018).

Partindo desse pressuposto, as informações públicas, que são informações geradas e em posse de entidades governamentais, tornam-se um bem público, pelo potencial de geração de novos conhecimentos sobre as operações governamentais, a partir da análise e combinação de dados. Estas informações são públicas, de acordo com Lei 12.527 de 2011[1], conhecida como Lei de Acesso à Informação (LAI), que dispõe sobre o direito constitucional de acesso dos cidadãos às informações públicas.

Cotidianamente nos deparamos com o termo "Informação" sendo usado para descrever nosso mundo, como por exemplo: a era da informação, a indústria da informação, a sociedade da informação, entre outros. Isso leva ao crescente interesse na Ciência de Dados, que surge do reconhecimento de que os dados são os ingredientes básicos do conhecimento (STANTON, 2012).

A atitude de permitir o acesso aos seus dados, por parte das entidades governamentais, ocorre por meio de um processo chamado de Governo Aberto. Processo este que se caracteriza pela disponibilização na Internet das informações de domínio público para utilização, compartilhamento e alteração pela a sociedade em geral. Esta disponibilização deve seguir as licenças abertas Creative Commons[2] e/ou Open Data Commons[3], com o intuito de prover suporte legal para o reuso de informações públicas.

O processo de disponibilização de dados em formato e licença aberta denomina-se Dados Abertos, sendo estes dados provenientes de qualquer fonte (privada ou pública). Ao se tratar da abertura de dados de governo, a denominação se altera para Dados Governamentais Abertos. A divulgação desses dados tem o intuito de trazer transparência e promover uma maneira de auxiliar na prestação de contas no serviço público, por meio

[1] Lei n 12.527, de 18 de novembro de 2011<http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>

[2] <https://creativecommons.org>

[3] <https://opendatacommons.org>

da disponibilização de informações governamentais, em formato e licença aberta, tornando o conteúdo disponível a qualquer um, em qualquer lugar, para qualquer objetivo.

A utilização desses Dados Governamentais Abertos, muitas vezes, esbarra na dificuldade de lidar com grandes massas de dados e, também, nas especificidades das técnicas e metodologias para análise e mineração dos mesmos. Ao encontro disso, vem a Ciência dos Dados provendo o suporte ferramental e metodológico para execução de análises e estudos baseados em dados.

Considerando que, atualmente, grandes quantidade de Dados Governamentais Abertos são disponibilizados, nos mais diferentes formatos esta dissertação portanto buscou responder à seguinte pergunta: como técnicas e metodologias de Ciência de Dados podem ser utilizadas para realizar as análises e transformações de dados necessárias a fim de produzir informação sobre as atividades das organizações governamentais tendo como princípio o prisma da Ciência da Informação?

## 1.1 Objetivo

Assim, esta pesquisa tem por objetivo geral explorar a aplicação de técnicas de Ciência de Dados a Dados Governamentais Abertos, de forma que viabilize a transformação de dados em informação, possibilitando a transformação de informação em conhecimento por meio dos consumidores das informações geradas.

Para tanto, apresentará as bases teóricas que sustentam o regime de informação focado nos Dados Governamentais Abertos. Além disso, serão apresentadas aplicações práticas de técnicas de Ciência de Dados a Dados Governamentais Abertos, com o objetivo de propiciar a transformação de dados em informação.

Os objetivos específicos da pesquisa são:

- Levantamento bibliográfico, a fim de determinar os principais conceitos envolvidos neste trabalho;
- Coleta e processamento de Dados Governamentais Abertos da Câmara Federal e Senado por meio de API;
- Coleta e processamento de Governamentais Abertos do Diário Oficial da União por meio de download de lote de arquivos XML;
- Aplicação de ciência de dados para análise dos dados governamentais abertos assim ilustrando assim possibilidades de transformação de dados em informação.

## 1.2 Justificativa

Apesar de ser uma disciplina emergente, a ciência de dados representa uma nova corrente vital da escola de educação em Biblioteconomia e Ciência da Informação - BCI. Ainda mais, Ciência dos Dados e Ciência da informação são disciplinas gêmeas por natureza, por conta das semelhanças entre suas missões e tarefas, que muitas vezes se sobrepõem, além da possibilidade de serem complementares (WANG, 2018).

Essas relações de complementaridade e similaridade podem atingir o campo acadêmico, pois Wang (2018) argumenta que a academia de BCI deveria integrar ambas as disciplinas e desenvolver a ambidestria organizacional. Pois a Ciência da Informação pode fazer contribuições únicas à pesquisa em ciência de dados, incluindo concepção de dados, controle de qualidade de dados, biblioteconomia de dados e dualismo de teorias. Além de que a teoria de documentos, como uma direção promissora da ciência da informação unificada, deve ser introduzida na ciência de dados para resolver a divisão disciplinar.

Quando apresenta-se que a missão e natureza das disciplinas de Ciência de Dados e Ciência da Informação são fortemente relacionadas é com base nas exposições de Stanton e Bush. De acordo com Stanton (2012) a missão da Ciência dos dados é transformar dados brutos e bagunçados em conhecimento, já a missão da Ciência da Informação de acordo, Bush e Bush (1945) é tornar mais acessível uma desconcertante reserva de conhecimento. Em outras palavras, sua missão é efetivamente e eficientemente organizar e utilizar o conhecimento.

## 1.3 Metodologia e Estrutura da Dissertação

A proposta dessa dissertação é apresentar a pesquisa explicativa que busca entender a Ciência de Dados pela ótica da Ciência da Informação se utilizando da pesquisa bibliográfica e experimental como suporte.

De forma que a aplicação de técnicas e metodologias de ciência de dados na transformação de dados em informação, utilizando como plano de fundo o regime de informação. Este por sua vez, é propiciado pela conjunção de aspectos legais, socioculturais e técnicos.

Como metodologia bibliográfica, se tem os aspectos legais e socioculturais podem ser entendidos como as bases teóricas, essas bases teóricas são apresentadas como o Governo Eletrônico, Acesso à Informação, Transparência e Movimento Aberto. Estes, representam como a informatização do serviço público apoiado pelo direito de acesso a informação, motivado pela transparência e pelo senso de aberto (*open*) leva a abertura do governo e dados.

Ainda se tratando da pesquisa bibliográfica, como fruto o Governo Aberto, Dados

Abertos que quando idealizados conjuntamente se apresenta os Dados Governamentais Abertos. Esta abertura refere-se ao livre acesso a dados licenciados por meio de *Creative Commons* ou *Open Data Commons*, em formato estruturado e, preferivelmente, padronizado para facilitar seu uso e reuso como o arcabouço teórico.

Já na pesquisa experimental, são apresentados temas como a Ciência de Dados e temas relacionados como ciclo de vida de dados, que é um conjunto de processos e passos que devem executados desde o momento da criação ou coleta até o descarte dos mesmos. Além disso, é utilizada a metodologia CRISP-DM, que provê uma base técnica de passos e métodos de organização em um projeto de análise ou mineração de dados.

Além disso, foram realizadas duas analises de dados, foram realizadas duas analises para enriquecer os cenários da simulação e por consequência enriquecer a comparação da analise e possibilidades de exploração das relações entre Ciência de Dados e a Ciência da Informação.

Complementando a metodologia do trabalho, entende-se que o mesmo se trata de uma pesquisa explicativa por abordar a disciplina de Ciência de Dados que ainda está em desenvolvimento pela ótica de um campo cientifico já sedimentado, o da Ciência da Informação. O estudo da relações entre essas disciplinas tem o potencial de aproximar acadêmicos e profissionais das áreas estimulando uma troca de conhecimentos e métodos que enriquecem ambos os campos.

Esta dissertação está organizada da seguinte forma. Após esta Introdução, o Capítulo 2, apresenta os conceitos de Regime de Informação, Cultura Aberta e Dados Abertos, os quais embasam o atual momento da civilização ocidental que valoriza os dados abertos. O Capítulo 3, aborda especificamente os dados governamentais abertos os conceitos de governo e transparência que embasam tal abordagem. No Capítulo 4 apresentam-se os conceitos e técnicas de *Data Science*.

No Capítulo 5, articulam-se as teorias dos capítulos precedentes com procedimentos técnicos e a fim de se propor um método para processamento dos dados governamentais abertos. No Capítulo 6 realizam-se alguns ensaios e explorações com dados dos poderes Legislativo e Executivo Federal baseados na metodologia proposta a fim de se ilustrar as potencialidades e limitações dos ciclo de vida dos Dados Governamentais Abertos. Por fim, no Capítulo 6 são apresentadas as conclusões do trabalho.

# 2 Regime de Informação, Acesso, Transparência e Abertura

Neste capítulo são apresentados e discutidos os conceitos que permeiam o tópico que dá sustentação à pesquisa. Inicia-se descrevendo o conceito de regime de informação para, então, apresentar o governo eletrônico e o acesso à informação. Também se apresentam os conceitos de transparência e cultura aberta, demonstrando como os temas se relacionam e o quanto são importantes para a pesquisa.

## 2.1 Regime de Informação

Regime de Informação vem sendo abordado por diversos autores, geralmente como um recurso interpretativo para abordar as relações entre política, informação e poder. Neste trabalho serão abordadas, majoritariamente, as contribuições dos seguintes autores: (FROHMANN, 1995), (BRAMAN, 2004), (GOMEZ, 2002), (GOMEZ, 2012) e (GOMEZ; CHICANEL, 2013).

O termo Regime de Informação - RI foi proposto em 1984 e correlacionado pela primeira "regime de informação" ou o "regime global de política de informação" são conceitos que vêm sendo trabalhados na Ciência da Informação como uma forma de se obter uma paisagem do campo de ação da política de informação relacionando atores, tecnologias, representações, normas, e padrões regulatórios que configuram políticas implícitas ou explícitas de informação (MAGNANI; PINHEIRO, 2011).

Depois, em 1995, conceitua o RI como um sistema ou rede, mais ou menos estável, na qual a informação flui através de canais determináveis de produtores específicos, via estruturas organizacionais específicas, a consumidores ou usuários específicos (FROHMANN, 1995).

Em adição ao conceito demonstrado acima, Frohmann também fala de redes mais ou menos definidas, que emergem e se estabilizam, mesmo sem qualquer interferência governamental. Essas redes formam os fluxos de informação que podem ser de diversos tipos como cultural, acadêmico, financeiro, industrial, entre outros ou seus híbridos. Além disso, também discorre sobre as cinco limitações da Biblioteconomia e Ciência da Informação em lidar com política da informação (FROHMANN, 1995).

Nesse sentido podem-se elencar dois pontos importantes como síntese das proposições de Frohmann:

- O Regime de Informação surge como uma alternativa aos estudos de Política da

Informação e como uma crítica ao reducionismo praticado pela academia de BCI, dessa forma distanciando-se criticamente das abordagens reducionistas apresentadas por parte dos estudiosos da área;

- Proposição de um neo-documentalismo, a partir a transição do poder informacional das instituições, como Estados e editoras, para a produção e distribuição do conteúdo como escrita, discurso, rádio e internet.

Quase dez anos depois da publicação de Frohmann sobre o Regime de Informação, em 2004, Sandra Braman expande a discussão, trazendo o tema com uma mudança de termo, escopo e aplicação. O novo termo em tradução livre seria "Regime Global Emergente de Política de Informação". Dessa forma, amplia o alcance do regime para global por incluir instituições Estatais de todos os níveis e setor privado. Para Braman, regime pode ser entendido como um conjunto de regras, definições, práticas e processos, com poder normativo e regulatório internacional, podendo ser menos rígido e menos formal que um sistema jurídico. Já o adjetivo emergente remonta o fato de que o campo da política de informação no caso, objeto do regime em análise e suas características, ainda esteja em evolução (MAGNANI; PINHEIRO, 2011).

Grande parte das teorias que fundamentam os regimes internacionais é originária da Ciência Política. Essa característica dinâmica, repleta de rearranjos empíricos que geralmente vêm sendo tratados isoladamente, deveria ser tratada dentro da Política da Informação (BRAMAN, 2004).

Para Braman (2004), a teoria de regime pode complementar e contextualizar qualquer análise de política global de informação. Assim, ela não é uma substituta para o estudo da Política da Informação e sim uma fonte complementar de conhecimento por conta dos seguintes fatores:

1. Permite a identificação de padrões e tendências por meio de análise histórica;
2. Colabora na criação de novas instituições, ações e instrumentos políticos;
3. Unifica um domínio de tomada de decisão, evitando a dispersão gerada pela pluralidade de canais, de meios e de fluxos de informação;
4. Oferece novos parâmetros para estimar o impacto das tecnologias de informação sobre as relações internacionais.

Dessa forma, Braman (2004) apresenta a composição de regimes por:

- Governo: as instituições, regras e práticas formais de entidades geopolíticas com base histórica;

- Governança: as instituições formais e informais, regras, acordos e práticas de atores estatais e não estatais cujas decisões e comportamentos têm um efeito constitutivo sobre a sociedade;
- Governamentalidade: o contexto cultural e social a partir do qual os modos de governança surgem e pelos quais são sustentados.

A Autora Maria Nelida Gonzáles de Gómez expande o conceito de RI proposto por Frohmann em diversos trabalhos, começando em 2002, passando por 2012 e, por fim, em 2013.

> "... um conjunto mais ou menos estável de redes sociocomunicacionais formais e informais, nas quais informações podem ser geradas, organizadas e transferidas de diferentes produtores, através de muitos e diversos meios, canais e organizações, a diferentes destinatários ou receptores, sejam estes usuários específicos ou públicos amplos" (GOMEZ, 2002, p. 34).

Em 2012, Gonzáles de Gómez e Chicanel, apresenta uma discussão voltada para a transversalidade do Regime de Informação e como o termo remete às figuras contemporâneas de poder.

> "O conceito de regime de informação poderia formar parte de uma família de palavras que tematizam as configurações contemporâneas de práticas, meios e recursos de informação, onde as tecnologias da linguagem, caracterizadas por sua transversalidade e expansão indefinida, encontram seu espaço de operacionalização. O regime de informação, como conceito analítico, remete às figuras contemporâneas do poder, mas colocando em questão os critérios prévios de definição e reconhecimento do que seja juntamente da ordem da política e da informação" (GOMEZ, 2012, p. 43).

Por fim, em 2013, Gonzáles de Gómez acrescenta o sentido de atores em rede e como se dá sua relação.

> "O regime de informação remete à distribuição do poder formativo e seletivo de "testemunhos" sociais entre atores e agências organizacionais, setores de atividades, áreas do conhecimento, regiões locais e redes internacionais e globais, seja na medida em que definem, constroem e estabilizam as zonas e recursos de visibilidade social regulada, seja pela sonegação e/ou substituição de informações, seja por efeitos não totalmente intencionais que resultantes daqueles atos seletivos de inclusão/exclusão de atores, conteúdos, ações e meios. " (GOMEZ; CHICANEL, 2013, p. 4).

Com o intuito de apresentar os fatores, atores e fundamentos do Regime de Informação, os pontos apresentados a seguir sustentam o regime de informação. Por serem diretamente relacionados, Governo eletrônico, Acesso à informação, Transparência e Cultura aberta são bases para o entendimento do regime de informação abordado neste trabalho.

## 2.2 Governo Eletrônico - E-Gov

Governo eletrônico surgiu nos Estados Unidos da América em 1993, durante a chamada Iniciativa Nacional de Infraestrutura Informacional. Iniciativa essa que, como o nome sugere, tinha por objetivo incentivar a criação de empresas de base de Tecnologia da Informação e Comunicação (TIC), focadas na fabricação de hardware e desenvolvimento de softwares. Como resultado desse programa governamental de estímulo à indústria de TIC, juntamente com o advento da web, pessoas, empresas, universidades e governo se interconectam por meio de sistemas de informação (LASALA CALLEJA et al., 2014).

No Brasil, a primeira iniciativa de E-Gov teve início em março de 2000, cobrindo três das sete linhas de ação do Programa Sociedade da Informação: Universalização de serviços, Governo ao alcance de todos e infraestrutura avançada Jardim e Almeida (2012). Já Diniz et al. (2009, p. 26, apud Bresser-Pereira, (2002)): expõe que processo de transformação em E-Gov se deu com "O movimento conhecido por reformada gestão pública teve como cerne a busca da excelência e a orientação dos serviços ao cidadão. Esse movimento se baseou em princípios gerenciais voltados a resultados, eficiência, governança e orientação da gestão pública para práticas de mercado".

Ao se tratar de E-Gov o país considerado como expoente é a Estônia, que por conta disso é conhecida como e-Estônia. A trajetória para para se tornar referência em E-Gov começa em 1997 quando o chamado Ato do Banco de Dados foi adotado, o qual regula os bancos de dados digitais desde a criação até sua manutenção (KOTKA; VARGAS; KORJUS, 2015).

Atualmente a e-Estônia oferece cidadania eletrônica, conhecida como e-Residencia, em inglês *e-Residency*. A e-Residencia não pode ser encarada como um fenômeno isolado, pelo contrário, é produto de anos de experimentação e desenvolvimento do Governo Eletrônico Estoniano. Dois elementos foram fundamentais para a consolidação da chamada e-Estônia e o surgimento da e-Residencia. Primeiro o Sistema de código de identificação da Estônia e o segundo a infraestrutura tanto legal, política e tecnológica (KOTKA; VARGAS; KORJUS, 2015).

Dentre as definições de Governo Eletrônico disponíveis na literatura, todas convergem para a utilização das Tecnologias da Informação e Comunicação (TICs) na Administração Pública como pilar central do conceito. Lasala Calleja et al. (2014), propõe que a

utilização das TICs nos serviços públicos têm o potencial de causar impactos positivos como:

- Facilitar o acesso a informações governamentais, aos cidadãos e empresas;
- Melhorar a qualidade dos serviços pelo aumento da velocidade, integridade e processos;
- Proporcionar a oportunidade de participação democrática dos cidadãos.

Ainda sobre a conceituação do termo Governo Eletrônico, Alves (2012) apresenta uma compilação de definições que pode ser muito importante para entender os mais diversos aspectos e pontos de vista. Nela, ele traz a definição de Jardim e Almeida (2012, apud Balutisiv, (1999)) que defende que o Governo Eletrônico é a soma do Comércio Eletrônico (S) com o *Customer Relationship Management* (CRM), o *Supply Chain Management* (SCM), a Gestão do Conhecimento, o *Business Intelligence* (BI) e as Tecnologias Colaborativas.

Na compilação de Alves (2012), aparece também a definição do *Department of Information Resources, State of Texas, January* (2001), que defende o Governo Eletrônico como as atividades governamentais que ocorrem em comunicações eletrônicas entre todos os níveis de governo, cidadãos e comunidade empresarial, incluindo: aquisição e fornecimento de produtos e serviços, colocação e recebimento de pedidos, fornecimento e obtenção de informações e realização de transações financeiras

Em Zweers e Planque (2001), o Governo Eletrônico é defendido como um conceito emergente que objetiva fornecer ou tornar disponível informações, serviços ou produtos, através de meio eletrônico, a partir ou através de órgãos públicos, a qualquer momento, local e cidadão, de modo a agregar valor a todos os stakeholders envolvidos com a esfera pública. A aplicação da tecnologia da informação, combinada com as mudanças práticas nos órgãos públicos, tornam as operações governamentais mais responsivas, eficientes e transparentes, como defendido pelo *Commitee on Computing and Communications Research to Enable Better Use of Information Technology in Government, National Research Council* (2002).

Entretanto, o governo eletrônico não se restringe à incorporação de novas tecnologias para ampliar a capacidade de conexão entre governo e cidadão. As relações dentro do próprio governo também se reinventam, já que, nas suas mais diferentes instâncias, passa a atuar em rede. Cada poder, cada esfera e seus respectivos desdobramentos, trabalham como extensões, atuando como nós desta rede de governo. O advento do e-governo é resultado da aproximação dos nós entre todos os atores: governo eletrônico, cidadãos, empresas, terceiro setor(POMAR et al., 2003).

Trabalhando com ambas visões de Governo Eletrônico Pinho (2008) argumenta que Governo eletrônico se aplica a ampla adoção das TICs pelo setor governamental,

representado pela informatização do serviço público em suas atividades internas e externas. Um exemplo dessa informatização são os portais governamentais que servem como forma de comunicação do setor público com a população e forma de disponibilização de serviços e informações, dessa forma aumentando a transparência e participação da sociedade.

Ao observar as definições de Governo Eletrônico listadas acima, pode-se chegar à conclusão que as definições seguem dois tipos. Sendo o primeiro tipo, enviesado para a tecnologia e o outro que aponta para a inter-relação entre governo e sociedade. Dessa forma, o primeiro "Tecnologia" abrange tópicos como sistemas computacionais, artefatos de software, processos tecnológicos e hardware, com claro enfoque mecanicista. Já o segundo, foca nas relações governo sociedade, transparência, direito de acesso à informação e participação popular (ALVES, 2012).

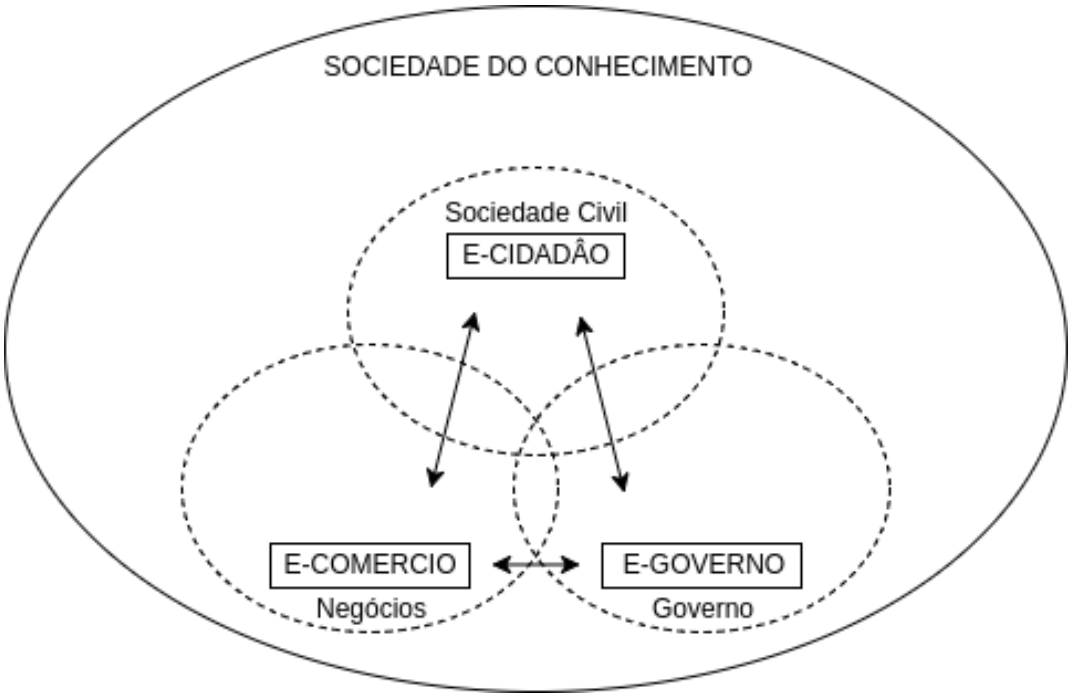
Os benefícios do E-Gov dependem diretamente da execução de operações de negócios eletrônicos (*E-Business*) que remetem ao conceito de comércio eletrônico (*E-Commerce*), essa inter-relação aponta para uma prestação de serviços em forma de transações eletrônicas. Apesar dessa aproximação de E-Gov de *E-Commerce*, o primeiro tem foco em melhorar o compartilhamento de informações, participação popular, disponibilização de serviços e a transformação dos relacionamentos internos e externos (FANG, 2002).

Partindo dessa inter-relação entre E-Government e E-Business/E-Commerce traz mais um elemento consigo o Cidadão eletrônico (E-Citizen), assim formando os três pilares da relação. Fang (2002) define esses três pilares como:

- E-Governo: trata das estruturas e processos que definem os relacionamentos entre órgãos, agências governamentais e entre funcionários e setor públicos como um todo (Legislativo, Executivo e Judiciário), em todos os níveis;
- E-Negócios: define as estruturas e processos que regem os relacionamentos entre governo, os mercados e indústrias setoriais, de forma geral, com o setor privado;
- E-Cidadão: trata das estruturas e processos que definem os relacionamentos entre a disponibilização de serviços governamentais, as expectativas do usuário e a relação entre países e instituições internacionais.

A expressão visual do relacionamento entre os pilares apresentados acima está representada no contexto da Sociedade do Conhecimento e pode ser examinada na Figura 1. O relacionamento entre os pilares gera diversos tipos de parcerias de governo eletrônico. Nesse sentido, Fang (2002), propõe oito modelos e seus benefícios: Governo para Cidadão - *Government to Citizen* (G2C); Cidadão para Governo - *Citizen to Government* (C2G); Governo para Negócios - *Government to Business* (G2B); Negócios para Governo - *Business to Government* (B2G); Governo para Funcionário Público - *Government to Employee* (G2E); Governo para Governo - *Government to Government* (G2G); Governo para Terceiro

Figura 1 – Relação entre pilares do governo eletrônico




Fonte: Adaptado de Fang (2002)

Setor - *Government to Nonprofit* (G2N);e Terceiro Setor para Governo - *Nonprofit to Government* (N2G).

Por este trabalho ser focado na utilização de dados abertos governamentais, as interações mais importantes para a discussão são:

- Governo para Cidadão - G(CAN, 2018)overnment - to - Citizen (G2C): Proporcionar a dinâmica para colocar serviços públicos on-line. Em particular, por meio da prestação de serviços eletrônicos para oferecer informações e comunicações;
- Cidadão para Governo - Citizen - to - Government (C2G): Proporcionar a dinâmica para colocar serviços públicos online, especialmente através da prestação de serviços eletrônicos para troca de informações e comunicação.

Essa interação entre governo e sociedade também é discutida por Clift (2004), onde é apresentado que o desafio do governo eletrônico não deve ser enfrentado apenas por tecnólogos (profissionais de tecnologia). Por se tratar de um problema político abordado por tecnologia, deve-se ser tratado por agentes públicos informados pela tecnologia e tecnólogos informados pela democracia. Dessa forma sendo possível criar uma era da informação que não apenas acomoda a vontade democrática do povo, mas também promove o bem público de maneira efetiva e sustentável.

Esses dois tipos de interação propiciam o acesso à informação na via governo - cidadão e na via cidadão – governo, que pode ser representada pela participação popular. Entretanto, para que essa relação seja verdadeira, efetiva e relevante, é essencial que todos tenham acesso à informação com transparência.

Esse é o conteúdo das seções que seguem, onde serão discutidas as conceituações e características do acesso à informação e, posteriormente, do Governo Aberto.

## 2.3 Acesso à Informação

O acesso à informação, nos moldes que conhecemos hoje, está diretamente vinculado à adoção da resolução 59 da ONU de 1946, que apresenta esse acesso como sendo um direito humano fundamental e pedra fundamental das liberdades nas quais a ONU é fundamentada. Além disso, o acesso à informação implica no direito de coletar, transmitir e publicar informações (ASSEMBLY, 1946).

Quando foi aprovado, juntamente como parte da Declaração Universal dos Direitos Humanos (DUDH) em Assembleia Geral da ONU, o Artigo 19 especificou os direitos de liberdade de expressão e de acesso à informação Mendel e Unesco (2008). Segue o Artigo 19 na íntegra: Todo ser humano tem direito à liberdade de opinião e expressão; este direito inclui a liberdade de, sem interferência, ter opiniões e de procurar, receber e transmitir informações e ideias por quaisquer meios e independentemente de fronteiras (ASSEMBLY, 1948).

Mendel e Unesco (2008) também aponta que apenas em 1966 o tratado legal internacional *The International Covenant on Civil and Political Rights* (ICCPR) foi assinado e em Julho de 2007, 160 países já tinham ratificado o acordo. Esse acordo garante o direito à liberdade de expressão e acesso à informação de forma similar ao Artigo 19 da DUDH.

Da mesma forma que a ONU, uma organização a nível global, as organizações regionais como Organização dos Estados Americanos (OEA), Conselho Europeu e União Africana também reconheceram formalmente o direito de acesso à informação (MENDEL; UNESCO, 2008).

A OEA adotou sua política de direitos humanos em 1969 e inclusa está o acesso à informação como direito humano fundamental, em seu Artigo 13 que dispõe sobre a o direito de liberdade de pensamento e expressão e onde está contido o direito de acesso à informação. O tratado foi assinado na Conferência Especializada Interamericana sobre Direitos Humanos, em 22 de novembro de 1969, conhecida como Pacto de San José. Segue trecho onde o direito ao acesso à informação é detalhado: Esse direito compreende a liberdade de buscar, receber e difundir informações e ideias de toda natureza, sem

consideração de fronteiras, verbalmente ou por escrito, ou em forma impressa ou artística, por qualquer outro processo de sua escolha (OEA, 1969).

Em Outubro de 2000 a OEA, em sua Comissão Interamericana de Direitos Humanos, aprovou a Declaração de Princípios sobre Liberdade de Expressão que apresenta diversos pontos relacionados com o acesso à informação, dentre eles serão apresentados alguns.

