ANDRÉ FILIPE DE MORAES BATISTA

MODELO DE AVALIAÇÃO DE CONJUNTOS DE DADOS CIENTÍFICOS POR MEIO DA DIMENSÃO DE VERACIDADE DOS DADOS

São Paulo
2018

ANDRÉ FILIPE DE MORAES BATISTA

MODELO DE AVALIAÇÃO DE CONJUNTOS DE DADOS CIENTÍFICOS POR MEIO DA DIMENSÃO DE VERACIDADE DOS DADOS

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Ciências.

São Paulo
2018

ANDRÉ FILIPE DE MORAES BATISTA

MODELO DE AVALIAÇÃO DE CONJUNTOS DE DADOS CIENTÍFICOS POR MEIO DA DIMENSÃO DE VERACIDADE DOS DADOS

Versão Corrigida

(Versão Original encontra-se na unidade que aloja
o Programa de Pós-graduação)

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Ciências.

Área de Concentração:

Engenharia da Computação

Orientador:

Prof. Dr. Pedro Luiz Pizzigatti Corrêa

São Paulo
2018

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

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Batista, André Filipe de Moraes
Modelo de Avaliação de Conjuntos de Dados Científicos por meio da Dimensão de Veracidade dos Dados / A. F. M. Batista -- versão corr. -- São Paulo, 2018.
167 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Ciência dos