Logo no início do documento são apresentadas convicções e pressupostos, dentre eles um é relacionada ao acesso à informação: Convencidos de que, garantindo o direito de acesso à informação em poder do Estado, conseguir-se-á maior transparência nos atos do governo, fortalecendo as instituições democráticas (OEA, 2000).

Em seguida, no seu terceiro princípio, é exposto o direito de acesso às informações pessoais, como pode ser observado a seguir: Toda pessoa tem o direito de acesso à informação sobre si própria ou sobre seus bens, de forma expedita e não onerosa, esteja a informação contida em bancos de dados, registros públicos ou privados e, se for necessário, de atualizá-la, retificá-la e/ou emendá-la (OEA, 2000).

Já em seu quarto princípio, é detalhada a expectativa quanto às atuações no acesso às informações governamentais, como pode ser observado no trecho a seguir:

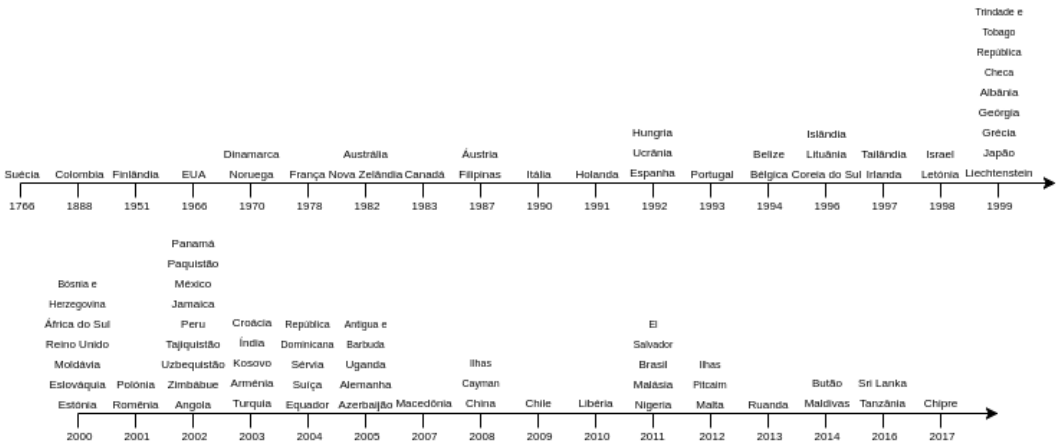
> "O acesso à informação em poder do Estado é um direito fundamental do indivíduo. Os Estados estão obrigados a garantir o exercício desse direito. Este princípio só admite limitações excepcionais que devem estar previamente estabelecidas em lei para o caso de existência de perigo real e iminente que ameace a segurança nacional em sociedades democráticas" (OEA, 2000).

Por fim, a OEA, em 2004, aprovou a Declaração de Nueva León. Nesta declaração, diversos Chefes de Estado reuniram-se e comprometeram-se a implantar as medidas acordadas, dentre elas está o acesso à informação. Segue trecho na íntegra:

> "O acesso à informação em poder do Estado, com o devido respeito às normas constitucionais e legais, incluindo aquelas sobre privacidade e confidencialidade, é condição indispensável para a participação do cidadão e promove o respeito efetivo dos direitos humanos. Comprometemo-nos a dispor de marcos jurídicos e normativos, bem como das estruturas e condições necessárias para garantir a nossos cidadãos o direito ao acesso à informação" (LEON, 2004).

Ao avaliarmos as implementações legislativas do direito de acesso à informação - *Freedom of Information* (FOI) em uma granularidade menor, podemos identificar que a adoção pelos países da Lei de Acesso à informação foi gradual. A Figura 2 mostra que o primeiro país a adotar uma Lei de Acesso à Informação foi a Suécia, em 1766, e com o

Figura 2 – Linha do Tempo da Adoção de Legislação FOI por Países




Fonte: Autor

passar do tempo, dezenas de outros países também adotaram legislação similar, até os dias atuais.

Conforme apresentado na Figura 2, a legislação brasileira de FOI, Lei n. 12.527, só foi aprovada em 2011, entrou em vigor em 2012 e ficou popularmente conhecida como Lei de Acesso à Informação. O Brasil tornou-se, assim, o 91o país do mundo e 13o país da América Latina a reconhecer e comprometer-se a promover o Direito de Acesso à Informação por meio de aprovação de legislação. Até 2015, na AL, apenas Venezuela e Costa Rica não tinham nenhuma legislação sobre Acesso à Informação (MICHENER; MONCAU; VELASCO, 2015).

É importante ressaltar que a Lei n. 12.527/2011, "Lei de Acesso à Informação" está em consonância com as decisões tomadas pela Organização dos Estados Americanos - OEA e Organização das Nações Unidas - ONU e aplica-se aos:

- Níveis governamentais: Federal, Estadual, Distrito Federal e Municipal;
- Poderes: Executivo, Legislativo e Judiciário;
- Entidades controladas direta ou indiretamente pelo Estado: Autarquias, Fundações e Empresas públicas, sociedades de economia mista e entidades sem fins lucrativos que recebam incentivos e/ou recursos governamentais.

Com o intuito de estabelecer os padrões Mendel (1999) propôs nove princípios sobre o direito de acesso à informação e eles são:

1. Divulgação máxima: as legislações que regulamentam o direito de acesso à informação devem ser guiadas pelo princípio da divulgação máxima. Esse princípio implica que

o âmbito do direito à informação deve ser amplo no que diz respeito ao leque de informações e organismos abrangidos;

2. Obrigação de publicar: órgãos públicos devem ser obrigados a divulgar informações de relevância pública. As instituições encarregadas de prover informações deveriam abrir as mesmas de forma proativa, sem a necessidade de um pedido formal de acesso;

3. Promoção de um Governo Aberto: instituições e órgãos públicos devem promover a abertura governamental. Dessa forma, é de extrema importância um amplo debate sobre a cultura do segredo sobre atividades governamentais;

4. Escopo limitado de exceções: os parâmetros necessários para uma exceção ser classificada devem ser claros e definidos precisamente. Assim, exceções ao direito de acesso à informação precisam ser cuidadosamente analisadas, para que não ocorra um desequilíbrio informacional;

5. Procedimentos e processos que facilitem o acesso: para que se possibilite um rápido e justo processamento de pedidos de acesso à informação, além de propiciar instâncias que permitam a reavaliação do pedido;

6. Custos: o alto custo não pode ser um impedimento para a requisição de informações, apesar de prover a informação ao cidadão onerar o erário público o equilíbrio entre custo de execução da atividade não pode impossibilitar a requisição da informação governamental;

7. Reuniões abertas: reuniões de instituições e órgãos públicos devem ser abertas à população, pois o acesso à informação não trata apenas do conteúdo de documentos, mas também daquele gerado em reuniões;

8. A divulgação tem precedência: outras leis que abordam o tema de acesso à informação devem estar em consonância com o 1o Princípio - Divulgação máxima ou ser revista e/ou revogadas;

9. Proteção para Denunciantes - *Whistleblowers*: pessoas que tornam públicas informações sobre atividades ilícitas e apresentam provas, segundo a legislação FOI, devem ser protegidas.

Depois, Mendel (2003) defende o acesso à informação de posse de autoridades públicas como um componente-chave para um governo transparente e accountable, exercendo um papel fundamental, por possibilitar que os cidadãos tenham visibilidade sobre assuntos internos dos governos. Desta forma, promove a redução da corrupção e mau uso de recursos públicos. Complementando seus pontos expostos em 2003, Mendel e Unesco (2008) apresenta que o progresso das TICs mudou inteiramente a forma como as sociedades

usam e se relacionam com a informação. Isso tornou o acesso à informação ainda mais importante para os cidadãos.

Ainda em sintonia com Mendel, quando relaciona Acesso à Informação com Transparência, Gruman (2012) apresenta que:

> "... acesso público à informação, ainda que primordial para a garantia de um Estado transparente e responsável, é instrumental no sentido de que os ganhos advindos das políticas de transparência governamental não se encerram em si mesmos, mas nos resultados trazidos por este tipo de política para a administração pública. A transparência e o acesso não garantem a eficácia do funcionamento da máquina pública, mas, pelo contrário, sua ausência, é garantia de mau uso dos recursos públicos porque livres de controle social. O acesso à informação é um instrumento, um meio para se alcançar um fim, a eficácia das políticas públicas"

## 2.4 Transparência

Colaborando com o apontamento de Mendel e Gruman, sobre a relação entre acesso à informação e transparência Birkinshaw (2006) aponta que acesso a informação é parte constituinte da transparência, mas que acarreta em conduzir os assuntos governamentais de forma aberta.

Conforme Zuccolotto, Teixeira e Riccio (2015) expõe o conceito acadêmico de transparência é fluido, isso se deve ao fato do conceito ter a potencialidade de ser utilizado em vários aspectos relacionados ao fluxo de informações. Além disso, os autores apresentam uma série de definições de transparência criadas em diversas áreas do conhecimento, como podemos verificar a seguir.

Para Black, Hashimzade e Myles (2012), a transparência pode ser entendida como oposição às políticas opacas, em que não se tem acesso às decisões, ao que elas representam, como são tomadas e o que se ganha ou se perde com cada uma delas. Já para Grigorescu (2003), ela está associada à divulgação de informações por parte dos governos para atores internos e externos.

Heald (2006), fala que a transparência é uma construção física, carregada de poder simbólico, independentemente de seu uso metafórico no discurso sobre as maneiras pelas quais o governo, os negócios e os assuntos públicos devem ser conduzidos. Segundo a International (2009), ela é característica de governos, empresas, organizações e indivíduos que são abertos e claros na disponibilização das suas regras informacionais, processos e ações.

Schultz (1999), defende que um país é transparente quando outro consegue obter

informações sobre as preferências da sua sociedade e o seu respectivo apoio às ações de governo. A abertura dos procedimentos de funcionamento para aqueles que não estão diretamente envolvidos (o público) demonstra o bom funcionamento de uma instituição, como defende Moser (2001). A transparência é tida como uma característica fundamental de um bom governo e um pré-requisito essencial para accountability, (MCGEE; GAVENTA, 2011).

A quando se trata da relação de transparência com outros temas, nota-se a inter-relação entre transparência, abertura e vigilância que é apontada como questão importante por Heald (2006), que além desse apontamento, o mesmo também destaca que não se pode estabelecer uma distinção entre transparência e abertura, em outras palavras, abertura está intimamente relacionado com a transparência. Essa proximidade entre abertura e transparência é reafirmada por Birkinshaw (2006) quando afirma de forma que, abertura é muito similar à transparência.

### 2.4.1 Características de Transparência

Após apresentar essa série de conceituações sobre transparência serão apresentadas as classificações de transparência, as classificações possíveis que serão analisadas neste trabalho são referentes as direções.

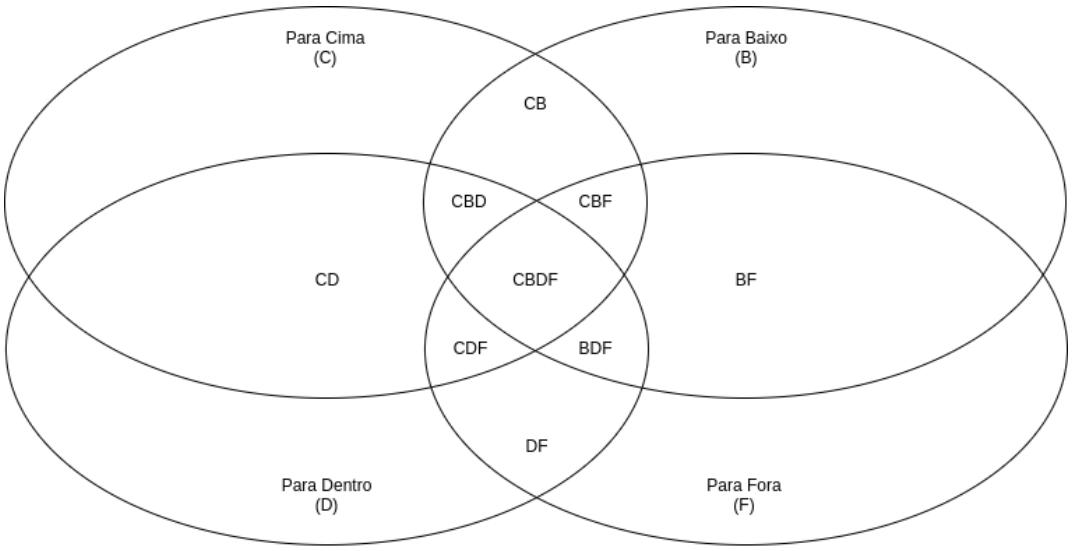
Heald (2006) apresenta sua versão das classificações como em direções, as direções podem ser representadas como apresentado na Figura 3 com o Diagrama Venn, que possibilita entender e discorrer sobre as intersecções entre as direções. São quatro as direções e essas direções podem ser classificadas como vertical e horizontal, a seguir as direções:

- Transparência para Cima (Vertical): demonstra relacionamentos hierárquicos, como de análise principal-agente, análise essa que é muito utilizada em modelagem econômica. A hierarquia apresentada é caracterizada pela possibilidade de monitoramento da conduta e comportamento em sistemas hierárquico superior/principal. Trata também de resultados dos subordinados em sistemas hierárquico hierarquias/agentes;

- Transparência para Baixo (Vertical): é quando os governados podem observar a conduta, comportamento e/ou os resultados dos seus governantes. O direito do governado em relação ao governante constitui uma democracia em teoria e prática e está diretamente relacionado com o *accountability*;

- Transparência para Fora (Horizontal): ocorre quando um subordinado hierárquico ou agente pode monitorar o que está acontecendo 'fora' da organização. A habilidade de observar o que ocorre fora da organização é de fundamental importância, pois

Figura 3 – Diagrama Venn Direções da Transparência




Fonte: Adaptado de Heald (2006)

possibilita entender o ecossistema onde a organização está inserida e monitorar os concorrentes;

- Transparência para Dentro (Horizontal): é quando agentes externos à organização podem observar o que ocorre dentro da organização. Esse tipo de transparência está relacionada diretamente com direito de acesso à informação.

No diagrama Venn acima (Figura 3), toda sobreposição de direções de mesma orientação (vertical e horizontal) representa um tipo de transparência simétrica. Por exemplo, (CB) representa uma transparência vertical simétrica e transparência horizontal assimétrica, por outro lado (DF) representa uma transparência horizontal simétrica e transparência vertical assimétrica. Por outro lado quando não ocorre uma sobreposição de direções da mesma orientação ocorre uma representação de transparência assimétrica. Desta forma (CBDF), representa uma transparência vertical e horizontal simétrica ou transparência completamente simétrica (ZUCCOLOTTO; TEIXEIRA; RICCIO, 2015).

## 2.5 Cultura Aberta

O ressurgimento do movimento ou cultura de aberto nos anos 80 se deu no meio da cultura de software e mais recentemente na cultura de rede, este movimento aberto se baseia em um novo conjunto de conceitos: colaboração, participação e transparência. Esse ressurgimento também pode ser entendido como a segunda vinda de aberturas que, no princípio dos anos 80 e 90, estavam direcionadas para discussões sobre sistema aberto e software aberto. Foco esse que depois foi generalizado e proliferado para outras áreas como

acesso, educação e cultura abertos, entre outros. Além dessas áreas, a política também foi afetada pelo surgimento de discussões e escritos sobre abertura, especialmente aqueles relacionados com a política institucional (TKACZ, 2012).

Transcorrendo sobre a generalização e proliferação da aplicação de abertura Powell (2015), apresenta o conceito de abertura mais focado à cultura aberta, mais especificamente o conhecimento aberto. Desta forma Powell defende um conceito de abertura como um valor intermediário entre código de software reutilizável, transparência institucional e oportunidades ampliadas de participação em culturas de produção de conhecimento.

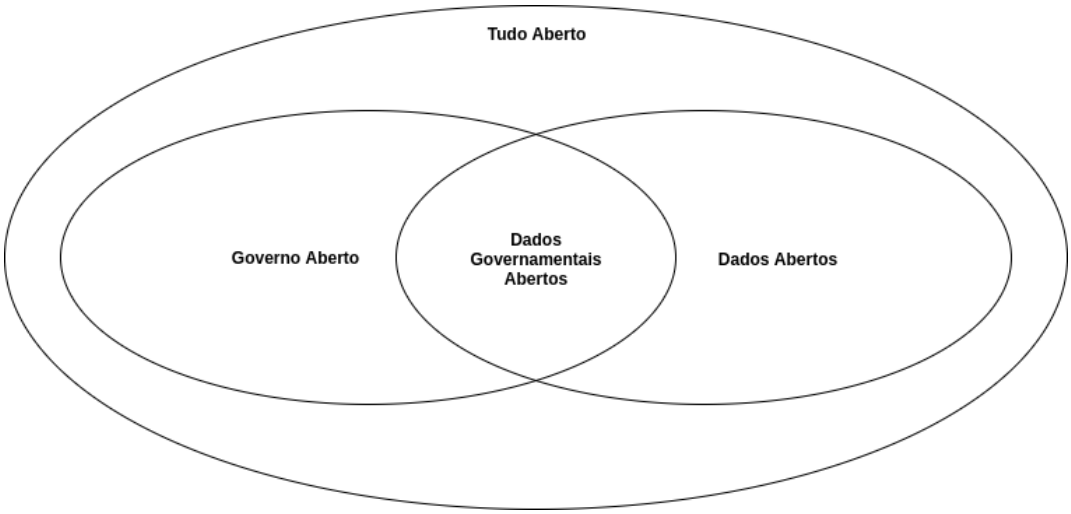
Quanto à definição de aberto, a *Open Knowledge Foundation* (OKF) um conceito (Versão 2.1) relacionado aos dados e conteúdo. Dessa forma, definindo de forma mais precisa o termo aberto ligado a dados abertos, conteúdo aberto e por consequência conhecimento aberto. Nesse sentido aberto pode ser definido como: aberto significa que qualquer um pode livremente acessar, utilizar, modificar e compartilhar com qualquer propósito OD (2019). E de acordo com o OKF (2018) Aberto tem três características e elas são:

- Disponibilidade e acesso: os dados devem estar disponíveis como um todo e não mais do que um custo razoável de reprodução, de preferência por download pela internet. Os dados também devem estar disponíveis de uma forma conveniente e modificável;
- Reutilização e redistribuição: os dados devem ser fornecidos em termos que permitam a reutilização e a redistribuição, incluindo a mistura com outros conjuntos de dados. Os dados devem ser legíveis por máquina;
- Participação universal: todos devem poder usar, reutilizar e redistribuir, não deve haver discriminação contra áreas de atuação ou contra pessoas ou grupos. Por exemplo, restrições "não comerciais"que impediriam o uso "comercial"ou restrições de uso para determinados fins (por exemplo, apenas em educação) não são permitidas.

Outro conceito importante apresentado por Tkacz (2012) é o tudo aberto - *open everything*, que pode ser entendido como uma conversa global sobre a arte, ciência e espírito de 'abertura'. Reunindo pessoas usando a modo aberto para criar e melhorar software, educação, mídia, filantropia, arquitetura, vizinhanças, locais de trabalho e a sociedade em que vivemos; tudo é sobre pensar, fazer e ser aberto. Além disso, o OFK afirma que a cultura aberta está mudando as regras do jogo e que Wikipédia e código aberto são um excelente exemplo disso.

Como apresentado, o movimento aberto é generalista e abrange diversas áreas. Dessa forma, a Figura 4 mostra que da intersecção de Governo Aberto e Dados Abertos surgem os Dados Governamentais abertos, e pode-se entender que a partir das características de cada um deles (Governo e Dados Abertos) forma-se um terceiro movimento, baseado

Figura 4 – Sobreposição dos termos abertos




Fonte: Autor

na abertura das informações governamentais em formato padronizado, de modo que seja legível por máquina.

# 3 Governo, Dados e Dados Governamentais Abertos

Este capítulo apresenta uma discussão mais aprofundada sobre o conceito "aberto" para áreas como governo e dados abertos, além disso, também apresenta como esses dois conceitos são base para os Dados Governamentais Abertos, que o objeto alvo do estudo.

## 3.1 Governo Aberto

Historicamente num primeiro momento o advento de governo aberto está ligado ao Iluminismo, com a Suécia em 1766 por meio do Ato de Liberdade de Imprensa a despontar como o primeiro país a adotar uma legislação que tendo base a esse conceito, fato que está diretamente relacionado com o direito de acesso à informação e transparência (LINDERS; WILSON, 2011).

Por outro lado alguns autores apresentam que o conceito de Governo Aberto está relacionado com a evolução do Governo Eletrônico, como apresentado por Albano (2014). Diametralmente, ele é oposto à ideia de que governo aberto seria uma evolução de governo eletrônico. NEVES JÚNIOR (2013) afirma que governo aberto ou governo 2.0, como é chamado por ele: "...vai além de uma mera evolução dos conceitos de governo eletrônico, constituindo-se em uma abordagem de atuação do governo totalmente direcionada ao cidadão, valendo-se das tecnologias comuns de mercado para promover tal direcionamento e aproximação".

A *Open Government Partnership* (OGP) é uma organização internacional que seu comitê diretor é composto por representantes de governos e organizações da sociedade civil em número igual de representantes, dessa forma promovendo e assegurando a participação social, além de rotacionar a liderança do grupo. A instituição foi lançada em 2011 para promover uma plataforma internacional para organizações nacionais que atuam na abertura governamental, em sua fundação era composta por 8 países crescendo para mais de 70. Para se tornar membro os países devem endossar a Declaração de Governo Aberto, apresentar um plano de ação e se comprometer com a fomentação de relatórios independentes sobre os avanços. E ao endossar esta Declaração, os países se comprometem a fomentar uma cultura global de governo aberto que capacita e entrega para os cidadãos e promove os ideais de um governo aberto e participativo do século XXI (OGP, 2018).

Segue alguns dos pontos mais importantes da declaração de governo aberto proposto pela OGP (2018) em tradução livre:

- Reconhecimento que pessoas em todo o mundo estão exigindo mais abertura no governo. Eles estão pedindo maior participação cívica nos assuntos públicos e buscando maneiras de tornar seus governos mais transparentes, responsivos, responsáveis e eficazes;

- Reconhecemos que os países estão em estágios diferentes em seus esforços para promover a abertura no governo, e que cada um de nós persegue uma abordagem consistente com nossas prioridades e circunstâncias nacionais e as aspirações de nossos cidadãos;

- Aceitamos a responsabilidade de aproveitar este momento para fortalecer nossos compromissos de promover a transparência, combater a corrupção, empoderar os cidadãos e aproveitar o poder das novas tecnologias para tornar o governo mais eficaz e responsável;

- Defendemos o valor da abertura em nosso compromisso com os cidadãos para melhorar os serviços, gerenciar os recursos públicos, promover a inovação e criar comunidades mais seguras. Adotamos princípios de transparência e de governo aberto com vistas a alcançar maior prosperidade, bem-estar e dignidade humana em nossos próprios países e em um mundo cada vez mais interconectado;

- Países que assinam esta declaração se comprometem a:

  – Aumentar a disponibilidade de informações sobre atividades governamentais;

  – Apoiar a participação cívica;

  – Implementar os mais altos padrões de integridade profissional em todas as nossas administrações;

  – Aumentar o acesso a novas tecnologias para abertura e responsabilidade.

### 3.1.1 Definição Governo Aberto

Quando se trata de definir ou conceituar Governo Aberto, é possível perceber tendências dentre a grande variedade de definições e conceituações. Dentre elas, algumas se destacam pela frequência com que são invocadas, elas seriam: a relação com transparência e/ou acesso à informação, além da relação com governo eletrônico e/ou avanços nas TICs e internet. Quanto à classificação sobre origem da definição, Chatwin e Arku (2017) defendem que as conceituações de Governo Aberto podem ser agrupadas, quanto a sua origem, como acadêmicas ou praticantes.

As definições de origem praticantes aparecem em publicações como a da OECD (2008), que defende que Governo aberto e responsivo se refere à transparência das ações

do governo, a acessibilidade dos serviços e informações governamentais e a capacidade de resposta do governo a novas ideias, demandas e necessidades.

Segundo a Chatwin e Arku (2017) o conceito de governo aberto possui três significados:

- Transparência da Informação: quando a população consegue perceber e entender as atividades de seu governo;
- Engajamento público: a capacidade do público de fazer parte do funcionamento de seu governo, influenciando os processos de políticas governamentais e programas de prestação de serviços;
- Responsabilidade: quando a população é capaz de responsabilizar o governo por sua política e desempenho de prestação de serviços.

Já segundo o Chatwin e Arku (2017), o governo aberto é um motor essencial do desenvolvimento no século XXI, pois traz maior transparência, participação do cidadão e colaboração entre as duas partes. Ele agrega benefícios significativos, como permitir o fortalecimento da democracia, dar maior voz à população, aumentar a eficiência da máquina pública e promover a integridade nas instituições, criando oportunidades econômicas que beneficiam a todos. E dá um exemplo claro quando fala de seus objetivos de criar um nível de abertura sem precedentes no governo. Trabalharemos juntos para garantir a confiança do público e estabelecer um sistema de transparência, participação pública e colaboração. A abertura fortalecerá nossa democracia e promoverá eficiência e eficácia no governo.

É importante ressaltar que o Governo aberto não trata apenas da transparência das informações, mas também da abertura em termos interativos. Esse paradigma coloca governo e população no mesmo nível, interagindo face a face. Essa estrutura se destaca, justamente, pela oposição ao governo "tradicional", onde as instituições estão hierarquicamente acima dos cidadãos, decidindo sobre as políticas e serviços unilateralmente (CHATWIN; ARKU, 2017).

Lathrop e Ruma (2010) apresenta o Governo aberto como uma instituição que aproveita o poder da colaboração da massa, valoriza a transparência e não se isola no poder decisório. Assim sendo, representa uma administração pública onde os cidadãos não apenas têm acesso a informações, documentos e procedimentos, mas também podem participar de uma forma significativa.

Assim como sintetiza Sandoval-Almazan e Gil-Garcia (2016),o governo aberto pode ser entendido como uma estratégia tecnológica e institucional que transforma o governo informação da perspectiva de um cidadão; Os cidadãos podem proteger, reutilizar, colaborar ou interagir com informações e dados de várias formas; e, como resultado dessa

transformação, os cidadãos têm o poder de examinar as decisões e ações dos funcionários públicos para aumentar a transparência e a prestação de contas e, consequentemente, propor diferentes alternativas para os serviços públicos e outras ações governamentais.

### 3.1.2 Princípios e Formas de Enxergar o Governo Aberto

Ao discorrer sobre os princípios do Governo aberto é importante destacar o conjunto de princípios, proposto por Linders e Wilson (2011) que são apresentados no Quadro 1. Onde de forma didática o autor demonstra quais são os objetivos e motivações de cada princípio e como estes propiciam o accountability por meio da transparência, engajamento dos cidadãos por meio da participação cívica e oferta colaborativa de serviços e parceria interagências por meio da colaboração.

Quadro 1 – Princípios do Governo Aberto

| | Transparência | Participação Cívica | Colaboração |
|---|---|---|---|
| Objetivo | Promover a prestação de contas e fornecer informações aos cidadãos sobre o que o governo está fazendo. | Melhorar a eficácia do governo e melhorar a qualidade da tomada de decisões. | Envolver os cidadãos no trabalho de seu governo, colaborando em todos os níveis e com organizações sem fins lucrativos, empresas e indivíduos. |
| Motivação | As informações mantidas pelo governo federal são um ativo nacional. | O conhecimento é amplamente disperso na sociedade o governo deve aproveitar este conjunto mais amplo de conhecimento. | Parcerias e cooperação melhoram a eficácia do governo. |

Fonte: Adaptado de Linders e Wilson (2011)

As formas de enxergar o Governo Aberto podem ser entendidas com lentes, como proposto por Linders e Wilson (2011, p. 264). Devido à variedade de definições apresentadas anteriormente, são necessárias classificações (LINDERS; WILSON, 2011).

- Lente da Transparência (Os defensores): promovida majoritariamente por instituições como (OMBWatch, a Sunlight Foundation, o OpentheGovernment.org, o Citizens for Responsibility and Ethics em Washington), além de verificadores e fatos independentes. Essa lente define “abertura” como transparência deliberada ou, pelo menos, anti-sigilo no funcionamento do governo;
- Lente Tecnológica (Os Futuristas): promovido por influentes instituições como (O’Reilly Media, Sunlight Labs, Crisis Commons, Center for Democracy and Technology, Google, and Participatory Politics Foundation), inspirados por movimentos

como software livre e código aberto defendem a tecnologia como meio para promover abertura;

- Lente da e-Democracia (Os Engajadores Cívicos): promovida por defensores de engajamento cidadão e democracia direta como instituições como (Open Society Institute, America Speaks e Personal Democracy Forum) por meio de fortes parcerias governamentais com os cidadãos. Defensores da e-Democracia apoiam o governo aberto como a pedra angular de uma sociedade mais democrática, fornecendo as ferramentas e a capacitação que ajudam o público a participar ativamente nas conversas e nos processos de tomada de decisão do governo;

- Lente da Conformidade (Os Burocratas): promovida por pessoas instituições ligadas ou interessadas no cumprimento de mandatos e normas por meio de dados concretos, como, por exemplo, órgãos federais, implementadores do governo e avaliadores públicos.

Ainda por conta da ambiguidade trazida pelo termo Governo Aberto, vêm sendo discutidas formas de avaliar e enxergar. Linders e Wilson (2011) apresentam uma matriz de relacionamento entre seus princípios e lentes propostos. Essa matriz representada no Quadro 2, onde podemos visualizar o alinhamento entre Princípios e Lentes onde "X" representa um relacionamento primário e "-" representa um relacionamento secundário.

Quadro 2 – Alinhamentos entre Princípio e Lente

| | Transparência | Tecnologia | e-Democracia | Conformidade |
|---|---|---|---|---|
| Transparência - Accountability | X | | - | - |
| Transparência - Reuso Público | - | X | | - |
| Participação - Engajamento Cidadão | - | | X | - |
| Participação - Cidadão Fonte | | X | - | - |
| Colaboração - Oferta Colaborativa de Serviços | | X | X | - |
| Colaboração - Parceria Interagências | | - | | X |

Fonte: Adaptado de (LINDERS; WILSON, 2011)

## 3.2 Dados Abertos

A junção do Movimento Aberto com a Cultura de Dados torna viável o surgimento de um movimento chamado de Dados Abertos - DA ou do inglês, *Open Data*. Este movimento DA, caracteriza-se, de forma geral, com as seguintes proposições:

- Acesso aberto: acesso livre e gratuito;
- Origem diversa: os dados podem ser provenientes de vários setores, indústrias e instituições;
- Interoperabilidade: padronização dos formatos para facilitar o reuso.

Partindo disso uma série de definições foram propostas para o termo Dados Abertos. Segundo a OD (2019), Dados abertos são aqueles que podem ser usados livremente, compartilhados e incorporados por qualquer pessoa, em qualquer lugar, para qualquer finalidade.

Gartner (2019), fala que DA são informações ou conteúdos disponibilizados gratuitamente para uso e redistribuição, sendo apenas necessária a referência autoral. A expressão ainda pode ser usada, de forma casual, para identificar quaisquer dados que sejam compartilhados fora da organização e além de seu uso pretendido original.

Já a conceituação de OKF (2010), fala que:

> "Os dados estão abertos se puderem ser livremente acessados, usados, modificados e compartilhados por qualquer pessoa para qualquer finalidade - sujeitos apenas, no máximo, aos requisitos para fornecer atribuição e / ou compartilhar da mesma forma. Especificamente, os dados abertos são definidos pela Definição Aberta e exigem que os dados sejam A. Legalmente abertos: isto é, disponíveis sob uma licença aberta (dados) que permite a qualquer pessoa acessar livremente, reutilizar e redistribuir B. Tecnicamente aberto: isto é, que os dados estejam disponíveis para não mais do que o custo de reprodução e em formato legível por máquina e a granel."

É importante ressaltar, como defende Manyika et al. (2013), que os Dados abertos não são exclusivamente aqueles divulgados pelos Governos, mas podem vir de qualquer instituição. O W3C (2011) conceitua o termo como "todo o conjunto de dados que podem ser publicados na web, não apenas dados governamentais". A classificação requer ainda que ele precisa estar disponível para todos, sem restrições.

Quando perguntado por que algum dado deveria ser aberto, devem-se levar em consideração três razões comuns para a abertura de dados. Essas razões de acordo com OKF (2018) são:

- Transparência: Em uma sociedade democrática que funcione bem, os cidadãos precisam saber o que seu governo está fazendo. Para fazer isso, eles devem poder acessar livremente dados e informações do governo e compartilhar essas informações com outros cidadãos. Transparência não é apenas sobre acesso, mas também sobre compartilhamento e reutilização - muitas vezes, para entender o material, ele precisa ser analisado e visualizado, e isso requer que o material seja aberto para que possa ser usado e reutilizado livremente;

- Liberando valor social e comercial: Na era digital, os dados são um recurso fundamental para atividades sociais e comerciais. Tudo, desde encontrar a agência postal local até a criação de um mecanismo de pesquisa, exige acesso aos dados, muitos dos quais são criados ou mantidos pelo governo. Ao abrir os dados, o governo pode ajudar a impulsionar a criação de negócios e serviços inovadores que ofereçam valor social e comercial;

- Participação e engajamento: governança participativa ou para negócios e organizações engajados com seus usuários e público-alvo. Na maior parte do tempo, os cidadãos só podem se envolver esporadicamente com seu próprio governo - talvez apenas em uma eleição a cada 4 ou 5 anos. Ao abrir os dados, os cidadãos podem ser mais diretamente informados e envolvidos na tomada de decisões. Isso é mais do que transparência: trata-se de criar uma sociedade de "leitura / gravação" completa - não apenas sobre o que está acontecendo no processo de governança, mas também como contribuir para isso.

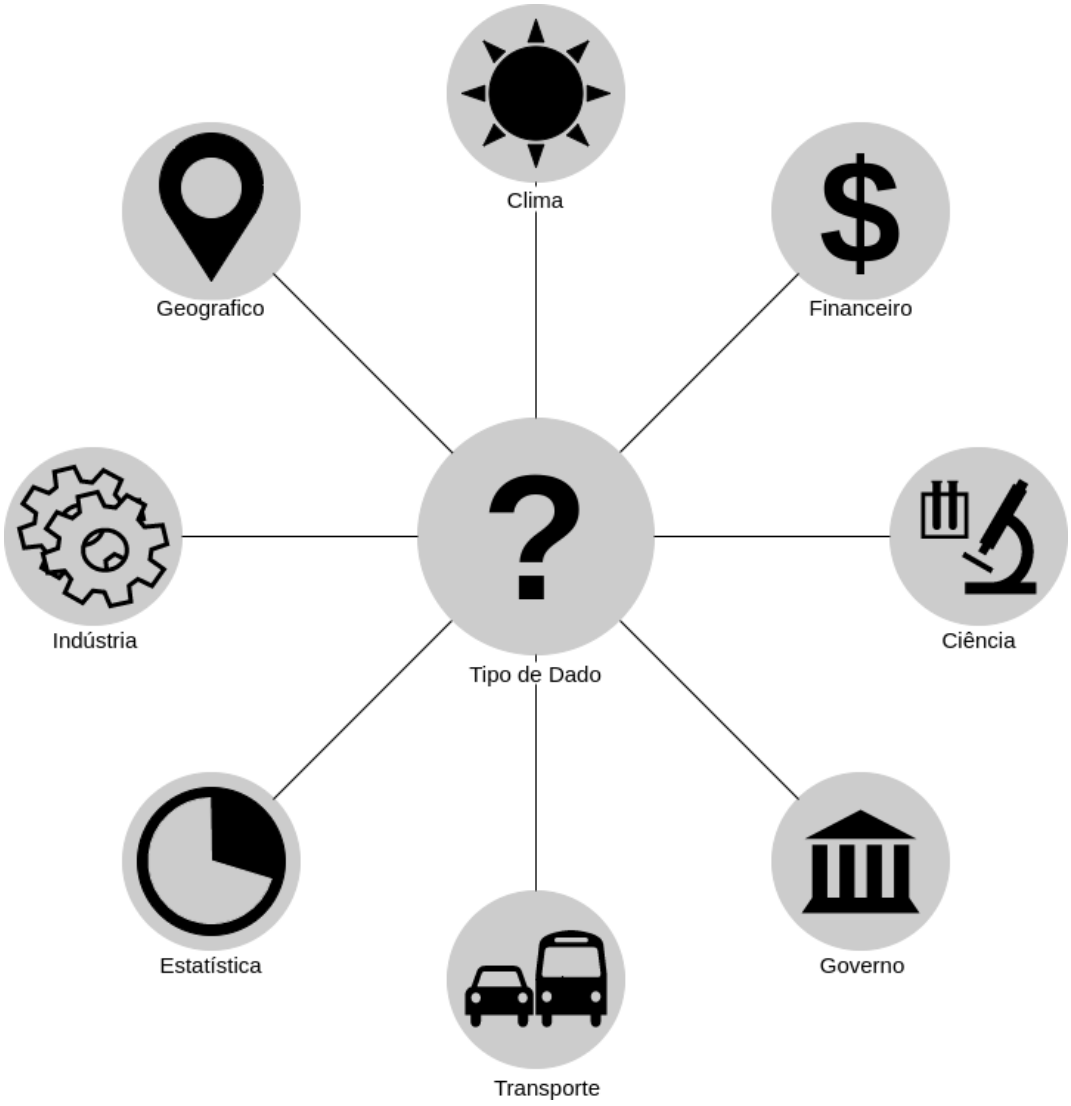
É importante salientar que quando se fala de Dados Abertos, está se falando de todos os tipos de dados, de qualquer assunto e de qualquer fonte (Pessoas, Instituições Públicas, Instituições Privadas, etc), O importante é ter a licença aberta para que possa ser reutilizado. Existem casos de dados abertos em diversas indústrias e setores (transporte, ciência, produtos, educação, sustentabilidade, mapas, legislação, bibliotecas, economia, cultura, desenvolvimento, negócios, design, finanças, entre outros). A Figura 5 apresenta alguns exemplos de tipos de dados que podem ser dados abertos (OKF, 2013).

### 3.2.1 Características e Princípios dos Dados Abertos

Para entender se um dado deveria ser aberto a OKF (2013) por meio da Open Definition OD (2019) apresenta uma série de características necessárias para o dado ser considerado aberto, essas características são:

- Disponibilidade e acesso: os dados devem estar disponíveis como um todo e não mais do que um custo razoável de reprodução, de preferência por download pela internet. Os dados também devem estar disponíveis de uma forma conveniente e modificável;

Figura 5 – Tipos de Dados Abertos




Fonte: Adaptado de OKF (2013)

- Reutilização e redistribuição: os dados devem ser fornecidos em termos que permitam a reutilização e a redistribuição, incluindo a mistura com outros conjuntos de dados. Os dados devem ser legíveis por máquina;
- Participação universal: todos devem poder usar, reutilizar e redistribuir - não deve haver discriminação contra áreas de atuação ou contra pessoas ou grupos. Por exemplo, restrições "não comerciais"que impediriam o uso "comercial"ou restrições de uso para determinados fins (por exemplo, apenas em educação) não são permitidas.

Mais tarde a OD (2019) ampliou esse grupo de características. Abaixo, listamos algumas das novas classificações:

- Quanto à Distribuição:

  - Licença Aberta: o trabalho deve ser de domínio público ou fornecido sob licença aberta;
  - Acesso: o trabalho deve ser fornecido como um todo e deve ser baixado via Internet sem custos;
  - Legibilidade da Máquina: o trabalho deve ser fornecido em um formato prontamente processável por um computador;
  - Formato Aberto - O trabalho deve ser fornecido em um formato aberto. Um formato aberto é aquele que não impõe restrições, monetárias ou de outra forma, após seu uso e pode ser totalmente processado com pelo menos uma ferramenta de software gratuita / de código aberto.

- Quanto ao Licenciamento Aberto:
  - Permissão de Uso e Permissão de Redistribuição: a licença deve permitir o uso gratuito do trabalho licenciado incluindo a venda, seja por conta própria ou como parte de uma coleção feita a partir de trabalhos de diferentes fontes;
  - Permissão de Modificação: permite a criação de derivativos do trabalho licenciado e sua distribuição sob os mesmos termos do trabalho original;
  - Permissão de Separação: qualquer parte da obra pode ser livremente utilizada, distribuída ou modificada. Todas as partes devem ter os mesmos direitos que aquelas concedidas em conjunto com a obra original;
  - Permissão de Compilação: autoriza que o trabalho licenciado seja distribuído junto com outros trabalhos distintos, sem colocar restrições a esses outros;
  - Permissão de Aplicação para qualquer finalidade: permite o uso, redistribuição, modificação e compilação para qualquer finalidade;
  - Condições de Atribuição: tratam da atribuição dos colaboradores, detentores de direitos, patrocinadores e criadores;
  - Condições de Integridade: exige que versões modificadas de um trabalho tenham versão diferente do trabalho original e indiquem quais alterações foram feitas;
  - Condições de *Share-alike*: a licença pode exigir que as distribuições do trabalho permaneçam sob a mesma licença ou semelhante;
  - Sem custo: a licença não deve impor nenhum arranjo de taxas, *royalty* ou outra compensação ou remuneração monetária como parte de suas condições;
  - Não discriminação: a licença não deve discriminar nenhuma pessoa ou grupo.

O trabalho da OD vem no mesmo sentido do proposto pela W3C que, em 2011, apresentou uma série de princípios dos dados abertos que podem ser observados abaixo:

- Acesso livre: qualquer pessoa na rede poderá acessar os dados, sem que haja qualquer forma de discriminação;
- Separar os documentos o máximo: separar os dados ao máximo em estruturas distintas. Em outras palavras, procurar reduzir a agregação da informação. A fim de facilitar o entendimento;
- Responsabilidade: promover dentro das instituições públicas a responsabilidade na publicação de dados com qualidade e rapidez, porém respeitando as legislações vigentes;
- Rápida integração: deve-se ter foco em uma integração rápida e para tanto é necessário ofertar ferramentas e especificações técnicas para facilitar o processo e integração;
- Compartilhamento de boas práticas: informações e experiências relacionadas a formas de melhor executar integrações devem ser compartilhadas;
- Formatos de arquivos: dar prioridade a adoção de formatos de arquivo abertos, assim evitando a dependência tecnológica;
- Serialização dos dados: procura disponibilizar os dados em mais de um formato assim facilitando operações entre os conjuntos de dados.

Conforme apresentado o licenciamento é um ponto muito importante dos Dados Abertos, e de acordo com OD (2018) as licenças que estão em conformidade com a definição de aberto proposto pela *Open Definition* estão listadas no Apêndice A. Onde o "Domínio" é referente tipo de material que a licença se aplica, *"By"* é referente à necessidade de citação e "SA" de *share-alike*.

### 3.2.2 A Carta Aberta à Comunidade de Dados Abertos

Em março de 2017 foi criado um grupo de *Chief Data Officers* – CDOs, com integrantes de grandes cidades e condados nos Estados Unidos da América e chamado de *Civic Analytics Network* - CAN. O CAN, em sua Carta Aberta à Comunidade de Dados Abertos, apresenta que, apesar de as cidades já terem liberado terabytes de dados, ele tem por objetivo melhorar a acessibilidade e facilitar a utilização e que os portais de dados abertos das cidades devem evoluir. Para tal, a *Civic Analytics Network* oferece as oito diretrizes a seguir que, se seguidas, promoverão as capacidades dos portais de dados do governo e ajudarão a cumprir a promessa de um governo transparente (CAN, 2017).

A primeira diretriz Melhorar a acessibilidade e a usabilidade, tem por intuito permitir o *download* de dados simples e rápidos, com design amigável para compreensão. Focar na necessidade do usuário e oferecer dados de forma intuitiva e responsiva. Prover

espaço para conversas sobre o uso dos dados, incluindo guias, painéis e comunicação de mídia social, além de disponibilizar informações de consumo.

Já a segunda diretriz Evitar a visualização centrada em um único conjunto de dados, apresenta que os dados são abertos como um conjunto de tabelas que podem ser combinados pelo usuário e que esse trabalho deve ser facilitado. A diretriz seguinte, a terceira Tratar dados geoespaciais como um tipo de dados de primeira classe, significa ter dados geográficos melhores e mais facilmente utilizáveis é uma prioridade entre os consumidores.

A quarta diretriz Melhorar o gerenciamento e a usabilidade dos metadados, visa permitir esquemas de metadados personalizados, métodos de API para definir e atualizar o esquema e o conteúdo, além de interfaces de usuário que exibem e suportem o usuário final. Como diretriz subsequente tem-se Diminuir o custo e o trabalho necessários para publicar dados. persegue a automatização de processos e viabilizar maior volume de publicações.

Já na diretriz seguinte Melhorar do gerenciamento de grandes conjuntos de dados, os fornecedores devem certificar-se de que seus sistemas podem gerenciar grandes quantidades de dados, em tamanho e velocidade.

Por fim, a oitava diretriz Definir preços transparentes com base na memória, não no número de conjuntos de dados. Se foca nos modelos de precificação baseados no número de conjuntos de dados publicados desestimulam a publicação de dados abertos e solapam o espírito e os objetivos do movimento.

Um ano e três meses depois da publicação da Carta Aberta à Comunidade de Dados Abertos o CAN publicou uma nova carta, a Carta Aberta à Comunidade de Dados Abertos: Um ano depois. Nesta nova carta a CAN reafirma seu compromisso com as Oito Diretrizes para Dados Abertos, oferecendo atualizações sobre os compromissos de dados abertos. E afirmam que essas oito diretrizes não apenas deram suporte às discussões internas sobre dados abertos, mas também fomentaram conversas com fornecedores, grupos cívicos e outras partes externas (CAN, 2018).

Os engajamentos ocorridos em decorrência da primeira carta aberta podem ser considerados positivos, pois levou a interação com diversas empresas e organizações interessadas em Dados Abertos, que por sua vez apresentaram produtos e serviços de valia no ciclo de vida dos Dados Abertos, exemplos desses produtos e serviços podem ser observados abaixo no Quadro 3 (CAN, 2018).

## 3.3 Dados Governamentais Abertos

Dados Governamentais Abertos (DGA) em inglês *Open Government Data* teve sua origem em Sebastopol, Califórnia em dezembro de 2007, quando trinta defensores do

Quadro 3 – Produtos e Serviços com foco em Dados Abertos

| **Organização** | **Proposta / Engajamento** |
|---|---|
| CivicActions - empresa de serviços digitais | Projeto DKAN uma plataforma de dados abertos de código aberto e disponível gratuitamente. |
| OpenDataSoft - provedora de plataformas online sediada em Paris | Publicou sua própria carta pública de dados abertos em resposta a CANs e conectou-se com membros da rede parademonstrar como seu trabalho está alinhado com os valores da carta original. |
| Socrata - provedora de softwares governamentais | Expressou seu apoio aos objetivos da carta, além disso, também disso trabalhou para instituir mudanças para alinhar com várias das principais prioridades da nossa carta. |
| Tyler Kleykamp - CDO do Estado de Connecticut | Expressou seu apoio ao Open Data Letter, observando que uma rede como a CAN com valores similares beneficiaria CDOs estaduais e conectada com a rede durante a Cúpula de 2017 sobre o Governo Inteligente de Dados. |
| Expressou seu apoio ao Open Data Letter, observando que uma rede como a CAN com valores similares beneficiaria CDOs estaduais e conectada com a rede durante a Cúpula de 2017 sobre o Governo Inteligente de Dados. | Disponibilizou o Censo de Dados Abertos, um relatório anual que acompanha quais conjuntos de dados estão abertos e disponíveis nas cidades dos Estados Unidos. O Censo de Dados Abertos serve como uma ferramenta valiosa que complementa a visão e o espírito da carta original da CAN. O Censo do Sunlight permite que moradores e funcionários da cidade compreendam melhor quais dados sua cidade disponibiliza, como os dados de sua cidade se comparam aos dados disponíveis de outras cidades americanas e quais conjuntos de dados sua cidade deve considerar em nome de um governo transparente e responsável. |

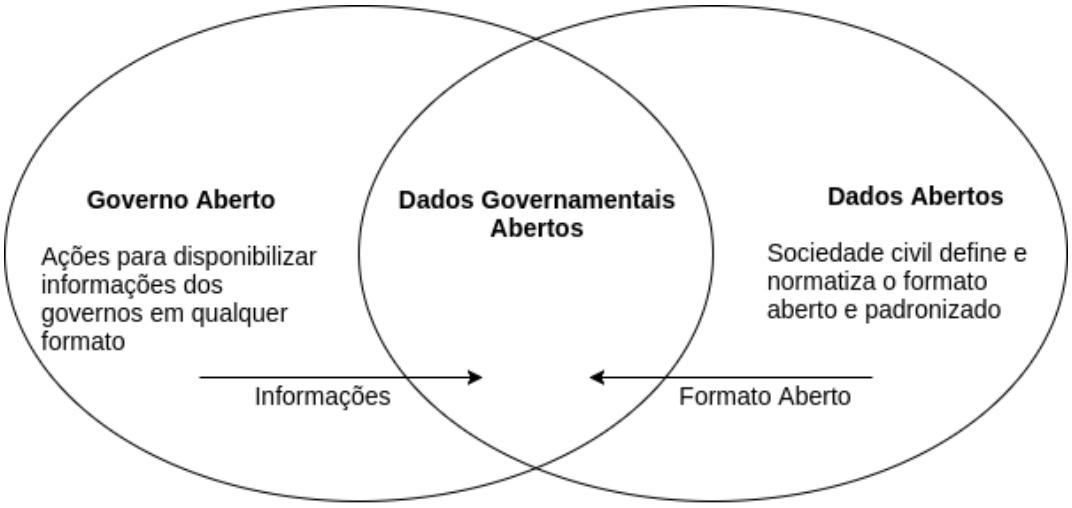
Fonte: Adaptado de (CAN, 2018)

governo aberto se reuniram com o intuito de desenvolver uma compreensão mais robusta do por que dados governamentais abertos são essenciais para a democracia (OGDWG, 2007).

Ainda focando na origem dos DGA, só que mais precisamente quanto as intersecções de conceitos e práticas Albano (2014), essas intersecções se apresentam como a intersecção entre Governo Aberto (GA) e Dados Abertos (DA) como pode ser observado na Figura 6. Essa intersecção demonstra que Governo Aberto, Dados Abertos e Dados Governamentais Abertos são diferentes, porém existe uma relação entre eles. Além disso, Albano apresenta de forma sucinta as definições para GA, DA e DGA:

- Governo Aberto - GA: é a disponibilização de informações em qualquer formato por parte de entidades governamentais e outras ações que promovam transparência;

Figura 6 – Governo Aberto, Dados Abertos e Dados Governamentais Abertos




Fonte: Adaptado de Albano (2014)

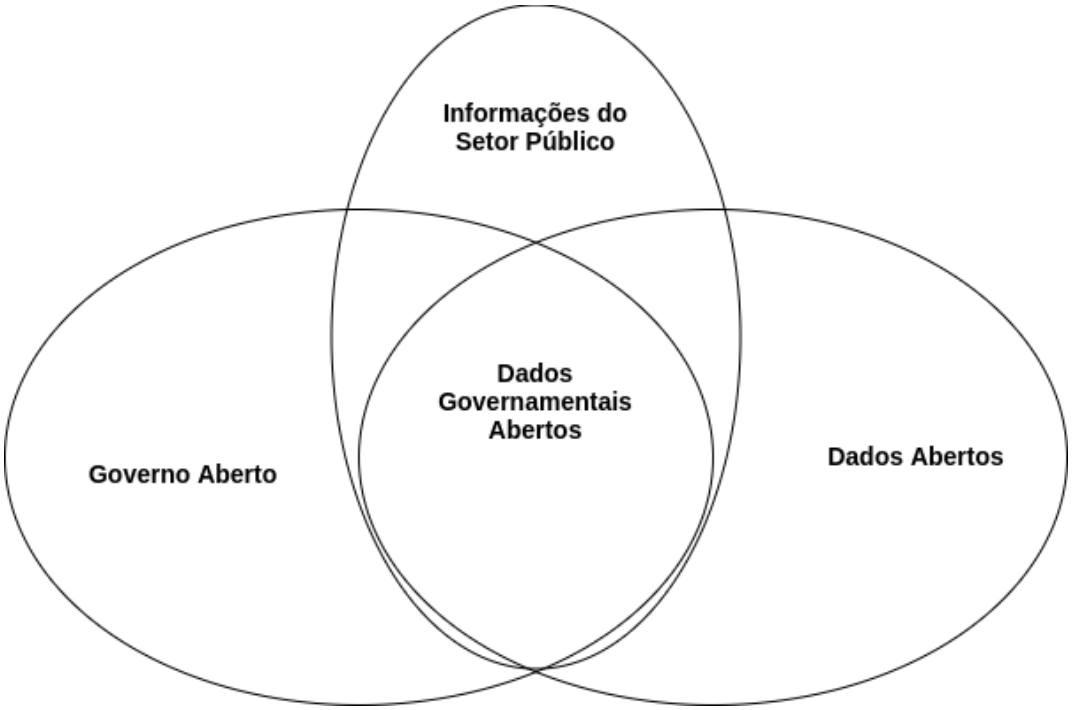
- Dados Abertos - DA: é a disponibilização de dados em formato padronizado por qualquer tipo de entidade (pública, privada, terceiro setor etc.);
- ados Governamentais Abertos - DGA: é a disponibilização de dados em formato padronizado por entidades governamentais.

Também em nesse sentido UN (2013) afirma que em adição a intersecção entre GA e DA, para dar forma aos DGA é necessário que haja a adição de mais um conceito nessa intersecção, esse conceito é Informações do Setor Público - ISP. Os dois principais elementos de dados governamentais abertos podem ser definidos da seguinte forma:

- Dados abertos são definidos como materiais que qualquer um pode usar para qualquer propósito, sem restrições;
- Dados do governo ou ISP são quaisquer dados e informações produzidos ou comissionados por Organismos do Setor Público.

Essas diferenças se apresentam por nem todo ISP é DGA. Na verdade, DGA é a interseção de PSI e dados abertos. Os dados são "abertos", não importa a fonte, somente se puderem ser acessados, reutilizados e redistribuídos por qualquer pessoa, para qualquer finalidade, incluindo reutilização comercial, gratuitamente e sem quaisquer restrições. A maioria desses dados, exceto dados pessoais e dados que podem ser personalizados ou dados classificados por motivos de segurança nacional, pode se tornar dados abertos. A Figura 7 demonstra a intersecção de conceitos descrita acima.

Figura 7 – Intersecções Dados Governamentais Abertos




Fonte: Adaptado de UN (2013)

### 3.3.1 Princípios e Leis dos Dados Governamentais Abertos

O encontro em Sebastopol, Califórnia onde o termo Dados Governamentais Abertos (DGA) foi cunhado, também propiciou a criação dos 8 princípios dos DGA. Esses 8 princípios são chamados de princípios originais, em adição a estes princípios também existem outros 7 que são chamados complementares. o Quadro 4 apresenta todos os 15 princípios, originais e complementares (OGDWG, 2007).

Com outra abordagem, mas também tentando estabelecer as bases para os DGA, Eaves (2009) elaborou três leis que caracterizam e norteiam o conceito de DGA. Essas leis dizem respeito a: disponibilidade e acesso, reúso e redistribuição e participação universal, conforme pode ser observado abaixo no Quadro 5 o detalhamento de cada uma dessas leis.

### 3.3.2 Impactos, benefícios, mitos e barreiras dos Dados Governamentais Abertos

A fim de analisar os impactos dos DGA na sociedade, de acordo com Davies e Perini (2016) é possível dividi-los em três grandes domínios: políticos, econômicos e sociais.

Na esfera política os dados abertos trarão mais transparência no governo, que por sua vez provoca uma maior responsabilização dos atores-chave para tomar decisões e aplicar regras de interesse público. Na economia, os dados abertos permitirão que os

Quadro 4 – Princípios dos Dados Governamentais Abertos

| **Tipo** | **Princípio** | **Descrição** |
|---|---|---|
| Original | Completos | Todos os dados públicos são disponibilizados. Dados públicos são dados que não está sujeita a limitações de privacidade, segurança ou privilégios válidos. |
| Original | Primários | Os dados são recolhidos na fonte, com o mais alto nível possível de granularidade, não em forma agregada ou modificados. |
| Original | Atuais | Dados são disponibilizados tão rapidamente quanto necessário para preservar o valor dos dados. |
| Original | Acessíveis | Os dados estão disponíveis para a mais ampla gama de usuários e propósitos. |
| Original | Processável por Máquina | Os dados estão razoavelmente estruturados para permitir o processamento automatizado. |
| Original | Acesso não Discriminatório | Os dados estão disponíveis para qualquer pessoa, sem necessidade de registro. |
| Original | Formato não Proprietários | Os dados estão disponíveis em um formato sobre o qual nenhuma entidade tem controle exclusivo. |
| Original | Livre de Licença | Os dados não estão sujeitas a qualquer direito autoral, patente, marca registrada ou regulamento segredo comercial. Privacidade razoável, restrições de segurança e de privilégio podem ser permitidas. |
| Comple mentar | Online e Gratuito | A informação não é forma significativa pública caso não esteja disponível na Internet, sem custo, ou pelo menos não mais do que o custo marginal de reprodução. Deve também ser fáceis de encontrar. |
| Comple mentar | Permanente | Os dados devem ser disponibilizados em um lugar com Internet estável, por tempo indeterminado e em um formato de dados estável por tanto tempo quanto possível. |
| Comple mentar | Confiável | As assinaturas digitais ajudam o público validar a fonte dos dados, de modo que eles podem confiar que os dados não foram modificado. |
| Comple mentar | Presunção de Aberto | Presunção de abertura baseia-se em leis como a Lei de Liberdade de Informação, incluindo procedimentos de gerenciamento de registros e ferramentas, tais como catálogos de dados. |
| Comple mentar | Documentário | Documentação sobre o formato e significado dos dados vai um longo caminho para tornar os dados úteis. |
| Comple mentar | Seguro para Abrir | Evitar conteúdos executáveis dentro de documentos, pois representa um risco de segurança para os usuários de dados porque o conteúdo executável pode ser malware. |
| Comple mentar | Projetado com a Participação Pública | O público está na melhor posição para opinar sobre quais as tecnologias de informação que serão mais adequadas para as aplicações voltadas a ele próprio. A contribuição do público é, portanto, crucial para a divulgação de informações com real valor. |

Fonte: Adaptado de (OGDWG, 2007)

Quadro 5 – Leis dos Dados Governamentais Abertos

| | Disponibilidade e acesso | Reuso e redistribuiçã | Participação universal |
|---|---|---|---|
| Original | "Se não pode ser indexado, não existe" | "Se não estiver disponível em formato aberto e legível por máquina, não será possível envolver" | "Se uma estrutura legal não permitir que ela seja reutilizada, ela não capacitará" |
| Explicação | Estar totalmente disponível na Internet, ter custo acessível e formato que permite sua reutilização | Além de poder ser reutilizado, seu formato deve permitir o cruzamento com outros dados | Estar disponível para todos sem restrição de nenhuma espécie. Ex.: uso somente para fins educacionais |

Fonte: Adaptado de (OGDWG, 2007)

atores não estatais possam agir para melhorar os serviços públicos ou construir produtos e serviços inovadores com valor social e econômico.

Já no aspecto social, os dados abertos eliminarão os desequilíbrios de poder que resultaram da assimetria da informação. Eles trarão novos interessados em debates políticos, dando a grupos marginalizados mais voz na criação e aplicação de regras e políticas.

A abertura de dados tem como objetivo estimular a transparência pública e o controle dos atos do governo, tornando possível a sociedade desenvolver produtos e serviços utilizando estas informações, estimulando novas atividades econômicas. Dentre as vantagens e benefícios dos Dados Governamentais Abertos destaca-se a explanação Janssen, Charalabidis e Zuiderwijk (2012), que de forma similar a Davies e Perini (2016) classificam também em três grupos, mas dessa vez como: políticos e sociais, econômicos e, por fim, operacionais e técnicos.

Entre os benefícios políticos e sociais está o aumento da transparência, responsabilidade democrática, da participação e auto responsabilização dos cidadãos. A igualdade no acesso aos dados estimula ainda a construção do conhecimento.

Novos serviços governamentais para os cidadãos e inovação social também são resultados da abertura. Além disso, a melhoria no atendimento, na satisfação dos cidadãos e nos processos de formulação de políticas públicas aumenta a confiança no governo.

Os dados abertos estimulam o crescimento econômico, a competitividade e a inovação. Eles facilitam o desenvolvimento e aprimoramento processos, produtos e serviços pelo uso do conhecimento coletivo e a disponibilização de informações para investidores e empresas.

No aspecto operacional e técnico, a abertura dos dados permite a reutilização destes, evitando a duplicidade de trabalho e custos associados. Também melhora as políticas públicas, por meio da otimização de processos administrativos. Por fim, permite a criação

de novas bases de dados por meio de combinação, se utilizando da capacidade de terceiros para resolver problemas.

É interessante fazer uma reflexão sobre a lacuna entre as promessas e barreiras dos DGA. Estas lacunas, de acordo com Janssen, Charalabidis e Zuiderwijk (2012) são apresentadas como mitos e podem desempenhar um papel importante na formulação de políticas públicas, por inspirarem a ação coletiva popular. Também de acordo com Janssen, Charalabidis e Zuiderwijk (2012), são cinco os mitos mais proeminentes sobre DGA:

Sendo o primeiro Mito A divulgação dos dados trará benefícios automaticamente, a abertura de dados deve ser estimulada pelo princípio de que, ao menos que haja argumentos para a não divulgação, os dados devem ser publicados. Porém, muitas instituições públicas estão divulgando dados sem uma política sólida de abertura. Esta ausência acarreta alguns problemas, como a falta de padronização de formatos, de um portal centralizado para publicação, pulverização das informações. Além disso, por conta dos dados terem um valor intrínseco baixo, apenas sua combinação é capaz de produzir valor (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

Ainda de acordo com Janssen, Charalabidis e Zuiderwijk (2012) o Mito seguinte se trata de Todas as informações devem ser divulgadas sem restrição. Essa crença, de acordo com os autores, ignora uma série de questões, como o fato de que dados privados não podem ser divulgados, de acordo com a legislação. Dados que não trazem benefícios, não precisam ser disponibilizados, assim economizando verba pública, por exemplo. Neste quesito, pode-se concluir que a divulgação de dados públicos deve seguir critérios rigorosos, que levam em conta questões como privacidade, utilidade dos dados, qualidade, complexidade, legislação, potencial financeiro de sua utilização.

Já o Mito 3, É uma questão simples de publicar dados públicos, apresenta que apesar da publicação do dado bruto ser sempre a forma principal de publicação, por muitas vezes não pode ser utilizado sem um tratamento prévio. É preciso ainda padronizar dos metadados para permitir a indexação dos dados, criando a ligação (link) com outros conjuntos e facilitando sua busca e interpretação (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

O penúltimo mito, Mito 4, Todos os constituintes podem fazer uso dos DGA. Para essa afirmação ser verdade, os dados precisam estar em formato simples ou em aplicativos de software de fácil utilização. Deve-se partir do princípio de que os usuários não possuem o conhecimento técnico necessário a fim de obter uma profunda compreensão dos dados (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

Por fim, O Mito 5, Dados abertos resultarão em um governo aberto. Conforme Janssen, Charalabidis e Zuiderwijk (2012) afirmam, isso se trata de um mito, pois, somente abrir os dados, sem proporcionar métodos para interpretá-los e processá-los é inútil. Também, é preciso contar com mecanismos de busca que favoreçam a pesquisa das bases

de dados. Outro fator é a necessidade da implantação de sistemas para coleta de feedback dos usuários, a fim de retroalimentar o processo de qualidade e manutenção do serviço.

Apesar dos DGA terem um impacto extremamente positivo no governo, na sociedade e na economia, eles enfrentam várias adversidades em suas tentativas de implantação. Quando se trata da classificação das barreiras e os fatores inibidores à implantação de DGA, Janssen, Charalabidis e Zuiderwijk (2012) dividem esses obstáculos e fatores inibidores em seis grandes grupos: institucionais, complexidade da tarefa, uso e participação, legislação, qualidade da informação e técnicas.

As barreiras institucionais são formadas, principalmente, pelo dilema entre transparência e privacidade, pela cultura de aversão ao risco e a falta de empreendedorismo. Isso bloqueia a inovação e negligencia as oportunidades.

Além disso, a falta de uma uniformização na política de divulgação de dados, a qualidade das entradas providas por usuários e falta de recursos para a publicação também são barreiras institucionais.

Quando o assunto é a complexidade da tarefa, falta habilidade para detectar os dados apropriados e acessar os originais. Os usuários também encontram dificuldades para utilizar os dados coletados, especialmente aqueles complexos.

A qualidade da informação é uma barreira no momento em que os usuários encontram dados obsoletos, incompletos ou não sabem como analisá-los em conjuntos.

As ferramentas de suporte também não são favoráveis. Algumas escondem a complexidade e os outros usos potenciais da informação. Tudo isso leva a resultados contraditórios e conclusões equivocadas. As questões técnicas acabam sendo mais uma das barreiras.

Quanto ao uso e participação, a falta de tempo para aprofundar-se nos detalhes, a cobrança ou necessidade de registro pelo acesso aos dados e os custos inesperados escaláveis desestimulam os DGAs.

A legislação não traz segurança na violação de privacidade e a inexistência de licença para utilização dos dados, limita seu uso. A necessidade de permissão prévia, por escrito, para ter acesso aos dados é outra barreira encontrada.

Em resumo, os Dados Governamentais Abertos têm potencial de trazer mais transparência para as relações, aumentar a participação dos cidadãos e estimular o crescimento econômico, entre tantos outros benefícios.

Para tanto, é preciso reconhecer as barreiras, mitos e obstáculos no caminho dos DDAs, conforme levantado pelo trabalho de (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

# 4 Ciência de Dados

Explorar os padrões e regras na natureza dos dados é necessário, mas difícil. Uma nova disciplina chamada Ciência dos Dados está chegando. Ele fornece um tipo de método de pesquisa inovador para ciências naturais e sociais e vai além da ciência da computação na pesquisa de dados (ZHU; XIONG, 2015).

Neste sentido, este capítulo tem por objetivo apresentar as bases mais voltadas para o campo tecnológico que dão sustentação ao trabalho, além de possibilitar um nivelamento de conhecimento no tocante do que é a Ciência de Dados e os itens relacionados a ela como: ciclo de vida dos dados, o ciclo de vida de um projeto de análise de dados e também as tecnologias envolvidas como linguagens de programação, formato de dados e métodos de comunicação.

Dessa forma, esse capítulo é focado em trazer os pontos tecnológicos para dar sustentação as bases socioculturais apresentadas nos capítulos anteriores, de forma que, as similaridades e encontros entre a parte técnica voltada a tecnologia e a base teórica possam ser sintetizadas no capítulo posterior.

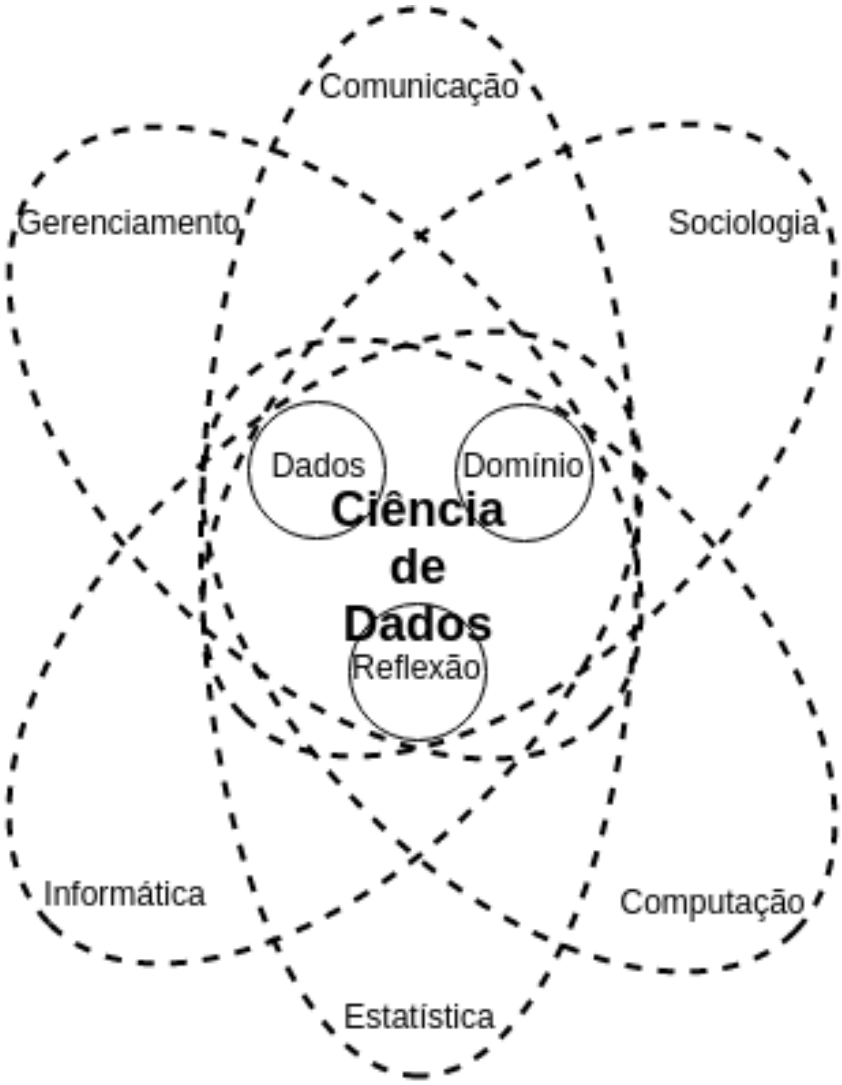
## 4.1 Ciência de Dados

O termo Ciência de Dados, em inglês, *Data Science*, foi cunhado há, aproximadamente, 15 anos. Os dados são um ativo crítico e a Ciência de Dados é o núcleo interdisciplinar que tem atraído cada vez mais atenção e debate nas áreas de estatística, análise, computação, Ciências Sociais e outros domínios e disciplinas científicas (CAO, 2016a).

Conforme afirmado por Cao (2016a) existem diferentes definições e interpretações, da Ciência de Dados. Como campo científico, que desenvolve metodologias, teorias, tecnologias e aplicativos relevantes para dados, desde a captura, criação, representação, armazenamento, pesquisa, compartilhamento, privacidade, segurança, modelagem, análise, aprendizagem, apresentação e visualização, até a integração de recursos complexos, heterogêneos e interdependentes para a tomada de decisões em tempo real, colaboração, criação de valor e suporte à decisão.

Sobre a definição Ciência dos Dados, Cao (2017b) apresenta uma proposta que, de acordo com a Figura 8, é centrada na transdisciplinaridade: “A ciência de dados é um novo campo transdisciplinar que constrói e sintetiza várias disciplinas e corpos de conhecimento relevantes, incluindo estatística, informática, computação, comunicação, gerenciamento e sociologia”.

Figura 8 – Ciência de Dados




Fonte: Adaptado de Cao (2017b)

Wang também trabalhou o conceito de Ciência dos Dados por meio da exposição de diversos outros conceitos elaborados por diferentes autores e de acordo com Wang (2018) “parece haver algum consenso de que a ciência de dados é um campo interdisciplinar que diz respeito à identificação e extração de padrões, convertendo dados em informação e conhecimento por meio de análise de dados e mineração".

Ainda se baseando em Wang (2018), que apresenta as contribuições de suma importância da Ciência da Informação para a Ciência de Dados como: Conceito de Dados, Controle de qualidade de dados, Bibliotecário de dados e Teoria dos documentos.

### 4.1.1 Contribuições da Ciência da Informação para a Ciência de Dados

O Conceito de Dados está no centro de toda a disciplina da Ciência de Dados, os cientistas subscreveram a crença de que os dados são objetivos e neutros. Já os estudiosos da informação argumentam que os mesmos têm um viés. Portanto, os dados são dependentes da teoria e podem estar cheios de vieses. Os "dados" da Ciência de Dados não são fatos objetivos e desinteressados, mas a coleta de informações de acordo com metas e pressupostos específicos (CARTER; SHOLLER, 2016)

O Controle de Qualidade é outro ponto onde existe uma contribuição da Ciência da Informação, como máxima apresentada sobre a qualidade dos dados: nem todos os dados são criados iguais. Além disso, as fontes são variadas, algumas sendo valiosas que outras e a Problemas de precisão e credibilidade dos dados existem quando os cientistas processam grandes volumes dos mesmos. É necessário executar a seleção e o controle da qualidade dos dados para desenvolver uma Ciência de Dados madura. A Ciência da Informação pode desempenhar um papel importante nesse aspecto. As teorias da CI qualificam-se como excelentes candidatas para fundamentar o campo da qualidade na Ciência de Dados (WANG, 2018).

Para contrapor a explosão de informações que vivemos na era do big data, nesse contexto surge o Bibliotecário de dados que trabalha no ecossistema de dados, com os princípios, teorias, conhecimentos e habilidades da Ciência da Informação que são aplicáveis e valiosas no gerenciamento e no suporte a Ciência de Dados. O que os bibliotecários de dados fazem abrange todos os ciclos de vida do gerenciamento de dados (WANG, 2018).

Seu trabalho começa nos planos de gerenciamento, coleta, preservação, curadoria, controle, acesso, metadados, documentação dos conjuntos de dados, compartilhamento, visualização, suporte à análise, avaliação de qualidade, referência, citação e treinamento em alfabetização de dados (WANG, 2018).

Os bibliotecários de dados devem ter competência técnica e tecnologia de gerenciamento de dados mestres; por outro lado, eles devem ter um conhecimento abrangente das implicações socioculturais e éticas da tecnologia de dados. Sua compreensão dos dados é sociotécnica (WANG, 2018).

Partindo do pressuposto que dados e documentos são informação-como-coisa, assim, estes tornam-se entidades intercambiáveis. Uma categoria crítica do conceito de informação é informação-como-coisa, significando "objetos físicos como dados e documentos que são referidos como informação porque são considerados informativos". A palavra "documento" pode ser usada para denotar informação-como-coisa (BUCKLAND, 1991).

Ainda de acordo com (BUCKLAND, 1991), devido às suas características físicas e tangíveis, sejam sistemas especialistas ou sistemas de recuperação de informações, podem manipular diretamente. Com base nessa equivalência conceitual, a pesquisa em

ciência de dados pode adotar teorias substanciais de documentos, incluindo teorias de coleta, organização, recuperação, disseminação, gerenciamento e uso de documentos, para enriquecer suas próprias teorias (WANG, 2018).

### 4.1.2 A Era da Ciência dos Dados

Conforme apresentado por Cao (2016a) e Cao (2016b)estamos na era do analytics, Ciência de Dados e do big data. Essa era é caracterizada por transformações e mudanças de paradigmas, onde podem ser destacados três indicadores (CAO, 2016b):

- Mudança de paradigma disciplinar para um paradigmas centrado em dados;
- Transformação tecnológica, ou avanços na tecnologia de dados de uma geração para outra;
- Inovação por meio do desenvolvimento de produtos de dados com aplicação técnica e prática.

Essa mudança de paradigma disciplinar pode ser identificada quando vemos as seguintes transições: da analítica descritiva para a análise profunda e da análise de dados para a Ciência de Dados. Também pode ser identificada nos avanços tecnológicos como as transições da *World Wide Web* para a *Wisdom Web* e da Internet para a Internet de Todas as Coisas (CAO, 2016b).

De acordo com Cao (2016b) às transformações na inovação, por sua vez, podem ser identificados, por exemplo, por meio da transição de economia digital para economia de dados, governo fechado para governo aberto e comércio eletrônico para negócio online.

Quando analisada a fundação dessa era da ciência de dados, Ayankoya, Calitz e Greyling (2014) argumenta que a *Datafication* exerce um papel fundamental nessa era de Ciência de Dados. A *Datafication* trata da transformação de todos os aspectos da vida em dados, indo muito além de converter informações analógicas existentes, como livros e fotografias, em formatos digitais.

Outro ponto que Cao (2017a) destaca nesta nova era é adoção e aceitação de modelos abertos. Isso se dá por conta destes poderem ser distribuídos livremente e por serem colaborativos em seu desenvolvimento. Neste sentido, andam de forma conjunta com dados abertos, compartilhamento de dados e acesso aberto.

### 4.1.3 Datafication e Quantificação de dados

O fenômeno da *Datafication* também é responsável por possibilitar a coleta de dados de formas não convencionais de áreas que anteriormente não era percebido valor na informação para ser coletada (AYANKOYA; CALITZ; GREYLING, 2014).

UUm exemplo prático da *Datafication* é a Quantificação dos dados, que pode ser percebida quando aplicativos móveis que agora são capazes de organizar os movimentos humanos, fornecendo dados e informações subsequentes sobre perda de peso e problemas de saúde (AYANKOYA; CALITZ; GREYLING, 2014).

Outras formas de *Datafication* e Quantificação dos dados são apresentados por (CAO, 2017a) como:

- Quantificação de tempo: quantificação a qualquer momento, desde o trabalho até o estudo, a vida cotidiana, o relaxamento, o entretenimento e a socialização;
- Quantificação de lugar: quantificação em qualquer lugar, de sistemas biológicos a sistemas e ambientes físicos, cibernéticos, ambientais, culturais, econômicos, entre outros;
- Quantificação de organismos: qualquer quantificação, de "eus" a outros, mundo, de indivíduos a grupos, organizações e sociedades;
- Quantificação de formas: quantificação de qualquer forma, desde a observação até os direcionadores, do objetivo ao subjetivo, do físico ao filosófico, do explícito ao implícito e das formas e aspectos qualitativos aos quantitativos;
- Quantificação de fontes:quantificação da fonte, como fontes e ferramentas que incluem sistemas de informação, digitalização, sensores, sistemas de vigilância e rastreamento, IoT, dispositivos móveis e aplicativos, serviços sociais e plataformas de rede e *wearable*;
- Quantificação de velocidade: quantificação à qualquer velocidade, de estática a dinâmica, de finita a infinita, e de geração incremental a exponencial de objetos de dados, conjuntos, armazéns, lagos e nuvens.

Por fim, Cao (2016b) apresenta que o resultado da quantificação de dados são os mesmos em todos os lugares, incluindo a Internet; IoT; redes de sensores; repositórios socioculturais, econômicos e geográficos, sensores personalizados, incluindo fontes móveis, sociais, vivas, divertidas e emocionais.

### 4.1.4 Tomada de decisões baseadas em dados

A Ciência de Dados envolve princípios, processos e técnicas para entender fenômenos através da análise de dados. Estes princípios, processos e técnicas tornam-se base para a tomada de Decisão Baseada em Dados - DBD. Esta, refere-se à prática de basear decisões na análise de dados que pode ser realizada manualmente ou de forma automatizada, em vez de apenas na intuição (PROVOST; FAWCETT, 2013).

A Ciência de Dados apoia a tomada de decisão orientada por dados, mas também se sobrepõe a ela. Isso destaca o fato de que, cada vez mais, as decisões de negócios estão sendo tomadas automaticamente por sistemas de computação (PROVOST; FAWCETT, 2013).

Por exemplo, um profissional de marketing pode selecionar anúncios baseados apenas em sua longa experiência no campo e seus olhos para o que funcionará. Outra opção seria basear sua seleção na análise de dados sobre como os consumidores reagem a diferentes anúncios. Ele também poderia usar uma combinação dessas abordagens (PROVOST; FAWCETT, 2013).

Ainda de acordo com Provost e Fawcett (2013), pesquisadores do MIT e Wharton School de Penn realizaram um estudo para avaliar o quando a utilização do DBD afeta o desempenho das empresas. Neste estudo eles desenvolveram uma medida de DBD que classifica as empresas quanto a quão fortemente elas usam dados para tomar decisões em toda a empresa. Eles mostram estatisticamente que quanto mais dados uma empresa é, mais produtiva ela é. Desta forma demonstrando de forma conclusiva que existem benefícios na tomada de decisão baseada em dados.

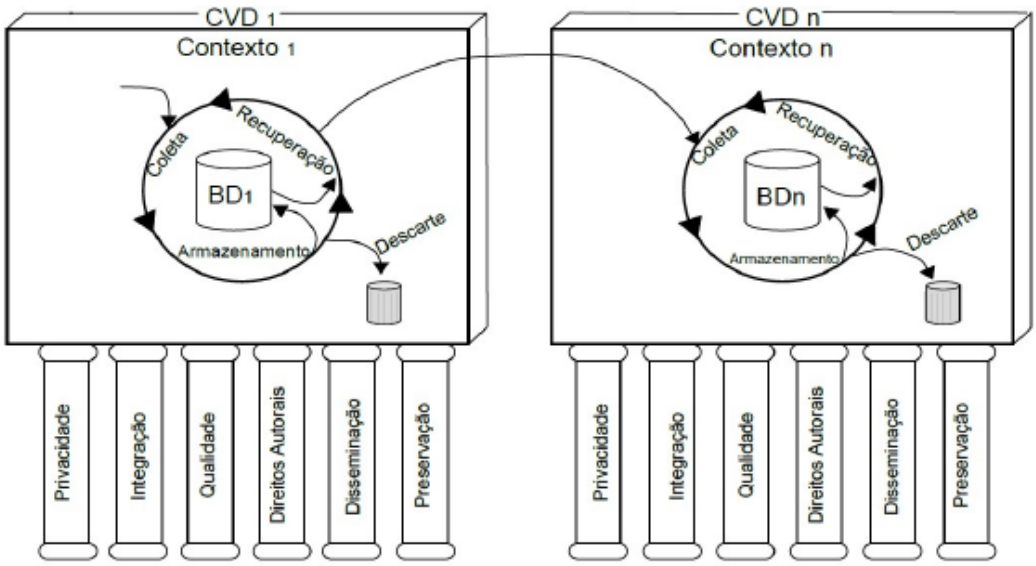
Como caso de uso pode-se apresentar, por exemplo, uma grande empresa de telecomunicações que pode ter centenas de milhões de clientes, cada qual candidato à deserção. Dezenas de milhões de clientes têm contratos que expiram a cada mês, então, cada um deles tem uma probabilidade maior de deserção no futuro próximo. Se pudermos melhorar nossa capacidade de estimar, para um determinado cliente, o quão lucrativo seria focarmos nela, poderemos potencialmente colher grandes benefícios aplicando essa capacidade aos milhões de clientes da população (PROVOST; FAWCETT, 2013).

Por fim, Provost e Fawcett (2013) argumenta que essa lógica apresentada no caso de uso se enquadra em muitas das áreas em que vimos a aplicação mais intensa da ciência de dados e mineração de dados: marketing direto, publicidade on-line, pontuação de crédito, negociação financeira, gerenciamento de help desk, detecção de fraude, classificação de pesquisa, recomendação de produto, e assim por diante.

## 4.2 Ciclo de Vida dos Dados - CVD pela perspectiva da Ciência da Informação

Na sociedade moderna, o acesso e uso de dados tem despontado com um fator chave para o sucesso e tem se estendido às mais variadas áreas. Esse acesso tem o potencial de transformar todas as áreas de atuação humana, assim, os processos de coleta, armazenamento e recuperação dos dados têm um papel central. Para entender mais profundamente esses processos/fases é necessário aprofundar a discussão sobre o ciclo de

Figura 9 – Ciclo de Vida dos Dados




Fonte: Adaptado de Sant'Ana (2016)

vida dos dados (SANT'ANA, 2016).

Levando em consideração a enorme quantidade de dados gerada atualmente, os processos de coleta, armazenamento e recuperação que vêm sendo utilizados são relacionados ao Big Data (SANT'ANA, 2016).

De acordo com Davenport (2014) "Big data se refere a dados que são grandes demais para um único servidor, muito diversos para se adequar a uma base de dados estruturada em linhas e colunas". De acordo com Akerlof, 1970 a posse e capacidade/possibilidade de análise de grandes volumes de dados pode acarretar efeitos colaterais como a assimetria informacional. Essa assimetria tem o potencial de anular os benefícios trazidos pelo Big Data como a possibilidade de acesso e transformação massiva de grande volume de dados (SANT'ANA, 2016).

Com a disponibilidade de dados crescendo dia a dia devido aos avanços tecnológicos, é importante delimitar fases envolvidas no acesso e uso dos dados. Tendo como ponto central o próprio dado, surge o CDV como forma de evidenciar os diferentes momentos e fatores envolvidos neste processo (SANT'ANA, 2016).

Todo conjunto de dados derivado do dado original tem seu próprio ciclo de vida conforme apresentado na Figura 9. Nela, pode-se observar que a criação de um conjunto de dados a partir de outro pré-existente, cria um novo ciclo de vida.

O Ciclo de Vida dos Dados proposto por Sant'Ana é composto por 4 fases e 6 pilares conforme apresentado abaixo (SANT'ANA, 2016):

**Fase 1 - Coleta**: fase de obtenção de dados que podem ser utilizados para atender uma necessidade específica ou uma demanda prevista de informações sobre um determinado contexto.

**Fase 2 - Armazenamento**: denominado de persistência dos dados leva a uma série de preocupações e aspectos que devem ser detalhadamente planejados.

**Fase 3 - Recuperação**: fase em que preocupações e esforços são voltados para que estes dados possam ser encontrados, acessados e interpretados. Trata-se de uma etapa em que o objetivo passa a ser a viabilização da recuperação destes dados.

**Fase 4 - Descarte**: responde pela limpeza ou, simplesmente, desativação de parte dos dados. Nesta etapa, identificada como fase de descarte, tem-se então a eliminação de parte dos dados que pode ocorrer em bloco, horizontalmente ou verticalmente.

**Pilares**: Privacidade, Integração, Qualidade, Direitos Autorais, Disseminação e Preservação.

De forma a apresentar o relacionamento das fases com os pilares, segue abaixo uma descrição das fases mais robusta e como as mesmas se relacionam com os pilares (SANT'ANA, 2016).

### 4.2.1 Coleta

A coleta de dados pode ser caracterizada de duas formas, em lote ou em tempo real. A diferença entre essas duas modalidades se dá pelo intervalo de coleta. Como os nomes sugerem, a coleta em lote é realizada periodicamente dentro de intervalos de tempo, portanto o dado é obtido de forma assíncrona à sua produção. Já a em tempo real, é feita uma coleta de forma síncrona com a produção ou em intervalos de tempo pequenos (SANT'ANA, 2016).

**Privacidade**: necessário identificar nas fontes de dados os aspectos/variáveis que possam caracterizar dados privativos de pessoas ou instituições.

**Integração**: deve-se identificar e validar os atributos que serão usados como chave para agregação com outras fontes de dados.

**Qualidade**: a confiabilidade é uma condição fundamental para que o dado seja útil, e para garantir a qualidade e a procedência, mecanismos de coleta e garantias de integridade física e lógica são elemento a serem considerados.

**Direitos Autorais**: respeitar os direitos autorais vinculados aos dados, em outras palavras, respeitar o arcabouço jurídico que sustenta a legalidade deste acesso.

**Disseminação**: identificar elementos contextuais dos dados que possam favorecer sua localização e interpretação na fase de recuperação.

**Preservação**: adição de metadados que propiciem a identificação, para que possam ser preservados mas também utilizados.

### 4.2.2 Armazenamento

Esta é uma fase com grande enfoque tecnológico, porque são defidos aspectos que garantem a reutilização dos dados, por meio de especificações físicas e lógicas sobre como os dados serão persistidos. Além disso, nesta fase também se faz necessário algumas definições, que são (SANT'ANA, 2016):

1. Dos dados coletados na fase anterior, qual serão os tipos utilizados (inteiro, caractere, data, entre outros), além de especificações semânticas como, por exemplo, unidade de medida. E também definir qual a estrutura de variáveis receberá esses dados coletados;

2. Estruturação das variáveis definidas no item anterior com seus valores, possibilitando sua interpretação como a estrutura básica <e, a, v>;

3. Principalmente quando se trata a dados sensíveis, mas também se referindo às outras variáveis, quem poderá acessar qual variável. Preferencialmente o permissionamento deve se dar em diversas granulometrias, como banco de dados, tabela e coluna. Além disso se deve definir qual o nível de anonimização necessário para cada variável para que ela possa ser disponibilizada em cada grupo de acesso;

4. Como se dará o acesso, este pode ser de forma direta ou como em um Sistema Gerenciador de Banco de Dados (SGBD) fazendo a interface entre o usuário e o dado. No caso do acesso direto, deve-se definir a semântica utilizada no dado como por exemplo, em CSV qual será o caractere usado para separação de campos;

5. Baseado na definição do acesso direto, o formato do dado deve ser definido nesta etapa, como, por exemplo, definição pela utilização de um formato JSON ou CSV;

6. Onde o dado será armazenado.

**Privacidade**: está fortemente relacionada com o item 3) onde ocorre a identificação e anonimização.

**Integração**: é baseada nos itens 3) e 4), devido as definições de quem pode acessar o recurso, a forma que se dará o acesso e qual o formato do arquivo.

**Qualidade**: é fundamental garantir que os dados mantenham sua integridade física e lógica, no caso de dados sensíveis a qualidade dos dados ganha dimensão extra já que decisões tomadas a partir destes podem gerar consequências graves.

**Direitos Autorais**: vinculados a fonte da qual os dados foram obtidos e de forma independente os metadados devem ter informações sobre a licença do recurso para que seja facilitado o entendimento sobre os limites de utilização do mesmo.

**Disseminação**: como intuito de propiciar a interpretação automatizada dos dados deve-se incorporar semântica aos dados além de prover meio com que estes dados sejam localizados e acessados.

**Preservação**: deve prever que esses dados sejam acessíveis no futuro, contexto do Big Data, deve-se levar em conta não somente os aspectos comuns ao processo de preservação, mas, também, fatores, como por exemplo, a vasta gama de formatos e de variedades de fontes de dados, além da diversidade de dispositivos de coleta.

### 4.2.3 Recuperação

As ações e estratégias dessa etapa tem por perspectiva o responsável pela manutenção, mas tem como foco os dados, passando a tornar os mesmos disponíveis para acesso e uso. Formas que ampliem e propiciem níveis de utilização dos dados, seja pelo aumento das possibilidades de acesso por meio de cópia ou obtenção por meio de outros recursos, como uma Interface de Programação de Aplicações em inglês *Application Programming Interface* - APIs (SANT'ANA, 2016).

**Privacidade**: os níveis de acesso e anonimização devem ser atribuídos para que cada nível de acesso tenha direito de recuperar dados relativos ao seu nível de permissão.

**Integração**: é necessário que se possibilite a análise de entidades distintas porém integradas de forma que representem o todo.

**Qualidade**: a Arquitetura da Informação deve ser levada em conta para que se ampliem a usabilidade e a acessibilidade, evitando possíveis erros derivados da própria interface.

**Direitos Autorais**: deve-se deixar explícito quem pode acessar o que e como os dados podem ser utilizados.

**Disseminação**: são necessárias estratégias que permitam sua localização, não bastando a simples possibilidade de acesso.

**Preservação**: está diretamente relacionada com sua interpretação, principalmente ao tempo. Deve-se manter um rígido controle sobre seus algoritmos e mecanismos de interação para não gerar resultados distintos a partir de uma base ao longo do tempo.

### 4.2.4 Descarte

Na era do Big Data existe uma discussão sobre a necessidade desta fase do CVD devido ao fato que está cada vez mais barato armazenar grandes volumes de dados, volume

esse que muitas vezes ultrapassa a capacidade de análise dos dados. Nesse sentido cabe uma discussão sobre o descarte de dados que não são mais necessários ou que estejam acima da capacidade de tratá-los com eficiência (SANT'ANA, 2016).

**Privacidade**: todo indivíduo tem por direito e pode requerer que seus dados sejam removidos de alguma base de dados, isso está sendo chamado de direito ao esquecimento.

**Integração**: é importante atenção ao relacionamento entre as tabelas, uma vez que remover os dados de uma tabela sem avaliar o impacto nas outras pode levar a degeneração de relacionamentos entre bases distintas levando a uma degradação do valor de uso da base como um todo.

**Qualidade**: diretamente relacionada com a integração, a remoção disforme de parte da base pode levar a conclusões equivocadas a partir de suas análises comparativas ao longo do tempo ou a partir de contextos diferentes.

**Direito Autoral**: mesmo após o descarte as informações sobre licenças devem ser mantidos pois os dados podem estar sendo ou ter sido utilizados por terceiros.

**Disseminação**: podem ocorrer impactos negativos nos pipelines de dados que contém informações e apontamentos a dados que podem ter sido removidos.

**Preservação**: estratégias de preservação podem ocorrer mesmo com o descarte, pode-se realizar um backup compactado em uma área de armazenamento ou formato físico distinto.

# 5 Procedimentos Metodológicos

Neste capítulo apresentaremos os procedimentos e tecnologias empregados na transformação de Dados Governamentais Abertos em Informação a partir das técnicas de *Data Science*.

Como parte inerente do Ciclo de Vida dos Dados, o consumo e análise desses dados, que ocorre na fase de recuperação, podem ser vistos como um dos principais pontos. O termo análise, apresentado acima, pode ser entendido como mineração de dados, pois tratam de um processo criativo que requer a aplicação de diversas habilidades e conhecimentos.

A utilização de um *framework* ou metodologia é útil para padronizar os projetos de mineração de dados, propiciando um maior controle e previsibilidade, por não depender de pessoas específicas. Essa abordagem padronizada possibilita a transcrição de problemas de negócio em atividades de mineração de dados como transformações e técnicas, além de prover meios para a avaliação da efetividade da análise (WIRTH; HIPP, 2000).

Como *framework* ou metodologia, esse trabalho apresenta o modelo de processos CRISP-DM (*CRoss Industry Standard Process for Data Mining*), que é um projeto que iniciou em 1997 e foi desenvolvido por *Daimler Chrysler*, NCR, SPSS (antiga ISL) e OHRA (Companhia de Seguros Holandesa). E parcialmente patrocinado pela União Europeia por meio do programa, esse projeto tinha por objetivo ESPRIT (WIRTH; HIPP, 2000) e (JACKSON, 2002).

O projeto do CRISP-DM tinha por objetivo definir e validar um modelo de processos que fosse aplicável em qualquer indústria, sem vínculos com ferramentas e podendo ser utilizando tanto *"big"* quanto *"small" data*. Dessa forma, tornando os projetos de mineração de dados mais rápidos, baratos, confiáveis e mais gerenciáveis (WIRTH; HIPP, 2000) e (JACKSON, 2002).

## 5.1 CRISP-DM - Modelo de Referência

O modelo de referência da metodologia CRISP-DM pode ser entendido também como o ciclo de vida de um projeto de mineração de dados e se inicia com o entendimento de negócio. É aí que a definição de escopo e a análise funcional é realizada e, dando suporte a essa primeira etapa, existe a segunda fase. O entendimento dos dados, onde hipóteses são validadas e essas validações são usadas para premissas para a criação de um plano do projeto de mineração de dados (WIRTH; HIPP, 2000).

Após o entendimento, a preparação de dados tem início. Esta etapa contempla a

aquisição e transformações de dados necessárias para o desenvolvimento do conjunto a ser utilizado na modelagem. A fase de preparação é fundamental para a etapa seguinte, de modelagem. Todo conjunto de dados utilizado nas modelagens requer atividades de ajustes, limpezas, transformações. A modelagem consiste na aplicação de diversas técnicas e a parametrização ideal para obtenção dos resultados mais assertivos (WIRTH; HIPP, 2000).

Na fase de validação os modelos mais promissores elaborados na etapa anterior são mensurados a fim de avaliar sua aplicabilidade, assertividade e qualidade. Caso passem a fase de validação os modelos estão prontos para serem implantados na etapa final do ciclo de vida do projeto, caso os modelos não sejam aprovados na validação retorna-se a etapa de entendimento de negócio para identificar as lacunas deixadas pela análise e o projeto recomeça. Essa implantação é a disponibilização do modelo ou análise para o cliente final. Isso pode se dar por meio da apresentação de um relatório ou da implantação de uma peça tecnológica que possibilite o reuso do modelo, entre outras formas (WIRTH; HIPP, 2000).

De forma geral, um projeto de mineração de dados pode ser entendido como tendo seis fases ligadas por setas que indicam as dependências mais importantes e frequentes. de forma que suas atividades antecessoras e sucessoras em forma de dependência são apresentadas na Figura 10. Ainda na Figura 10, o ciclo de vida é representado pelo círculo de setas formado em volta das fases (WIRTH; HIPP, 2000) e (JACKSON, 2002).

### 5.1.1 Detalhamento do Modelo de Referência CRISP-DM

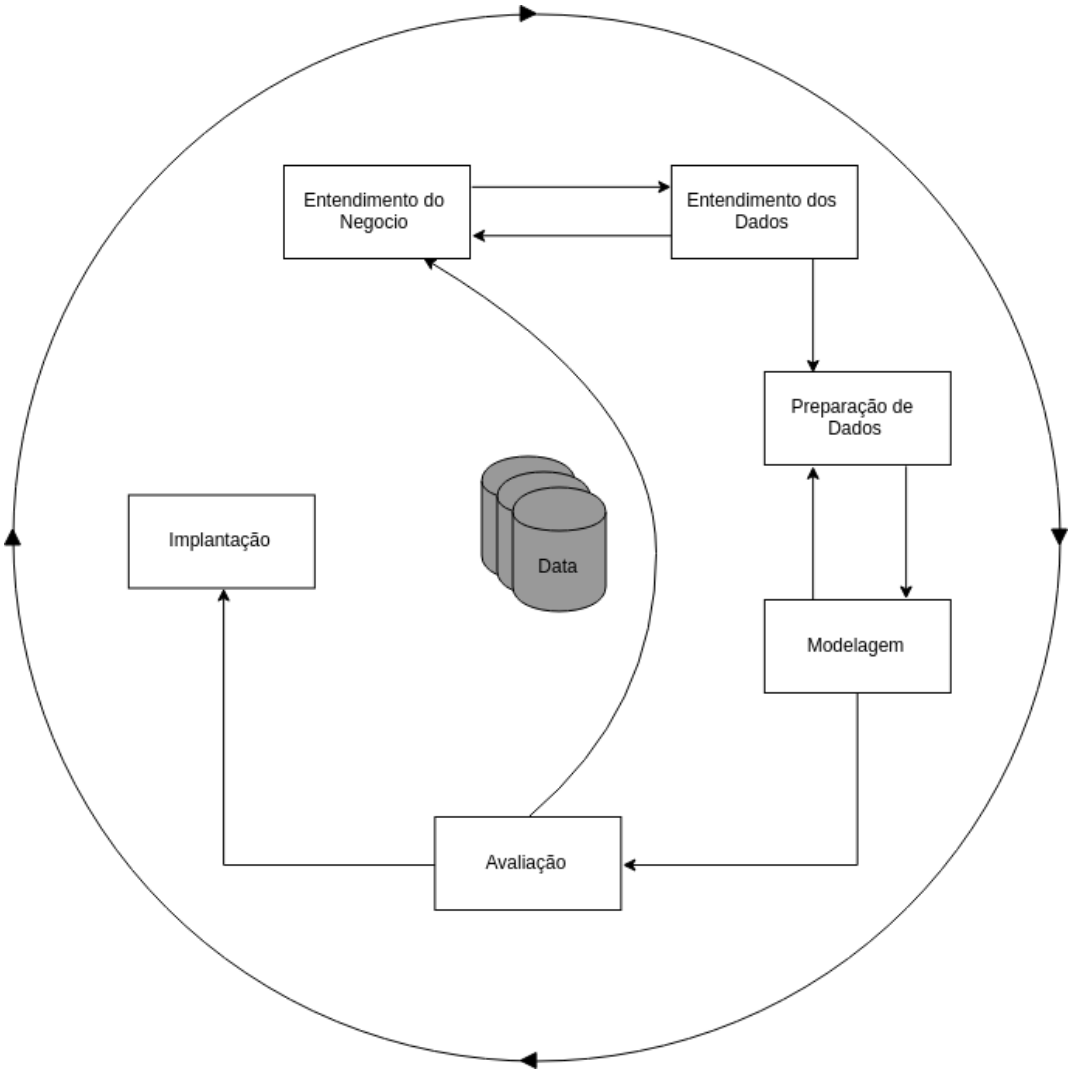
Agora de forma mais específica cada uma das seis fases apresentadas acima têm suas atividades, essas fases e atividades podem ser observadas na Figura 11 e foram descritas por (JACKSON, 2002).

**Entendimento do Negócio**: esta primeira fase tem por objetivo entender e definir o escopo do projeto, além de levantar os requisitos do ponto de vista de negócios. Como entregáveis, essa fase apresenta a conversão dos requisitos de negócio na definição de problema de mineração de dados e a criação de um plano para o projeto visando atingir os objetivos levantados.

Determinar objetivos de negócio (escopo): levantamento de requisitos e negociação do escopo, com viés de negócio. Avaliar situação: detalhamento de requisitos, cruzamento de requisitos técnicos, funcionais e de negócio para entendimento dos requisitos no detalhe. Determinar objetivos da mineração de dados: definição de objetivos técnicos. Produção de plano do projeto: elaboração de um plano de projeto para alcançar os objetivos técnicos e de negócio.

**Entendimento dos Dados**: esta fase está ligada fortemente a fase anterior, pois para determinar os problemas de mineração de dados e o pré projeto é necessário um

Figura 10 – Modelo de Referência CRISP-DM

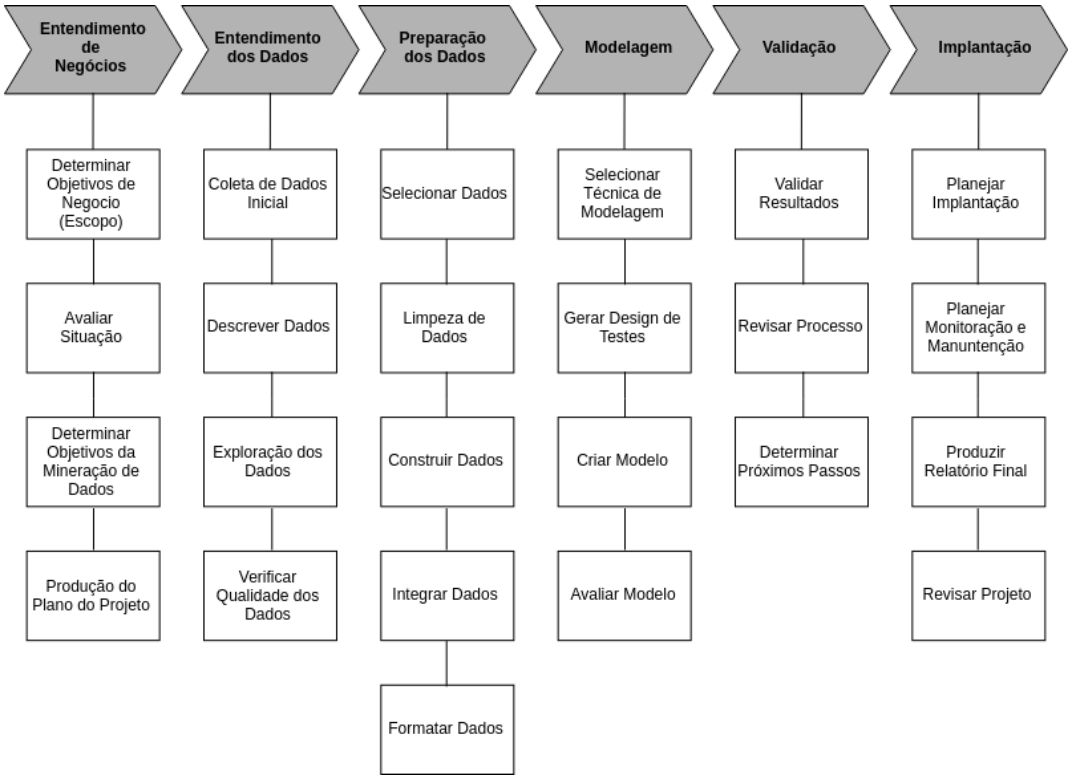



Fonte: Adaptado de Wirth e Hipp (2000)

entendimento mínimo dos dados disponíveis. Além disso, essa etapa começa com a coleta de dados e tem continuidade com a aplicação de diversas técnicas e conhecimentos que facilitem a descoberta de novos conhecimentos a partir dos dados, como também a validação de hipóteses.

Coleta de Dados: como o nome sugere, esse é um processo de aquisição de dados. Descrever Dados: análise descritiva dos dados brutos e suas variáveis, gerando relatório dos resultados. Exploração dos Dados: endereçando questões de mineração de dados, são aplicadas consultas, visualizações e relatórios, com o objetivo de responder às questões de mineração de dados. Verificar Qualidade dos Dados: a validação da qualidade dos dados visa responder aos seguintes questões: a validação da qualidade dos dados visa responder aos seguintes questões: se os dados estão completos, se são corretos e se existem valores

Figura 11 – Modelo de Referência CRISP-DM Detalhado




Fonte: Adaptado de Jackson (2002)

em branco.

**Preparação de Dados**: todas as atividades relacionadas com o desenvolvimento do conjunto de dados final (conjunto de dados que será usado na modelagem). Dentre as atividades realizadas nessa fase estão: seleção de atributos, limpeza de dados, transformação de dados, entre outras.

Selecionar Dados: decidir quais dados farão parte da análise, pelos critérios de relevância, qualidade e restrições técnicas. Limpeza dos Dados: atividade ligada a qualidade dos dados, por melhorar a qualidade dos dados. Construir Dados: criação de atributos derivados, transformação de valores, entre outras operações que viabilizam a preparação dos dados. Integrar Dados: se dá pela fusão dos dados ou pela geração de valores agregados. Formatar Dados: se trata primariamente de modificações sintáticas no dado.

**Modelagem**: nesta fase as técnicas de modelagem são selecionadas e aplicadas e seus parâmetros ajustados para os melhores valores. E para viabilizar a aplicação dessas técnicas de modelagem selecionadas com suas especificidades diversas interação com a etapa de preparação de dados são necessárias.

Selecionar Técnica de Modelagem: baseado nas fases anteriores e nas ferramentas utilizadas no projeto um ou mais modelos podem ser selecionados. Geração de Design de Testes:

antes de construir o modelo é necessário criar o modelo de testes para validar a qualidade e validade da modelagem. Construir Modelo: modelos tem por objetivo ser usado para previsões e apoiar as decisões de negócio e tem como seu principal destaque sua estabilidade. Avaliar Modelo: avaliação técnica dos modelos, geralmente por meio de um gráfico de elevação ou uma matriz de confusão, com foco na precisão e generalização do modelo.

**Validação**: a validação desta fase está ligada aos objetivos de negócio, dessa forma, a avaliação está voltada para a aderência da solução à necessidade do negócio e a capacidade de suprir a necessidade.

Validar Resultados: avalia a aderência da solução a necessidade de negócio e determina as evidências de deficiência de negócio do modelo. Revisar Processo: após a validação técnica e de negócios do modelo é importante revisar o processo que gerou o modelo para garantir que não haja nenhuma tarefa que não tenha sido executada por completo. essa revisão tem um caráter de garantia da qualidade. Determinar Próximos Passos: ponto onde pode-se optar por três principais próximos passos: A) Terminar o desenvolvimento e ir para implantação, B) Refazer modelagem, por meio de nova iteração e C) Começar novo projeto.

**Implantação**: uma implantação é a criação de uma maneira de apresentação do resultado final do projeto, dependendo do escopo pode ser a apresentação de um relatório ou até mesmo o desenvolvimento de uma peça tecnológica que permita a replicação da modelagem indefinidamente.

Plano de Implantação: passo onde se cria e documenta o plano de implantação que deve ser validado pelos *stakeholders*. Plano de Monitoração e Manutenção: uma vez que um artefato tecnológico resultado de uma projeto de mineração de dados se torna parte do dia a dia do negócio a manutenção e monitoração se tornam essenciais para garantir a qualidade do serviço. E é esta a atividade designada para a criação do plano que será utilizado. Produzir Relatório Final: o líder do projeto é responsável por escrever o relatório final e projeto e que pode ser desde um sumário executivo do projeto até uma apresentação profunda sobre os resultados do modelo e o tipo de implementação. Revisão do Projeto: avaliação sobre com as coisas ocorreram, o que deu certo e o que não deu e como melhorar os pontos negativos.

## 5.2 Tecnologias

Esta seção tem por objetivo apresentar algumas das peças tecnológicas que possibilitam a aplicação prática dos conceitos apresentados nos capítulos anteriores, essas peças tecnológicas são passíveis de serem utilizadas em uma análise de dados. Dentre essas peças estão as linguagens de programação Python e R, um exemplo de formato de dados - XML.

Essas são as tecnologias empregadas para viabilizar a utilização de dados governamentais abertos em análises, dessa forma dando subsídios para que dados sejam transformados em informação, possivelmente conhecimento e ação. Essa utilização dos DGA remete ao direito de acesso à informação, transparência, movimento aberto, governo eletrônico e por fim o regime de informação.

### 5.2.1 Python - Linguagem de Programação

Python é uma linguagem de programação de interpretável, software livre, propósito geral e de múltiplos paradigmas, seu lançamento se deu em 1991 pelo programador Guildo van Rossum, como principais destaques estão sua sintaxe simples, riquíssimo ecossistema de bibliotecas online e a grande ênfase na comunidade de desenvolvedores (PERKEL, 2015).

Já a versão 1.0 do Python foi disponibilizada em janeiro de 1994 acrescentava outras funcionalidades da paradigma de programação funcional como funções lambda, map, filter e reduce. Em 2000 Python 2.0 foi lançado com novos recursos como list comprehensions e garbage collection capaz de coletar ciclos de referência. Mais recentemente em 2008 a versão 3.0 do Python foi disponibilizado e devido a uma grande revisão de código não é inteiramente retro compatível com as versões anteriores.

Quanto ao propósito Perkel (2015) afirma que o Python acaba por ser capaz de se ser aplicado aos mais diversos propósitos como automação de pequenos processos, desenvolvimento de web sites, aplicações inteiras e, além disso, pode ser aplicada na programação científica devido sua excelente capacidade de trabalhar com dados. Ainda na programação científica, Perkel afirma que Python pode ser aplicado em todas as fases do desenvolvimento de pesquisas.

Sua sintaxe simplificada torna Python uma linguagem menos dolorosa de ser aprendida quando comparada a C ou C++, além disso, também é mais fácil usar Python pois, a sua execução é iterativa, de forma que digitando um comando obtém-se uma resposta de sua execução (PERKEL, 2015).

A comunidade por trás do desenvolvimento do Python é muito forte, por possuir uma organização sem fins lucrativos, chamada Python Software Foundation e é responsável por promover a linguagem por meio de conferências e pela evolução do Python PSF (2019).

Dessa forma se torna possível manter um ciclo virtuoso, novos usuários estendem a linguagem a novas áreas por meio da criação de novas bibliotecas e melhorias na própria linguagem, o que atrai novos usuários (PERKEL, 2015).

### 5.2.2 Linguagem de Programação "R"

R é um sistema para computação estatística e gráficos que consiste de uma linguagem e um ambiente de desenvolvimento com gráficos. R tem suas origens em duas outras linguagens, sendo elas S de onde se inspira na linguagem e a semântica da *Schema*. A linguagem R foi desenvolvida em 1997 por Ross Ihaka e Robert Gentleman e sua equipe do Departamento de Estatística da Universidade de Auckland, Nova Zelândia (HORNIK, ).

O núcleo R pode ser entendido como uma linguagem interpretada, que permite programação modular com funções. As funções do R nativas ou importadas via bibliotecas geralmente são escritas em R, porém em alguns casos por conta de eficiência outras linguagens podem ser empregadas, sendo essas linguagens: C, C++ e FORTRAN (HORNIK, ).

A *The R Foundation* é uma organização sem fins lucrativos responsável por manter e desenvolver melhorias para a linguagem. Além disso, provê um ponto de referência para indivíduos e organizações para interação em comunidade e administra a licença e documentação do software (HORNIK, ).

A comunidade de R é uma comunidade ativa e por R ter uma alta capacidade de extensibilidade. Existe uma grande disponibilidade de bibliotecas que dão suporte aos mais variados tipos de análises, como regressões lineares e não lineares, testes estatísticos clássicos, análises de série temporal, classificação, clusterização, entre outras (HORNIK, ).

### 5.2.3 XML - Formato de Arquivo

A eXtensible Markup Language (XML) como o próprio nome sugere é uma linguagem de marcação criada por Jon Bosac, na época vinculado a empresa de software Sun®. Definida como padrão de marcação para ser utilizado na Internet, é uma versão simplificada do SGML que tem como objetivo prover uma maneira de definir e criar marcadores e atributos, em vez de estar condicionada aos esquemas de marcação do HTML. Assim como a linguagem HTML a linguagem XML é um padrão aberto e independente de plataforma, porém preocupa-se em criar estruturas para representar objetos informacionais (FURGERI et al., 2006).

Na XML cada unidade de informação é delimitada por meio de uma tag que fornece significado, fazendo com que seja interpretável tanto por pessoas como por máquinas em decorrência da facilidade na identificação das estruturas chave-valor que compõe a unidade de informação (FURGERI et al., 2006).

De acordo com Rodriguez (2002, apud Furgeri et al. (2006)), a utilização de marcadores (tags) possibilita o detalhamento das informações sobre livros, dessa forma são criados dados sobre os livros individualmente, isto é, metadados. Quanto ao nível de detalhamento (Rodriguez 2002 apud Furgeri et al. (2006)) afirma que:

> "O nível de detalhamento dos dados pode ser ampliado na medida da necessidade. Por exemplo, a *tag* editora poderia conter outros elementos filhos, como nome, endereço, estado e assim, detalhando ainda mais a estrutura e ampliando a representação da informação. Com esse pequeno exemplo é possível observar que a XML constitui um importante recurso na criação de metadados, que por sua vez, constituem-se num recurso de vital importância para a representação e recuperação da informação" (RODRIGUEZ, 2002 apud (FURGERI et al., 2006)).

### 5.2.4 Fontes de dados

Neste trabalho as técnicas de Ciência de Dados foram empregadas para analisar dados de duas fontes governamentais: poder legislativo e poder executivo federal.

A análise dos dados do poder legislativo federal, dedicou-se a apresentação de um "raio-X" dos representantes da Câmara de Deputados Federais e do Senado. Em termos de recursos tecnológicos, na presente análise, foi utilizada a linguagem de programação R R Core Team et al. (2013) e pacote congressbr.

Já o congressbr, conforme apresentado pelos desenvolvedores, é um pacote para extração de dados das APIs do Senado Federal Brasileiro e Câmara Federal de Deputados, respectivamente. Em outras palavras, ainda de acordo com os autores, o pacote faz o download e trata os dados das APIs da Câmara e Senado (MCDONNELL et al., 2017).

A principal função do pacote congressbr é encapsular as chamadas da API, dessa forma abstraindo a complexidade das mesmas. Mais informações sobre as APIs da câmara[1] e do senado[2].

A análise dos dados do poder executivo federal foi um processamento análise textual, também conhecida como processamento de linguagem natural, dos dados do Diário Oficial da União, obtidos no portal[3] de DGA oficial do Governo Federal Brasileiro.

Os dados do DOU são disponibilizados mensalmente, em pacotes compactados separados por sessão. Dentro de cada pacote compactado existem milhares de arquivos XML, onde cada um representa uma publicação é importante ressaltar que ter os arquivos em formato XML que é um padrão aberto e é possível de ser lido por máquinas.

Esta análise foi desenvolvida utilizando duas linguagens de programação, sendo a primeira Python, para a extração do dos dados dos arquivos XML e conversão para o formato tabular, e a outra linguagem de programação utilizada foi R, utilizada para realizar a mineração de texto e construção do *wordcloud*.

[1] <https://dadosabertos.camara.leg.br/swagger/api.html>
[2] <http://legis.senado.gov.br/dadosabertos/docs/>
[3] <http://www.dados.gov.br>

# 6 Apresentação dos Resultados

O objetivo deste capítulo é apresentar as inter-relações entre os tópicos socioculturais, apresentados nos Capítulos 2 e 3, com os tópicos técnicos, apresentados no capítulo 4 e operacionalizados no Capítulo 5. Nesse sentido, é importante demonstrar que os pontos podem se complementar e até mesmo estender como, por exemplo, quando se pensa no ciclo de vida dos dados e uma metodologia como a CRISP-DM sendo aplicados aos DGA.

Durante a discussão propostas neste capítulo, quando possível, serão apresentados exemplos obtidos na prática da execução de duas análises, disponíveis como Apêndices B e C. Ambas as análises foram baseadas em Dados Governamentais Abertos - DGA e seguiram a metodologia CRISP-DM em seu desenvolvimento, utilizando software livre e trazendo como motivação e referenciais socioculturais baseados nos tópicos previamente apresentados nos Capítulos 2 e 3.

Alem disso, a motivação para a realização de duas analises é a diversificação de cenários, como o da coleta de dados, formato do dado e qualidade do dado. No Apêndice B a coleta do dado é facilitada por poder ser realizada mecanicamente pelo uso de um artefato de software, tendo seu formato bem definido pelo uso de linguagem de programação e uma excelente qualidade. Em contra partida, na analise do Apêndice C o dado é coletado manualmente, apesar de estar em formato XML peca muito na qualidade.

Para contextualizar essas análises, segue uma rápida descrição das mesmas.

## 6.1 Exploração de Dados do Congresso e Senado Brasileiro utilizando APIs para coleta de dados

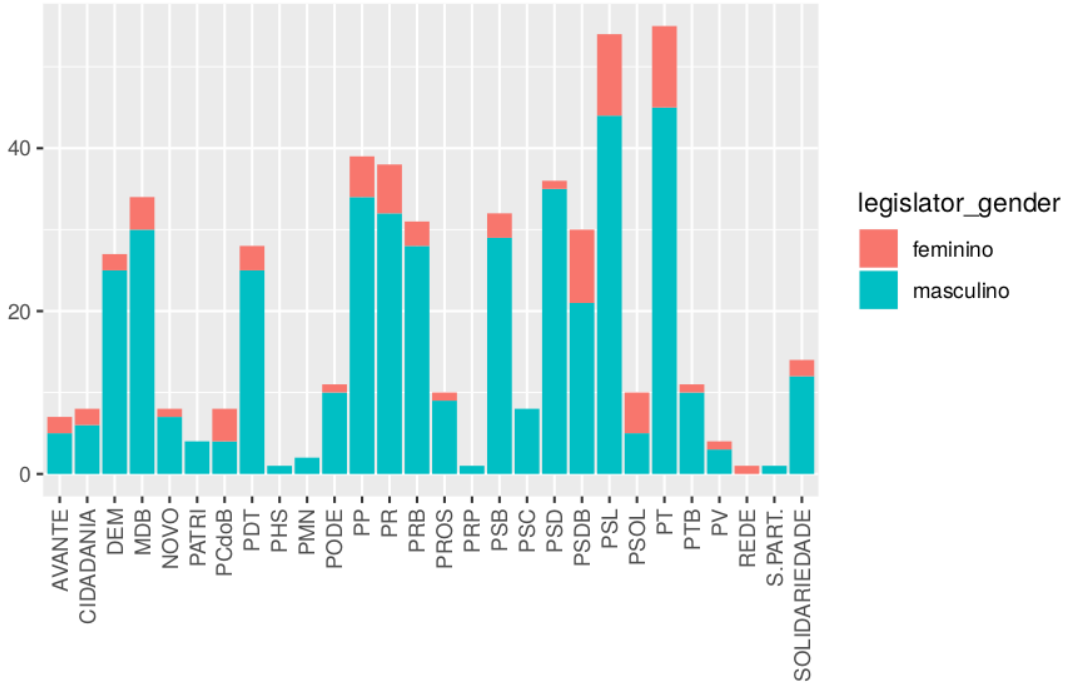
Este projeto é uma análise simples, focada em realizar uma exploração de dados sobre a composição das bancadas do congresso e senado. Portanto, foi utilizada a linguagem R e a biblioteca congressbr para coletar os dados.

Com os dados coletados, alguns indicadores foram analisados e apresentados de forma gráfica. Para essa apresentação visual foi utilizada a biblioteca ggplo2, que se trata de uma biblioteca amplamente utilizada e é especializada em gráficos, provendo diversas possibilidades de customização.

Um dos indicadores apresentados é a proporção de representantes de cada gênero por casa, por casa e partido e, no caso do Senado, por casa e coalizão. Outro indicador interessante é a composição das coalizões por partidos e como se dá essa distribuição.

Com o intuito de demonstrar parte da exploração de dados realizada, serão apre-

Figura 12 – Quantidade de Parlamentares por Partido e Gênero na Câmara Federal




Fonte: Autor

sentados exemplos da exibição gráfica dos indicadores. Podem-se observar as Figuras 12 e 13, onde são apresentadas as quantidades de parlamentares por partido e gênero da Câmara e Senado respectivamente.

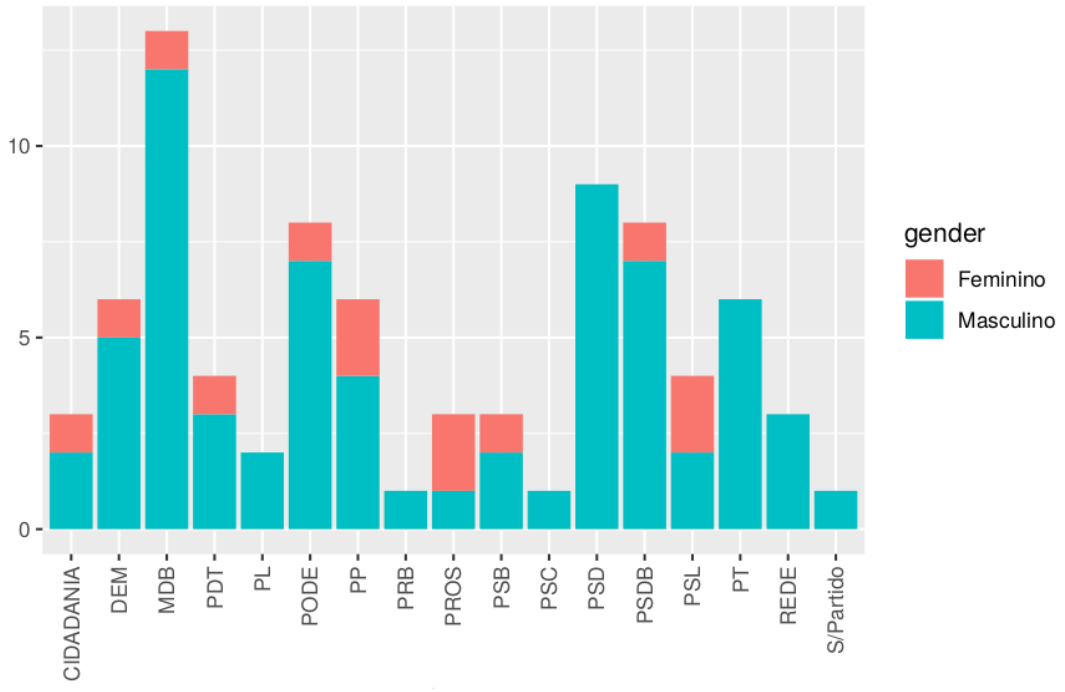
Em adição as métricas apresentadas acima, também foram disponibilizadas outras como, no caso do Senado conforme apresentado na Figura 14. Além disso, a representatividade dos partidos nas comissões. Ainda nas comissões, foi demonstrado o número de representantes por comissão, focado naquelas 10 com mais membros e, por fim, os 15 senadores que mais participam de comissões.

Pensando na execução desta análise sob a perspectiva da metodologia CRISP-DM, pode-se afirmar que é uma análise de ciclo incompleto pois contempla apenas as duas primeiras fases Entendimento do Negócio e Entendimento dos Dados.

## 6.2 Processamento de Linguagem Natural dos Dados do Diário Oficial da União (DOU)

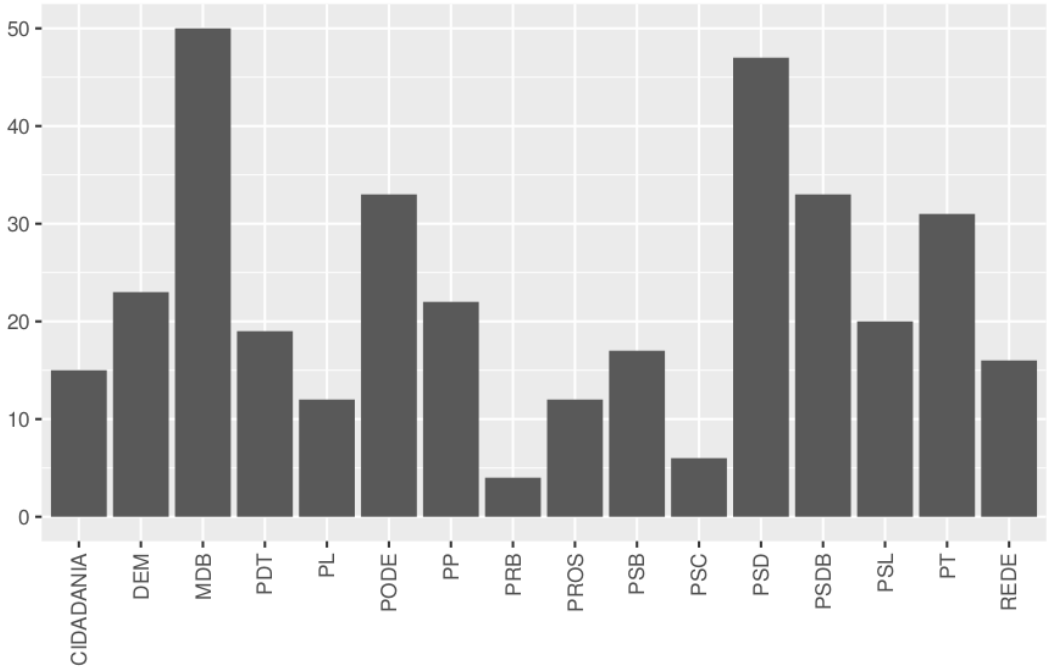
Atualmente, os dados do DOU são disponibilizados em formato XML. Anteriormente, os dados eram disponibilizados em PDF, que é um formato proprietário criado pela ADOBE, que não é amigável para a extração dos dados, mas que é mais facilmente lido

Figura 13 – Quantidade de Parlamentares por Partido e Gênero no Senado Federal




Fonte: Autor

Figura 14 – Quantidade de Parlamentares por Partido em Comissões do Senado Federal




Fonte: Autor

por pessoas. Entretanto, apesar do avanço pela adoção do formato XML, alguns pontos continuam longe do ideal, como por exemplo:

- Não existe um XML *Schema* disponível para facilitar o intercâmbio da informação;
- Não existe um dicionário de dados que explique o cada variável representa;
- Cada matéria em um XML individual torna o processo de extração dos dados lento e mais complexo do que se estivessem todas as matérias em um único arquivo;
- Apesar das imagens que acompanham algumas das publicações serem disponibilizadas juntamente com os arquivos XML, não existe nenhuma forma de correlacionar as imagens com a publicação a que pertence.

Outro ponto a ser destacado, são os metadados. Pelo lado positivo, eles existem parcialmente e estão dispostos no cabeçalho de cada matéria (arquivo XML). Porém, pelo lado negativo, pode-se apontar que faltam metadados do pacote mensal de dados, dificultando a catalogação do arquivo original de dados.

Como produto final da análise foi elaborado um relatório usando a biblioteca Knirt e como ponto alto do relatório pode-se destacar a nuvem de palavras (*wordcloud*) que pode ser observada na Figura 15. Essa nuvem de palavras representa os termos mais importantes contidos nas três sessões do DOU do dia 31/01/2019.

## 6.3 Modelo de CDV utilizando CRISP-DM voltado ao DGA

Tanto o modelo de ciclo de vida dos dados, como a metodologia CRISP-DM, foram pensados de forma genérica. Isso por um lado é excelente, pois com pequenas adaptações podem atender requisitos diferentes. Já por outro lado, essa abordagem genérica acaba por permitir extrair todo o proveito possível do segmento, que neste trabalho são os DGA.

Nesse sentido, utilizando-se dos referenciais teóricos dos capítulos 2 e 3, é possível identificar uma interdependência e relações de precedência entre os temas teóricos com os práticos. Essa relação pode ser notada quando analisamos os princípios de acesso à informação com os princípios do Governo Aberto e as características e princípios dos DGA.

Essa aproximação ou precedência é observada quando, nos princípios do acesso à informação, existem itens como o princípio da divulgação máxima, obrigação de publicar, promoção de um Governo Aberto e a divulgação tem precedência. Dessa forma, levando ao Governo Aberto que, dentre seus princípios, exalta a transparência como forma de prestação de contas, a participação dos cidadãos e a oferta colaborativa de serviços e parceria.

Figura 15 – Nuvem de palavras do DOU do dia 31-01-2019


Fonte: Autor

Indo mais um passo à frente chegando aos Dados Governamentais Abertos, suas características apresentam forte tendência a abertura, como observado a seguir:

- Os dados devem estar disponíveis de uma forma conveniente e modificável;
- Para reutilização, os dados deve ser legíveis por máquina;
- Participação universal: todos devem poder usar, reutilizar e redistribuir.

Continuando na relação dos DGA, seus princípios também estão altamente de acordo com temas referentes do acesso à informação e movimento aberto como um todo. Estes princípios impõem que os dados sejam:

- Completos;
- Primários;
- Atuais;
- Acessíveis;
- Processáveis por máquina;
- De acesso não discriminatório;
- Formato não proprietário.

Ainda demonstrando o entrelaçamento dos temas, o grupo *Civic Analytics Network* – CAN, mencionado anteriormente, apresenta diretrizes que estimulam o Governo Eletrônico, Governo Aberto, Dados Abertos e Dados Governamentais Abertos. Sendo essas diretrizes:

- Melhorar a acessibilidade e usabilidade aos dados;
- Tratar dados geoespaciais;
- Melhorar o gerenciamento e usabilidade dos metadados;
- Manter histórico de versões e temas referentes aos custos operacionais;
- Cobrança justa pela utilização dos dados.

Assim, iniciando no regime de informação, entende-se que os temas que ele abarca, como Governo Eletrônico, direito de acesso à informação e todo o movimento aberto, culminam nos DGA. Entendido que o regime de informação abarcará todos os outros temas relacionados, entende-se que ele precede e é base para o ciclo de vida dos dados.

Além disso, como os resultados de análises de dados suportados pela metodologia CRISP-DM se encaixam no CVD e como são início a um novo processo de CVD de dados. Dessa forma, com as bases teóricas do Regime de Informação provendo a sustentação para o desenvolvimento, criação e idealização de um Ciclo de Vida dos Dados que tem por objeto os DGA e como método o CRISP-DM.

O Ciclo de Vida dos Dados Governamentais Abertos (CVDGA) trata-se de uma extensão especializada do CVD e é fruto de tópicos como os Dados Governamentais Abertos e as demais bases teóricas, apresentadas capítulos 2 e 3, quando analisadas em perspectiva do Ciclo de Vida dos Dados.

Essa análise em perspectiva se faz possível quando as sobreposições e extensões de discussão de temas correlatos são examinadas de perto. Como a aproximação dos pilares do CVD com os princípios do Acesso à Informação, que temos os procedimentos e processos

que facilitem os acessos como um princípio do Acesso à Informação e os pilares de CVD Disseminação e Preservação.

Em contrapartida, existe o pilar da privacidade, que é colocado em perspectiva pelos princípios da divulgação máxima e com precedência. Além de que, de forma a dar exceções a esses princípios, existe outro princípio que é a existência de um escopo limitado de exceções que atenta para casos diferenciados.

Ao contrapor os princípios do CVD com algumas características da Cultura Aberta e, por consequência, dos DGA, observa-se a consonância do princípio da disseminação com a característica de disponibilidade de acessos e a da característica da cultura aberta de reutilização e acesso com o princípio da integração.

Dessa forma como representado na Figura 16, o CVDGA é responsabilidade compartilhada. Por um lado, o poder público, que é a fonte de informação, deve atuar para que os dados estejam disponíveis respeitando todas as características, premissas, princípios e pilares apresentados nesse trabalho. Por outro lado, a sociedade deve ter uma atuação ativa para garantir que o CVDGA exista e seja respeitado, além de também contribuir para que o processo, sempre que possível, seja por meio de atuação direta transformando dados, criando análises e gerando valor ou, de maneira indireta, apoiando essas iniciativas.

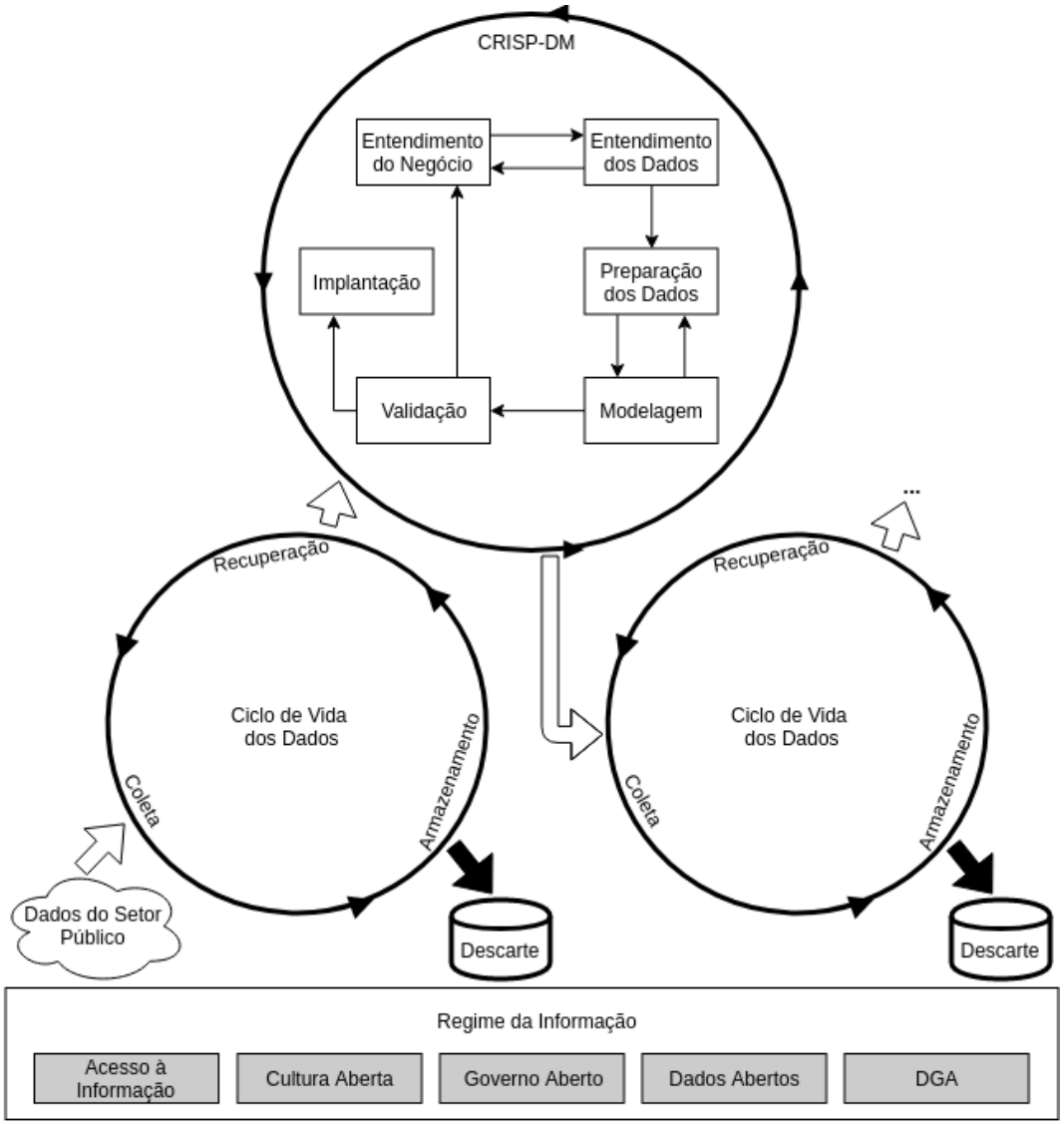
Conforme apresentado na Figura 16, pode-se entender que transformação de dados apresentada acima pode ser substituída pela Metodologia CRISP-DM, sendo aplicada a Dados Governamentais Abertos. Dessa forma agregando uma metodologia de mineração de dados ao CVDGA, que tem por objetivo padronizar, estabelecer e servir como referência para projetos que visem a análise e transformação de dados públicos.

## 6.4 Exemplificação do processo CVD com a Aplicação da Metodologia CRISP-DM

O processo de ciclo de vida dos dados governamentais abertos se inicia na geração ou captação do dado pela agência governamental, passa por todos os processos usuais do ciclo de vida de dados como armazenamento, recuperação e descarte. Durante esse processo, mais especificamente na fase de recuperação os dados foram coletados a fim de serem analisados e com isso se inicia um novo ciclo de vida de dados a partir da transformação dos dados originais.

Para o desenvolvimento das análises apresentadas Exploração de Dados do Congresso e Senado Brasileiro utilizando APIs para coleta de dados - Apêndice B e Processamento de Linguagem Natural dos Dados do Diário Oficial da União (DOU) - Apêndice C foram considerados as implicações do CVD juntamente com a Metodologia CRISP-DM, dessa forma, pode se entender que o processo CVDGA foi utilizado como base para estes

Figura 16 – Ciclo de Vida dos Dados Governamentais Abertos




Fonte: Autor

desenvolvimentos.

Focando na aplicação da metodologia CRISP-DM, abaixo segue a exemplificação de como a metodologia foi aplicada nas análises anexas.

### 6.4.1 Exploração de Dados do Congresso e Senado Brasileiro utilizando APIs para coleta de dados - CVDGA

Conforme já apresentado a primeira etapa da metodologia CRISP-DM é o entendimento do negócio e nesse sentido o escopo da análise foi determinado como uma exploração de dados como objetivo de apresentar indicadores comparativos sobre a composição das casas Câmara e Senado. Além disso, foi determinado o uso da biblioteca R cogressbr de

McDonnell et al. (2017), por abstrair o uso da API que provê esses dados.

Na fase seguinte, entendimento dos dados, foram realizadas diversas consultas a API com o auxílio da biblioteca congressbr para determinar quais dados seriam utilizados na análise, validando sua qualidade e disponibilidade. A etapa subsequente a terceira, preparação dos dados envolve atividades de seleção, limpeza e integração dos dados para possibilitar a análise.

Como o objetivo da análise é uma análise exploratória, essas três primeiras etapas abrangem quase todo o trabalho realizado, faltando apenas a última fase que é a implantação. Que ocorreu por meio da disponibilização de um relatório que apresenta as comparações de indicadores, gerado na linguagem R utilizando a biblioteca Knitr, disponível como Apêndice C.

### 6.4.2 Processamento de Linguagem Natural dos Dados do Diário Oficial da União (DOU) - CVDGA

A primeira fase entendimento do negócio foi muito importante para obter um entendimento do objeto de estudo da análise, que são as publicações no diário oficial da união. Durante esse passo foi definido que uma análise focada em processamento de linguagem natural seria aplicada para melhor entender e representar os dados e informações contidas nos DGA do DOU.

Sucedida pelo entendimento de dados, onde ocorreu a coleta de dados da página oficial do governo federal de DGA (http://www.dados.gov.br), fechando o escopo nos dados das três sessões do mês de Janeiro de 2019. Além da coleta, foi realizada a exploração de dados e validações de qualidade a fim de garantir a consistência da análise.

Na terceira etapa, preparação dos dados, foram realizados diversos processos de limpeza de dados, criação de variáveis calculadas e geração do conjunto de dados à ser usado na fase seguinte a modelagem. Na modelagem, como o nome diz o modelo foi criado, por meio da remoção de palavras não desejadas na análise como, por exemplo, preposições, artigos entre outras palavras. Após isso, foi realizada uma indexação de termos para determinar aqueles mais frequentes e apresenta-los em forma de nuvem de palavras.

Na quinta fase validação, foram executados diversos testes para analisar o resultado final e aprimorar a modelagem a fim de obter a análise mais representativa de termos de interesse. A implantação como a análise anterior, foi criado um relatório utilizando R e a biblioteca Knitr, disponível como Apêndice C.

# 7 Considerações Finais e Trabalhos Futuros

Partindo do título, “Ciência de Dados aplicada a Dados Governamentais Abertos sob a ótica da Ciência da Informação”, este trabalho visa apresentar o contexto, bases teóricas que abrangem as origens e motivação para aplicação de técnicas e metodologias ligadas à Ciência de Dados.

Esses contextos e bases teóricas podem ser divididos em dois grandes grupos: teórico e técnico. O primeiro aborda o Regime de Informação, sendo que este pode ser entendido sucintamente como o relacionamento de atores, tecnologias, representações, normas e padrões regulatórios que configuram políticas implícitas ou explícitas de informação.

Estes temas tratam do Governo Eletrônico, que aborda a adoção das TICs pelo setor governamental, representado pela informatização do serviço público em suas atividades internas e externas. Além disso, o advento do e-gov é resultado da aproximação dos nós entre todos os atores: governo eletrônico, cidadãos, empresas, terceiro setor. Essa informatização do serviço público possibilitou maior acesso dos atores externos com o governo, o que remete ao direito de acesso à informação.

O acesso à informação é um direito humano universal, desde 1946, quando sua resolução foi aprovada na ONU e implica no direito de coletar, transmitir e publicar informações. Dessa forma, o acesso à informação é parte constituinte da transparência e acarreta em conduzir os assuntos governamentais de forma aberta. De forma a dar apoio a Transparência e Acesso à Informação o Movimento Aberto ou a Cultura Aberta ressurge nos anos 80 em meio ao movimento do software livre com um conjunto de conceitos como colaboração, participação e transparência.

A primeira parte apresenta as bases teóricas do trabalho, além de servir como plataforma para os capítulos subsequentes. De forma mais específica, o Movimento ou Cultura Aberta pavimenta caminho para algumas das especialização do conceito de aberto são discutidas. Além disso, também existe uma relação entre os tópicos subsequentes ao Regime de Informação com todos os temas que o seguem.

Em seguida, são apresentadas algumas especializações do conceito de aberto que são relevantes a este trabalho, sendo elas o Governo Aberto, Dados Abertos e os Dados Governamentais Abertos. Estes temas têm alto grau de correlação entre si e com os temas discutidos anteriormente, como E-Gov, que traz a informatização necessária para viabilização da aplicação dos sub temas de aberto. O Acesso à Informação e a Transparência colocam as fronteiras legais e socioculturais para o estímulo a aplicação do GA e DGA.

O Governo Aberto, primeira especialização de aberto abordada, pode ser entendido

como a relação com transparência e/ou acesso à informação, além da relação com governo eletrônico e/ou avanços nas TICs e internet. Além disso, transparência das ações do governo, a acessibilidade dos serviços e informações governamentais e a capacidade de resposta do governo a novas ideias, demandas e necessidades.

Já os Dados Abertos são uma forma de abertura focada no compartilhamento de informações de forma agnóstica, em relação à origem, e são definidos como dados que podem ser usados livremente, compartilhados e incorporados por qualquer pessoa, em qualquer lugar, para qualquer finalidade.

Da junção de Governo Aberto e Dados Aberto surgem os Dados Governamentais Abertos que herdam as características dos seus dois antecessores. Eles podem ser entendidos como dados do governo ou Informações do Setor Público (ISP) e são quaisquer dados e informações produzidos ou comissionados por Organismos do Setor Público que qualquer um pode usar para qualquer propósito, sem restrições.

Em outras palavras, Governo Aberto - GA é a disponibilização de informações em qualquer formato por parte de entidades governamentais e outras ações que promovam transparência. Dados Abertos - DA é a disponibilização de dados em formato padronizado por qualquer tipo de entidade (pública, privada, terceiro setor etc.). Dados Governamentais Abertos - DGA é a disponibilização de dados em formato padronizado por entidades governamentais.

De forma a sintetizar as aproximações e interrelações apresentadas acima, pode-se entender que E-Gov sustenta o GA e DGA, por conta do processo informatização do setor público. Os temas do Acesso à Informação e Transparência demonstram a necessidade da abertura por meio de legislações e acordos. Somando-se a isso, a Cultura Aberta ponto de partida do GA, DA e DGA como norteadores dos princípios de abertura. Todos esses tópicos com suas características individuais e em suas aproximações compõe Regime de Informação apresentado neste trabalho e também formam as bases teóricas do trabalho.

Com o Regime de Informação delineado e apresentado juntamente com seus temas correlatos, são trazidos os aspectos técnicos que, sustentados pelas bases teóricas, propiciam a aplicação da técnica. Estes aspectos cobrem a Ciência de Dados e seus tópicos correlatos, como Ciclo de Vida de Dados, Metodología CRISP-DM e tecnologias envolvidas.

A Ciência de Dados é apresentada neste trabalho como campo científico, que desenvolve metodologias, teorias, tecnologias e aplicativos relevantes para dados. Ela aborda desde a captura, criação, representação, armazenamento, pesquisa, compartilhamento, privacidade, segurança, modelagem, análise, aprendizagem, apresentação e visualização, até a integração de recursos complexos, heterogêneos e interdependentes para a tomada de decisões em tempo real, colaboração, criação de valor e suporte à decisão.

Sua aproximação com a Ciência da Informação começa no objeto de estudo que é o

Dado, esta aproximação torna-se relevante pelo contraponto que a CI traz no tocante da crença dos Cientistas de Dados na neutralidade e objetividade dos dados. A CI apresenta que os dados têm vieses e não são fatos objetivos e desinteressados, mas a coleta de informações de acordo com metas e pressupostos específicos. Outros pontos de interseção entre CI e CD são o controle de qualidade de dados, os bibliotecários de dados suas atividades e funções e a teoria dos documentos.

Ainda tratando de aproximações entre Ciência da Informação e Ciência de Dados, o Ciclo de Vida dos Dados é um ponto de interesse mútuo que a CI contribui de forma decisiva. O CVD pode ser entendido como um conjunto de 4 fases e 6 pilares que determinam como os processos devem ser executados para garantir um ciclo de vida de dados saudável. Suas fases são Coleta, Armazenamento, Recuperação e Descarte, seus pilares são Privacidade, Integração, Qualidade, Direitos Autorais, Disseminação e Preservação. Além disso, outro ponto importante no CVD é que um novo dado gerado a partir de uma origem inicia um novo CVD próprio.

Como metodologia para a Ciência de Dados é apresentada a Metodologia para Mineração de Dados - CRISP-DM (*CRoss Industry Standard Process for Data Mining*), que se trata de um *framework* para projetos de mineração de dados. É útil para padronizar os projetos de mineração de dados, propiciando um maior controle e previsibilidade, por não depender de pessoas específicas. O modelo é composto de 6 fases, sendo elas Entendimento do Negócio, Entendimento dos Dados, Preparação de Dados, Modelagem, Validação e Implantação.

Tratando das tecnologias relacionadas a Ciência de Dados, foram apresentadas as seguintes: Linguagens de Programação: R e Python e o Formato de arquivo XML. Dessa forma encerrando o arcabouço técnico trabalho e abrindo espaço para a aplicação da técnica baseado na teoria com um objetivo prático de demonstrar a aplicação de técnicas de Ciência de Dados aplicada a Dados Governamentais Abertos sob a ótica da Ciência da Informação.

Tendo tudo isso em vista, este trabalho procurou responder ao problema de pesquisa por meio da apresentação de fundamentação teórica e técnica, culminando na aplicação práticas dos conceitos apresentados. Isso com o intuito de demonstrar que é possível utilizar de técnicas e metodologias de Ciência de Dados para realizar as análises e transformações de dados necessárias a fim de produzir informação sobre as atividades das organizações governamentais tendo como princípio a Ciência da Informação.

Como trabalhos futuros, a partir deste, existem três grandes vertentes. A primeira, que é o aprofundamento da pesquisa sobre a relação entre a Ciência de Dados e a Ciência da Informação e as suas influências. A segunda está ligada a parte técnica, aplicação de métodos específicos em conjuntos de dados de DGA específicos, como, por exemplo, alguma técnica de clusterização usando os resultados de votações da Câmara e Senado

como entrada de dados. Por fim, a terceira vertente é relacionada a interação DGA, CI e CD por meio de aplicações práticas.

# Referências


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) — Universidade de São Paulo, 2014.

ALVES, M. V. C. Portais de governo uma avaliação na câmara dos deputados. *Brasília: Câmara dos Deputados, Edições Câmara*, 2012.

ASSEMBLY, U. G. Calling of an international conference on freedom of information. *Resolution*, v. 59, n. 1, p. 14, 1946.

ASSEMBLY, U. G. Universal declaration of human rights. *UN General Assembly*, New York, NY, USA:, 1948.

AYANKOYA, K.; CALITZ, A.; GREYLING, J. Intrinsic relations between data science, big data, business analytics and datafication. In: ACM. *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology*. [S.l.], 2014. p. 192.

BIRKINSHAW, P. Freedom of information and openness: Fundamental human rights. *Admin. L. Rev.*, HeinOnline, v. 58, p. 177, 2006.

BLACK, J.; HASHIMZADE, N.; MYLES, G. *A dictionary of economics*. [S.l.]: OUP Oxford, 2012.

BRAMAN, S. The emergent global information policy regime. In: *The emergent global information policy regime*. [S.l.]: Springer, 2004. p. 12–38.

BUCKLAND, M. K. Information as thing. *Journal of the American Society for information science*, Wiley Online Library, v. 42, n. 5, p. 351–360, 1991.

BUSH, V.; BUSH, V. As we may think. *Resonance*, Indian Academy of Sciences Bengaluru, v. 5, n. 11, 1945.

CAN, C. A. N. *An Open Letter to the Open Data Community | Data-Smart City Solutions*. 2017. Disponível em:<https://datasmart.ash.harvard.edu/news/article/an-open-letter-to-the-open-data-community-988>. (Accesso em: 07/04/2019).

CAN, C. A. N. *An Open Letter to the Open Data Community: One Year Later | Data-Smart City Solutions*. 2018. Disponível em:<https://datasmart.ash.harvard.edu/news/article/letter-open-data-community-one-year-later>. (Accesso em: 07/04/2019).

CAO, L. *Data science and analytics: a new era*. [S.l.]: Springer, 2016.

CAO, L. Data science: Nature and pitfalls. *IEEE Intelligent Systems*, IEEE, v. 31, n. 5, p. 66–75, 2016.

CAO, L. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, ACM, v. 50, n. 3, p. 43, 2017.

CAO, L. Data science: challenges and directions. *Communications of the ACM*, ACM, v. 60, n. 8, p. 59–68, 2017.

CARTER, D.; SHOLLER, D. Data science on the ground: Hype, criticism, and everyday work. *Journal of the Association for Information Science and Technology*, Wiley Online Library, v. 67, n. 10, p. 2309–2319, 2016.

CASTELLS, M. *The power of identity*. [S.l.]: John Wiley & Sons, 2011. v. 14.

CHATWIN, M.; ARKU, G. Beyond ambiguity: Conceptualizing open government through a human systems framework. *eJournal of eDemocracy & Open Government*, v. 9, n. 1, 2017.

CLIFT, S. L. E-government and democracy. *Representation and citizen engagement in the information age*, v. 40, 2004.

DAVENPORT, T. *Big data at work: dispelling the myths, uncovering the opportunities*. [S.l.]: Harvard Business Review Press, 2014.

DAVIES, T.; PERINI, F. Researching the emerging impacts of open data: revisiting the oddc conceptual framework. *The Journal of Community Informatics*, v. 12, n. 2, 2016.

DINIZ, E. H. et al. O governo eletrônico no brasil: perspectiva histórica a partir de um modelo estruturado de análise. *Revista de Administração Pública*, v. 43, n. 1, p. 23–48, 2009.

EAVES, D. The three laws of open government data. *Eaves. ca*, v. 30, p. 8, 2009.

FANG, Z. E-government in digital era: concept, practice, and development. *International journal of the Computer, the Internet and management*, v. 10, n. 2, p. 1–22, 2002.

FROHMANN, B. Taking information policy beyond information science: applying the actor network theory. In: CITESEER. *ANNUAL CONFERENCE OF THE CANADIAN ASSOCIATION FOR INFORMATION SCIENCE/ASSOCIATION CANADIENNE DES SCIENCES DE L'INFORMATION*. [S.l.], 1995. v. 23.

FURGERI, S. et al. Representação de informação e conhecimento: estudo das diferentes abordagens entre a ciência da informação e a ciência da computação. Pontifícia Universidade Católica de Campinas, 2006.

GARTNER. *Open Data - Gartner IT Glossary*. 2019. Disponível em:<https://www.gartner.com/it-glossary/open-data/>. (Accesso em: 07/04/2019).

GOMEZ, M. N. G. d. Novos cenários políticos para a informação. Ibict, p. 34, 2002.

GOMEZ, M. N. G. d. Regime de informação: construção de um conceito. *Informação & sociedade: estudos*, v. 22, n. 3, 2012.

GOMEZ, M. N. G. d.; CHICANEL, M. A mudança de regimes de informação e as variações tecnológicas. 2013.

GRIGORESCU, A. International organizations and government transparency: Linking the international and domestic realms. *International Studies Quarterly*, Blackwell Publishing Ltd Oxford, UK, v. 47, n. 4, p. 643–667, 2003.

GRUMAN, M. Lei de acesso à informação: notas e um breve exemplo. *Revista debates*, v. 6, n. 3, p. 97, 2012.

HEALD, D. A. Varieties of transparency. In: *Transparency: The Key to Better Governance?: Proceedings of the British Academy 135*. [S.l.]: Oxford University Press, 2006. p. 25–43.

HORNIK, K. *R FAQ*. (Accesso em: 2019). Disponível em: <https://CRAN.R-project.org/doc/FAQ/R-FAQ.html>.

INTERNATIONAL, T. *The Anti-Corruption Plain Language Guide*. [S.l.]: Transparency International Berlin, 2009.

JACKSON, J. Data mining; a conceptual overview. *Communications of the Association for Information Systems*, v. 8, n. 1, p. 19, 2002.

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.

JARDIM, J. M.; ALMEIDA, C. H. M. d. Políticas de informação governamental: a construção de governo eletrônico na administração federal do brasil. 2012.

KOTKA, T.; VARGAS, C.; KORJUS, K. Estonian e-residency: Redefining the nation-state in the digital era. *University of Oxford Cyber Studies Programme working paper*, v. 3, 2015.

LASALA CALLEJA, P. et al. *Electronic Government*. [S.l.]: Prensas de la Universidad de Zaragoza, 2014.

LATHROP, D.; RUMA, L. *Open government: Collaboration, transparency, and participation in practice*. [S.l.]: "O'Reilly Media, Inc.", 2010.

LEON, D. de N. Cúpula extraordinária de chefes de estado e de governo das américas http://www. sice. oas. org/ftaa/nleon. *Nleon_p. asp*, 2004.

LINDERS, D.; WILSON, S. C. What is open government?: one year after the directive. In: ACM. *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*. [S.l.], 2011. p. 262–271.

MAGNANI, M. C. B.; PINHEIRO, M. M. K. "regime" e "informação": a aproximação de dois conceitos e suas aplicações na ciência da informação|"regime"and"information": the dialogue between two concepts and their application in the information science. *Liinc em revista*, v. 7, n. 2, 2011.

MANYIKA, J. et al. Open data: Unlocking innovation and performance with liquid information. *McKinsey Global Institute*, v. 21, 2013.

MATHEUS, R.; JANSSEN, M.; MAHESHWARI, D. Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*, Elsevier, 2018.

MCDONNELL, R. M. et al. congressbr: An r package for analysing data from brazil's chamber of deputies and federal senate. *SocArXiv*, p. 1–11, 2017. Disponível em: <https://osf.io/n5jd8>.

MCGEE, R.; GAVENTA, J. Shifting power? assessing the impact of transparency and accountability initiatives. *IDS Working Papers*, Wiley Online Library, v. 2011, n. 383, p. 1–39, 2011.

MENDEL, T. *The public's right to know: principles on freedom of information legislation*. [S.l.]: Article 19, 1999.

MENDEL, T. Freedom of information as an internationally protected human right. *Comparative Media Law Journal*, Temple University Press, v. 1, n. 1, p. 39–70, 2003.

MENDEL, T.; UNESCO, N. *Freedom of information: a comparative legal survey*. [S.l.]: Unesco Paris, 2008. v. 149.

MICHENER, G.; MONCAU, L. F.; VELASCO, R. B. *Estado brasileiro e transparência avaliando a aplicação da Lei de Acesso à Informação*. [S.l.], 2015.

MOSER, C. How open is' open as possible''? three different approaches to transparency and openness in regulating access to eu documents. AUT, 2001.

NEVES JÚNIOR, O. Sobre uma arquitetura da informação do governo brasileiro: Aigov-br. 2013.

OD, O. D. *The Open Definition - Defining Open in Open Data, Open Content and Open Knowledge*. 2019. (Accessoem: 07/03/2019). Disponível em: <http://opendefinition.org/>.

OEA, C. I. D. D. H. Convenção americana sobre direitos humanos. In: *Assinada na Conferência especializada interamericana sobre direitos humanos, San José, Costa Rica, em*. [S.l.: s.n.], 1969. v. 22.

OEA, C. I. D. D. H. Declaração de princípios sobre liberdade de expressão. *Aprovado pela Comissão Interamericana de Direitos Humanos em seu 108º período ordinário de sessões, celebrado de*, v. 16, 2000.

OECD. *OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information*. 2008. (Accesso em: 07/03/2019). Disponível em: <http://www.oecd.org/internet/ieconomy/40826024.pdf>.

OGDWG, O. G. D. W. G. *The Annotated 8 Principles of Open Government Data*. 2007.

OGP, O. G. P. *Open Government Partnership | Committed to making governments more open, accountable, and responsive to citizens*. 2018. (Accessed on 07/03/2019). Disponível em: <https://www.opengovpartnership.org/>.

OKF, O. K. F. *The Open Data Handbook*. 2010. (Accesso em: 2019). Disponível em: <http://opendatahandbook.org/guide/en/>.

OKF, O. K. F. *Defining Open Data – Open Knowledge Foundation Blog*. 2013. (Accesso em: 2019). Disponível em: <https://blog.okfn.org/2013/10/03/defining-open-data/>.

OKF, O. K. F. *Open Knowledge Foundation*. 2018. (Accesso em: 2019). Disponível em: <https://okfn.org/>.

PERKEL, J. M. Programming: pick up python. *Nature News*, v. 518, n. 7537, p. 125, 2015.

PINHO, J. A. G. d. Investigando portais de governo eletrônico de estados no brasil: muita tecnologia, pouca democracia. SciELO Brasil, 2008.

POMAR, C. et al. O governo eletrônico respondendo às propensões da presença da administração pública no ciberespaço. *II CIBERÉTICA, Simpósio Internacional de Propriedade Intelectual, Informação e Ética; VIII ENIDJ, Encontro Nacional de Informação e Documentação Jurídica, Florianópolis*, 2003.

POWELL, A. B. Open culture and innovation: integrating knowledge across boundaries. *Media, Culture & Society*, Sage Publications Sage UK: London, England, v. 37, n. 3, p. 376–393, 2015.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. *Big Data*, v. 1, n. 1, p. 51–59, 2013. PMID: 27447038. Disponível em: <https://doi.org/10.1089/big.2013.1508>.

PSF, P. S. F. *About the Python Software Foundation | Python Software Foundation*. 2019. (Accesso em: 07/04/2019). Disponível em: <https://www.python.org/psf/about/#how-do-i-reach-the-psf>.

R Core Team et al. R: A language and environment for statistical computing. Vienna, Austria, 2013.

SANDOVAL-ALMAZAN, R.; GIL-GARCIA, J. R. Toward an integrative assessment of open government: Proposing conceptual lenses and practical components. *Journal of Organizational Computing and Electronic Commerce*, Taylor & Francis, v. 26, n. 1-2, p. 170–192, 2016.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. *Informação & Informação, Londrina*, v. 21, n. 2, p. 116–142, 2016.

SCHULTZ, K. A. Do democratic institutions constrain or inform? contrasting two institutional perspectives on democracy and war. *International Organization*, Cambridge University Press, v. 53, n. 2, p. 233–266, 1999.

STANTON, J. Data science: What's in it for the new librarian. *Information Space*, v. 9, n. 20, 2012.

TKACZ, N. From open source to open government: A critique of open politics. *Ephemera: Theory & politics in organization*, v. 12, n. 4, 2012.

UN, U. N. *Guidelines on open government data for citizen engagement*. [S.l.]: New York: United Nations, 2013.

W3C, W. W. W. C. Manual dos dados abertos: desenvolvedores. *Cooperação técnica científica entre Laboratório Brasileiro de Cultura Digital e o Núcleo de Informação e Coordenação do Ponto BR (NIC. br). São Paulo: Comitê Gestor da Internet no Brasil*, 2011.

WANG, L. Twinning data science with information science in schools of library and information science. *Journal of Documentation*, Emerald Publishing Limited, v. 74, n. 6, p. 1243–1257, 2018.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: CITESEER. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. [S.l.], 2000. p. 29–39.

ZHU, Y.; XIONG, Y. Towards data science. *Data Science Journal*, Ubiquity Press, v. 14, 2015.

ZUCCOLOTTO, R.; TEIXEIRA, M. A. C.; RICCIO, E. L. Transparência: reposicionando o debate. *Revista Contemporânea de Contabilidade*, Universidade Federal de Santa Catarina, v. 12, n. 25, p. 137–158, 2015.

ZWEERS, K.; PLANQUE, K. Electronic government in the us. from an organization-based perspective towards a client oriented approach. *LAW AND ELECTRONIC COMMERCE*, Kluwer Law International, v. 12, p. 91–120, 2001.

# Apêndices

# APÊNDICE A – Licenças Dados Abertos

| Licença | Domínio | By | SA |
|---|---|---|---|
| Creative Commons CCZero (CC0) | Conteúdo e Dados | Não | Não |
| Open Data Commons Public Domain Dedication and Licence (PDDL) | Dados | Não | Não |
| Creative Commons Attribution 4.0 (CC-BY-4.0) | Conteúdo e Dados | Sim | Não |
| Open Data Commons Attribution License (ODC-BY) | Dados | Sim | Não |
| Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0) | Conteúdo e Dados | Sim | Sim |
| Open Data Commons Open Database License (ODbL) | Dados | Sim | Sim |

# APÊNDICE B – Consumo das APIs do Senado e Congresso Nacional

## Consumo das APIs do Senado e Congresso Nacional

*Paulo Cardoso*

*17/05/2019*

### Introdução

Este documento tem por objetivo apresentar algumas possibilidades do consumo das APIs do Senado e Congresso Nacional por meio do uso da biblioteca 'congressbr". Essa biblioteca é uma abstração das chamadas, encapsulando a chamada do serviços em funções parametrizáveis. As bibliotecas utilizadas nessa analise foram"congressbr","ggplot2","dplyr" e "plyr", as mesmas são instanciadas abaixo:

```
require(congressbr)
```

```
## Loading required package: congressbr
```

```
require(plyr)
```

```
## Loading required package: plyr
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

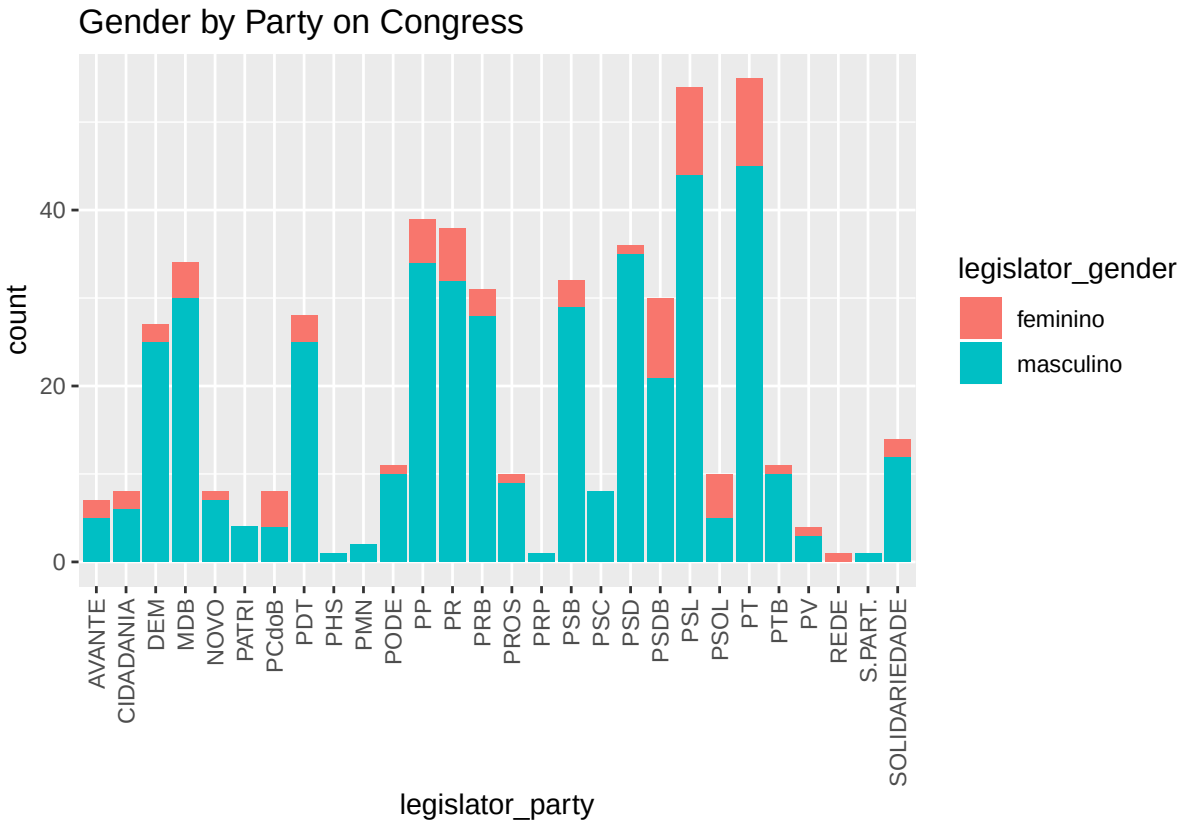
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

### Exploração de dados

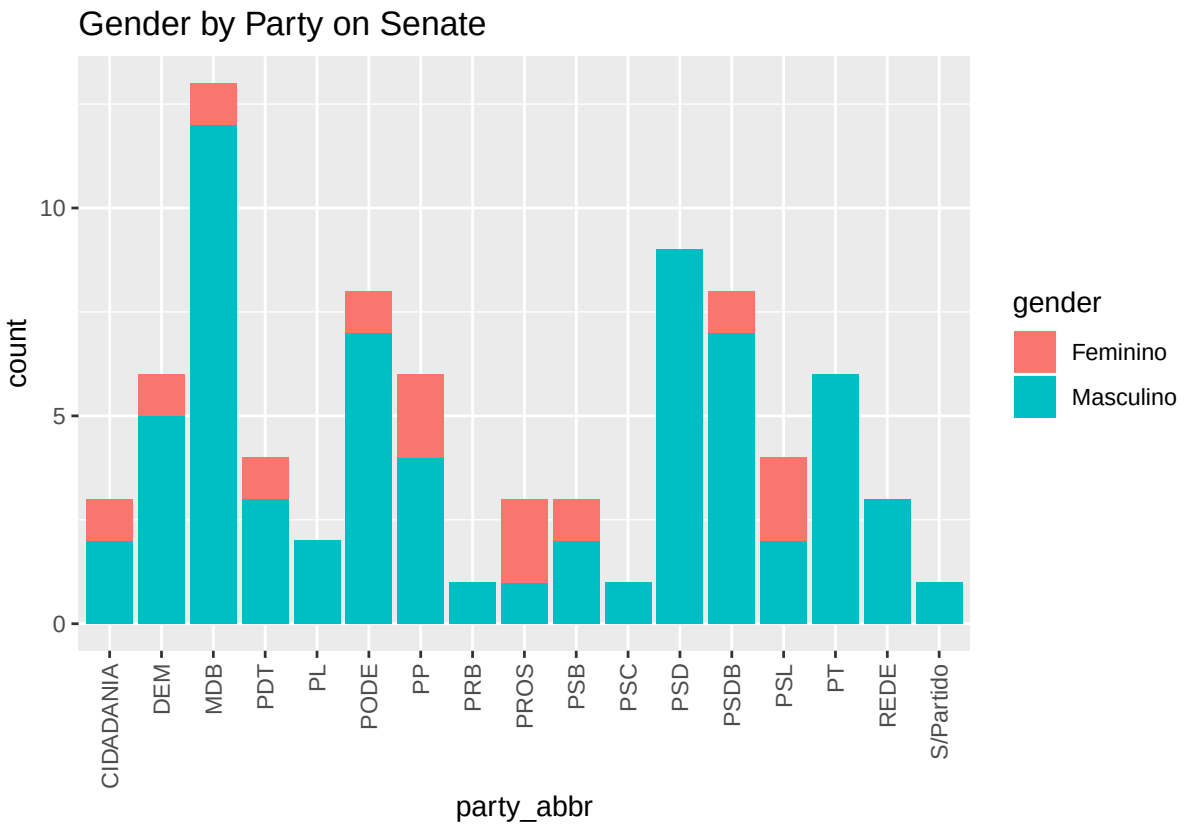
Esta sessão demonstrará algumas estatiscias basicas dos dados do Senado e Congresso Nacional, como proporção de homens e mulheres, proporção de partidos, entre outras estatisticas basicas.

#### Representatividade de Genero entre Senadores e Deputados Federais

```
t <- ggplot(conList, aes(x = legislator_party, fill = legislator_gender)) + geom_bar() + theme(axis.t
print(t + ggtitle("Gender by Party on Congress"))
```




```
t <- ggplot(senList, aes(x = party_abbr, fill = gender)) + geom_bar()  +  theme(axis.text.x = element
print(t + ggtitle("Gender by Party on Senate"))
```




### Coalisões no Senado

Dentre os dados possiveis de serem extraidos utilizando as APIs do Sentado está as coalisões que compõe as bancadas do Senado. Com isso é possível extrair a lista das coalisões ativas.

```
sen_coalitions(ascii = TRUE)
```

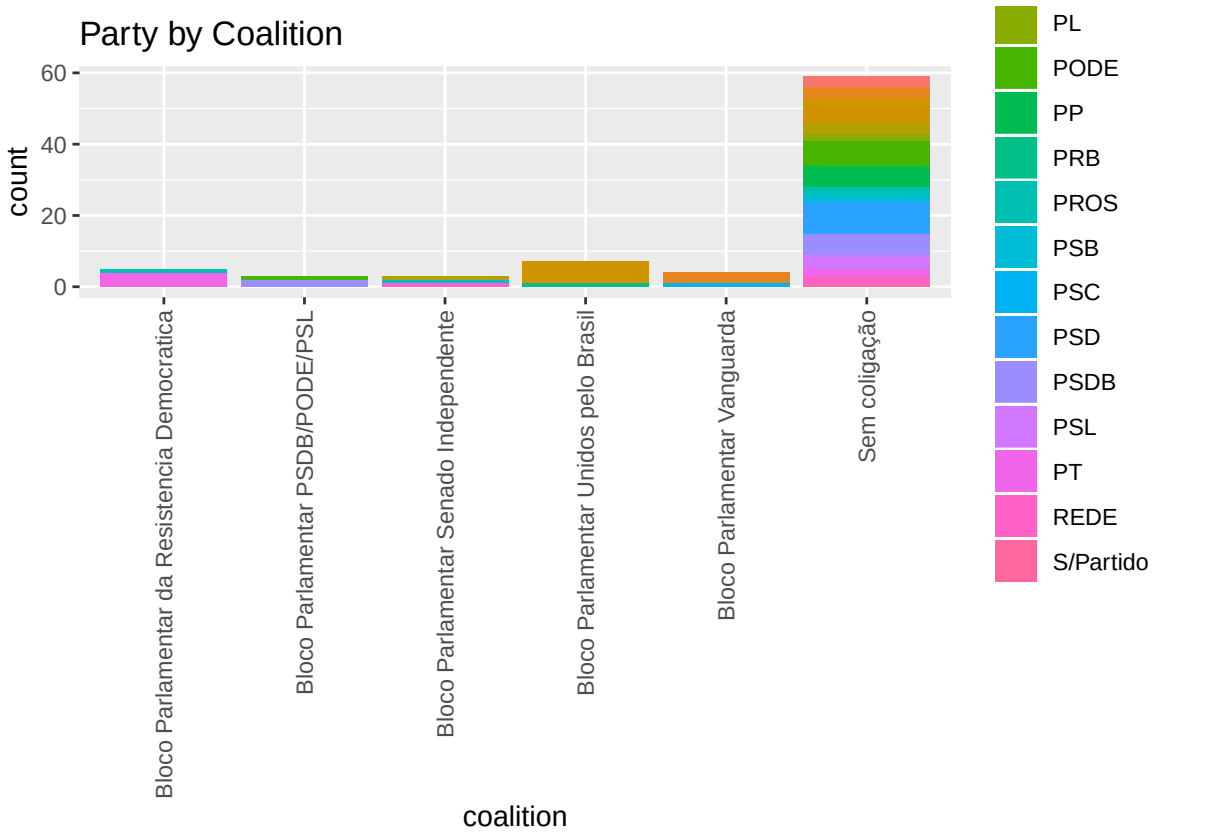
```
## # A tibble: 7 x 4
##   bloc_code bloc_name                  bloc_label       date_created
##   <chr>     <chr>                      <chr>            <dttm>
## 1 278       Bloco Parlamentar PSDB/PO~ Bloco PSDB/PODE~ 2019-02-12 00:00:00
## 2 272       Bloco Parlamentar Senado ~ Bloco REDE/PDT/~ 2019-02-06 00:00:00
## 3 277       Bloco Parlamentar Unidos ~ Bloco MDB/PRB    2019-02-11 00:00:00
## 4 274       Bloco Parlamentar Vanguar~ Bloco DEM/PR/PSC 2019-02-06 00:00:00
## 5 273       Bloco Parlamentar da Resi~ Bloco PT/PROS    2019-02-06 00:00:00
## 6 284       Maioria                    Maioria          2019-02-19 00:00:00
## 7 281       Minoria                    Minoria          2019-02-13 00:00:00
```

Então baseando-se nessa lista de coalisões, foi criado uma nova variavel no conjunto de dados para identificar a coalisão do parlamentar.

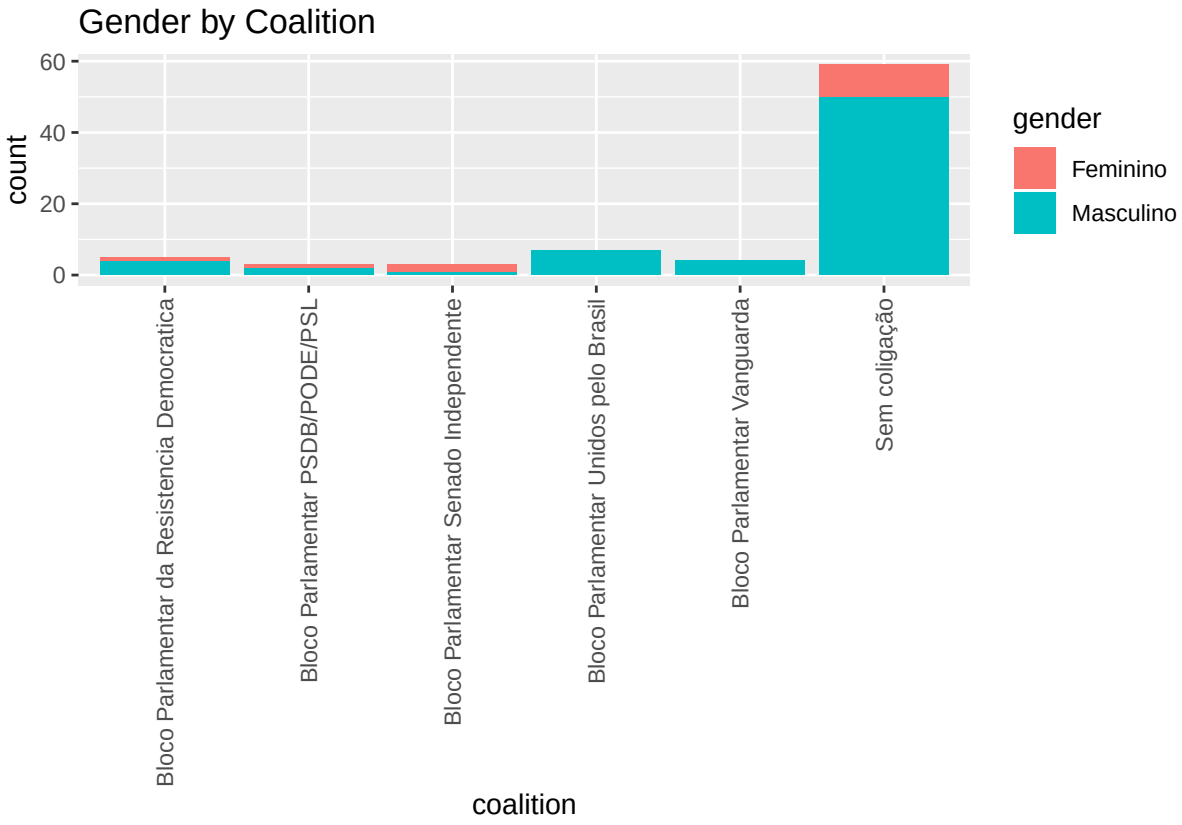
```
senList$coalition[senList$party_abbr == c("PSDB","PODE","PLS")] <- "Bloco Parlamentar PSDB/PODE/PSL"
senList$coalition[senList$party_abbr == c("REDE","PDT","PPS", "PSB")] <- "Bloco Parlamentar Senado Ir
senList$coalition[senList$party_abbr == c("MDB","PRB")] <- "Bloco Parlamentar Unidos pelo Brasil"
senList$coalition[senList$party_abbr == c("DEM","PR","PSC")] <- "Bloco Parlamentar Vanguarda"
senList$coalition[senList$party_abbr == c("PT","PROS")] <- "Bloco Parlamentar da Resistencia Democrat
senList$coalition[is.na(senList$coalition)] <- "Sem coligação"
```

Utilizando-se dessa nova variavel no conjunto de dados, abaixo é apresentado a distribuição de partidos e genero pela coalizão.

```
t <- ggplot(senList, aes(x = coalition, fill = party_abbr)) + geom_bar()  +  theme(axis.text.x = elem
print(t + ggtitle("Party by Coalition"))
```




```
t <- ggplot(senList, aes(x = coalition, fill = gender)) + geom_bar()  +  theme(axis.text.x = element_
print(t + ggtitle("Gender by Coalition"))
```




## Comissões no Senado

Existem tres tipos de comissões, comissões do senado, do congresso nacional e mistas. Nesta analise foram consideradas apenas comissões do senado, alem disso, outro ponto de destaque é que nem todas as comissoes tem dados disponiveis na API. Sendo assim a analise está incompleta por impossibilidade de acesso a parte dos dados.

Abaixo podemos observar a coleta de dados das comissões, que contem a lista e todas as comissões e o filtro que redus do conjunto de dados para comissões apenas do Senado.

```
com <- sen_commissions(active = "Yes", ascii = TRUE)
com <- com[com$commission_house == "SF",]
t <- NULL
t1 <- NULL
```

Apos a coleta da lista de comissões, pode-se recuperar a lista de senadores membros de cada comissão como pode ser observado abaixo:

```
for(i in 1:nrow(com)){
    if(is.null(t) ) {
        res <- try(t <- sen_commissions_senators(code = com$commission_abbr[i], ascii = TRUE))
        if(inherits(res, "GET request failed")) next
    } else if(nrow(t) == 0){
        res <- try(t <- sen_commissions_senators(code = com$commission_abbr[i], ascii = TRUE))
        if(inherits(res, "GET request failed")) next
        #t <- rbind(t,t1)
    } else {
        res <- try(t1 <- sen_commissions_senators(code = com$commission_abbr[i], ascii = TRUE))
```

```
        if(inherits(res, "GET request failed")) next
        t <- rbind(t,t1)
    }
}
```

```
## Error in status(req) :
##   GET request failed. Please check the validity of the information you requested.
```

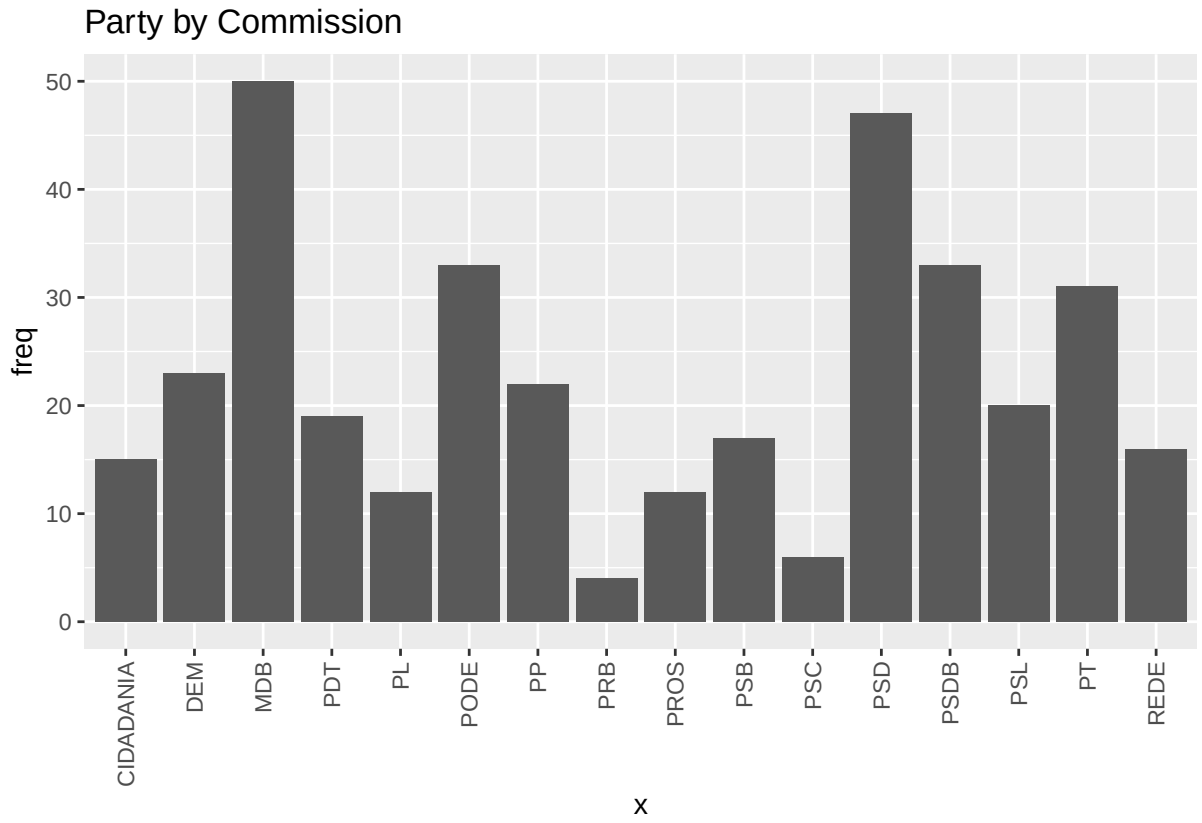
```
com1 <- t
head(com1)
```

```
## # A tibble: 6 x 6
##   commission commission_abbr senator_id senator_name senator_party
##   <chr>      <chr>           <chr>      <chr>        <chr>
## 1 Comissao ~ CRA             374        Paulo Rocha  PT
## 2 Comissao ~ CRA             5988       Soraya Thro~ PSL
## 3 Comissao ~ CRA             1173       Wellington ~ PL
## 4 Comissao ~ CRA             5748       Veneziano V~ PSB
## 5 Comissao ~ CRA             1249       Katia Abreu  PDT
## 6 Comissao ~ CRA             742        Marcelo Cas~ MDB
## # ... with 1 more variable: senator_state <chr>
```

Com a lista de comissões e membros extraidas anteriormente é possivel apresentar indicadores como a distribuição de membros participantes das comissões por partido, comissão e os 10 senadores que mais participam de comissões.

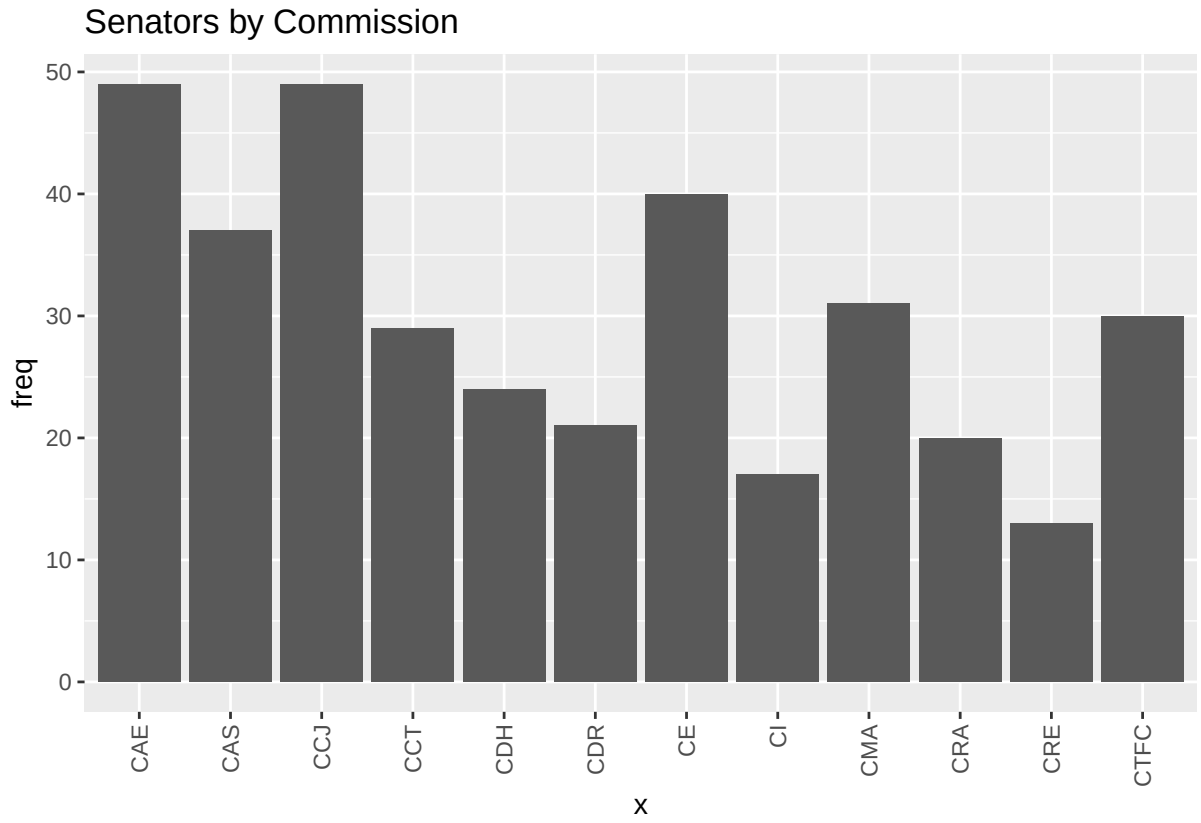
Este primeiro grafico apresenta a distribuição de partidos nas comissoes.

```
a <- plyr::count(com1$senator_party)
a$x <- as.character(a$x)
t <- ggplot(a, aes(x = x, y = freq)) + geom_bar(stat = "identity")  +  theme(axis.text.x = element_te
print(t + ggtitle("Party by Commission"))
```



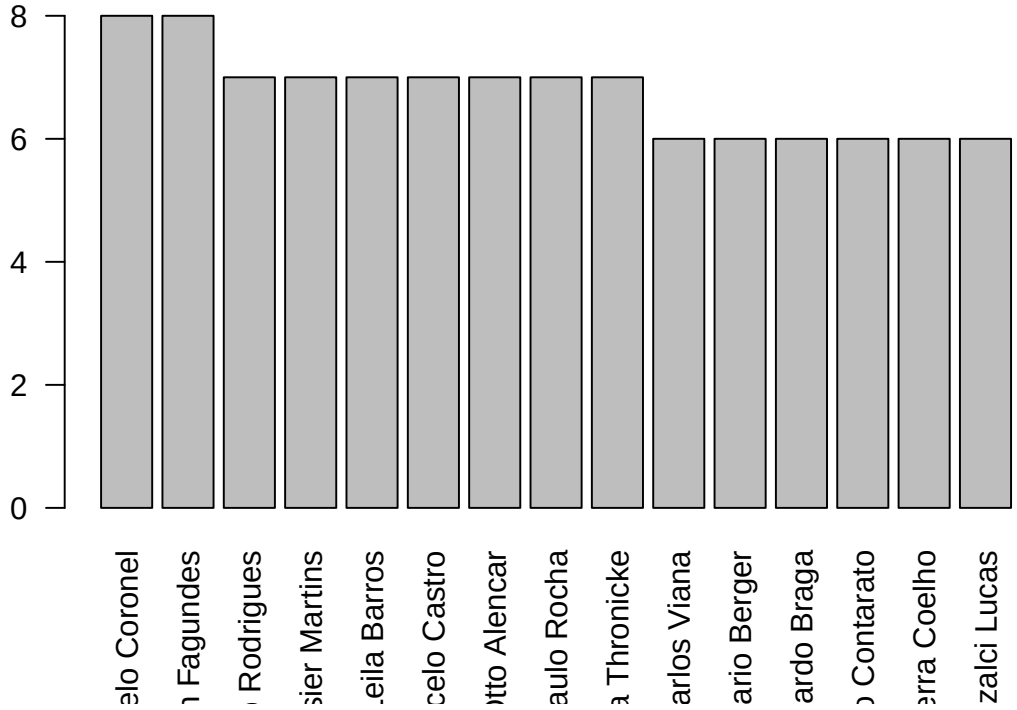

Ja este segundo grafico apresenta a distribuição de parlamentares nas comissoes.

```
a <- plyr::count(com1$commission_abbr)
a$x <- as.character(a$x)
p <- ggplot(data=a, aes(x=x, y=freq)) + geom_bar(stat="identity") +  theme(axis.text.x = element_text
print(p + ggtitle("Senators by Commission"))
```



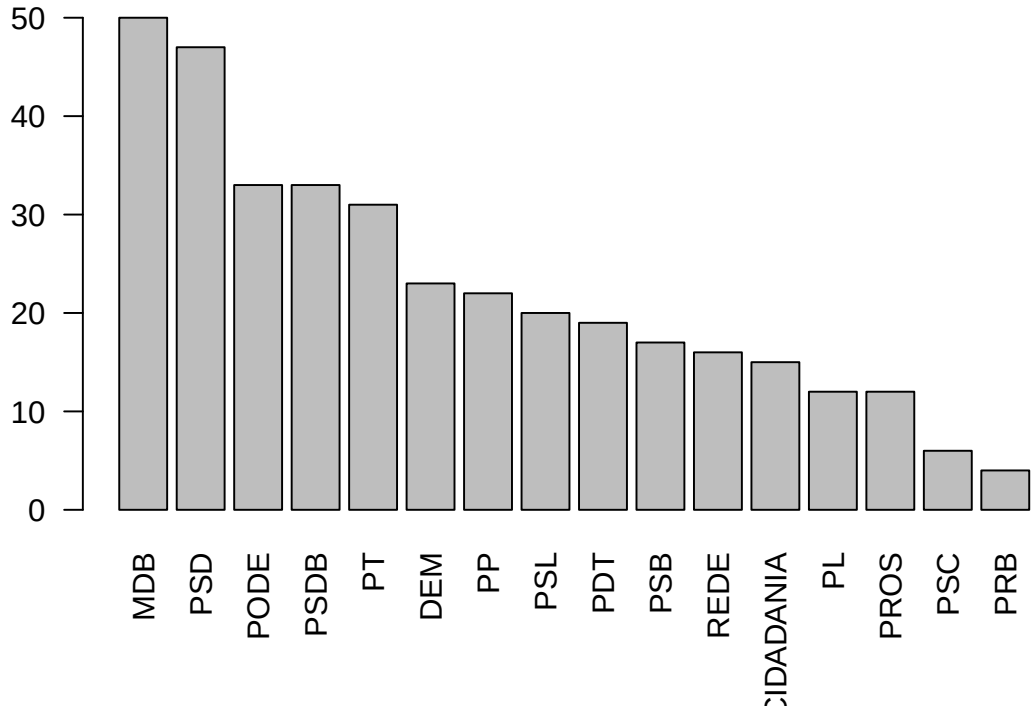

O Top 15 Senadores em mais comissões pode ser observado abaixo.

```
t <- table(com1$senator_name)
t <- t[order(-t)]
barplot(head(t,15),las = 2)
```




Para finalizar, é apresentada a distribuição dos partidos por comissão.

```
t <- table(com1$senator_party)
t <- t[order(-t)]
barplot(t,las = 2)
```



# APÊNDICE C – DOU - Natural Language Processing

## DOU - Natural Language Processing

*Paulo Cardoso*

*20/05/2019*

### Introdução

O objetivo deste documento é apresentar uma analise textual, tambem conhecida como processamento de linguagem natural. Como objeto da analise estão os dados abertos do diario oficial da união, obtidos no portal www.dados.gov.br.

Esta analise foi desenvolvida utilizando duas linguagens de programação, sendo a primeira Python que foi utilizado para a extração do dos dados dos arquivos XML e conversão para o formato tabular e a otra linguagem de programação utilizada foi R, que foi utilizado para realizar a mineração de texto e construção do wordcloud.

Apos realizar o download dos dados, e colocar os arquivos .zip no mesmo diretorio dos scripts a primeira parte da analise pode ser excecutada usando python. Para possibilitar a utilização dalinguagem Python nesse documento R Markdown foi utilizada a biblioteca reticulate.

```
require(reticulate)
```

```
## Loading required package: reticulate
```

### 1 - Descompactação, extração e transformação dos dados

Como primeiro passo, as bibliotecas Python são importadas:

```
# Loading Libraries
import os, zipfile, inspect, glob, re
import xml.etree.cElementTree as et
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Apos importar as bibliotecas a descompactação pode ser realizada, como pode ser observado abaixo:

```
# getting file name and directory
filename = inspect.getframeinfo(inspect.currentframe()).filename
dir_name = os.path.dirname(os.path.abspath(filename))

# Unzip
for item in os.listdir(dir_name):
    if item.endswith(".zip"):
        zip_ref = zipfile.ZipFile(item)
        zip_ref.extractall()
        print(item + " descompactado.")
```

```
## S03012019.zip descompactado.
## S01012019.zip descompactado.
```

```
## 0  7311944  ...   ATO Nº      DE    DE JANEIRO DE       Expede ...
## 1  6913734  ...    ANEXO II   a  QUADRO DEMONSTRATIVO DOS CARGO...
## 2  7159090  ...   PORTARIA Nº     DE    DE JANEIRO DE       Div...
## 3  6983174  ...     PORTARIA Nº    DE   DE JANEIRO DE       Dis...
## 4  7062206  ...                                      Unimed Norte...
##
## [5 rows x 23 columns]
```

```
# describing dataset
df.describe()
```

```
##              id  ...                                                Texto
## count     63504  ...                                                63504
## unique    63504  ...                                                63113
## top     7253784  ...   RETIFICAÇÃO  Subrogada pela UASG           UN...
## freq          1  ...                                                   10
##
## [4 rows x 23 columns]
```

Como encerramento da etapa de descompactação, extração e transformação de dados ocorre a persistencia do dataframe criado em disco para futura utilização.

```
# writing data on disc
df.to_csv('DOU1901.csv', sep = ',')
```

## 2 - Processamento de Linguagem Natural

Esta fase da analise apresenta tecnicas de mineração de texto, limpeza de dados textuais, tokenização e apresentação de wordclouds. Alem disso, foi desenvolvida utilizando a linguagem R com o suporte de bibliotecas relacionadas a text mining e datas.

Como primeiro passo, serão carregadas as bibliotecas utilizadas.

```
# loading libraries
library(tm)
```

```
## Loading required package: NLP
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(readr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

Os dados criados na primeira etapa e salvos em disco são carregados.

```
# data load
df <- read_csv("DOU1901.csv", col_types = cols(X1 = col_skip()))
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Warning: 1 parsing failure.
##  row       col               expected actual          file
## 2533 idMateria no trailing characters     -2 'DOU1901.csv'
```

### 2.1 - Limpeza e tratamento de Dados

Para que os dados do DOU estejam prontos para a analise é nessario que passem por diversas limpezas e transformações. Dentre os processos realizados estão: correção de formato de datas, criação de variaveis, agrupamento de variavais, remoção de acentuação e conversões de formato.

```
# substituindo / por - em pubdate, convertendo para data e criando campo de numero da semana paras eg
df$pubDate <- gsub('/', '-', df$pubDate)
df$pubDate <- dmy(df$pubDate)
df$week <- week(as.Date.character(df$pubDate))

# agrupando sessões
df$sessao <- df$pubName
df$sessao[df$sessao %in% c('DO1A','DO1','DO1E')] <- 1
df$sessao[df$sessao %in% c('DO2','DO2E')] <- 2
df$sessao[df$sessao %in%c('DO3','DO3E')] <- 3

# removendo acentos e caracteres especiais por meio de conversão utf-8 para ascii
df$Texto <- iconv(df$Texto,from="UTF-8",to="ASCII//TRANSLIT")

# removendo diversos espaços
df$Texto <- gsub("\\s+", " ", df$Texto)
```

### 2.2 - Wordcloud

Serão apresentados tres exemplos de wordclouds, sendo o primeiro contendo dados de todas as sessão do dia 31-01-2019 plotado inteiramente em preto já o segundo grafico é colorido.

Sendo que para apresentar o wordcloud são necessarias mais algumas limpezas e transformações de dados.

#### Mineração de Texto e Limpeza de dados

```
#df1 <- df[df$sessao == 3,] # filtro por sessão
#df1 <- df[df1$week == 3,] # filtro por semana
df1 <- df[df$pubDate == "2019-01-31",] # filtro apenas dia 2019-01-31
# creating corpus
corpus <- Corpus(VectorSource(df1$Texto))
```

```
# Convert all text to lower case
corpus <- tm_map(corpus, content_transformer(tolower))
```

```
## Warning in tm_map.SimpleCorpus(corpus, content_transformer(tolower)):
## transformation drops documents
```

```
# remover toda a pontuação
corpus <- tm_map(corpus, removePunctuation)
```

```
## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation
## drops documents
```

```
# Remove numbers
corpus <- tm_map(corpus, removeNumbers)
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeNumbers): transformation drops
## documents
```

```
# remover Stopwords
corpus <- tm_map(corpus, removeWords, stopwords('pt'))
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeWords, stopwords("pt")):
## transformation drops documents
```

```
corpus <- tm_map(corpus, removeWords, c('serao', 'meses', 'hora', 'caixa', 'cep', 'mail', 'WWW', 'usc
```

```
## Warning in tm_map.SimpleCorpus(corpus, removeWords, c("serao", "meses", :
## transformation drops documents
```

```
# generalizando termos para raiz
#corpus <- tm_map(corpus, stemDocument)
```

```
# Replacing "/", "@" and "|" with space
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
corpus <- tm_map(corpus, toSpace, "/")
```

```
## Warning in tm_map.SimpleCorpus(corpus, toSpace, "/"): transformation drops
## documents
```

```
corpus <- tm_map(corpus, toSpace, "@")
```

```
## Warning in tm_map.SimpleCorpus(corpus, toSpace, "@"): transformation drops
## documents
```

```
corpus <- tm_map(corpus, toSpace, "\\|")
```

```
## Warning in tm_map.SimpleCorpus(corpus, toSpace, "\\|"): transformation
## drops documents
```

```
td_mtx <- TermDocumentMatrix(corpus, control = list(minWordLength = 6))
```

```
v <- sort(rowSums(as.matrix(td_mtx)), decreasing=TRUE) #ordena as palavras
```

```
fdf <- data.frame(word=names(v), freq=v[]) #organiza um novo banco
```

### Wordcloud Monocolor (Preto)

```
wordcloud(fdf$word, fdf$freq, min.freq=333)
```


**Wordcloud Colorido**

```
pal1 <- brewer.pal(12,"Paired")
pal3 <- brewer.pal(12,"Set3")

wordcloud(fdf$word, fdf$freq, min.freq=333, random.order=FALSE, rot.per=.15,colors=pal1)
```


processo
valor
objeto
contrato
licitacao
termo
pregao
servicos
portaria
inciso
edital
federal
nacional
municipal
fundamento
ministerio
vigencia
extrato
cargo
artigo
assinatura
secretaria
publicada
fazenda
aquisicao
comissao
publicacao
contratacao
conforme
recorrente
estado
oliveira
global
resolve
aditivo
publico
geral
aviso
conceder
partir
unico
rio
nivel
sicon
centro
pessoal
padrao
legal
gov
matricula
saude
siape
servico
gestao
atribuicoes
prefeitura
presencial
termos
itens
fonte
registro
tendo
instituto
maria
decreto
desta
reais
interno
demais
entra
contratante
social
presidente
eletronico
jose
deste
sera
quadro
recorrida
ato
uasg
precos
classe
oficial
ata
mil
dou
tcu
chefe
trabalho
relator
contratada
emenda
constitucional
eireli
municipio
uniao
empresa
item
brasil
secao
vista
site
prazo
total
codigo
acordao
contratado
substituto
publica
sede
unico
diario
paulo
resultado
aposentadoria
ocupante
recursos
autos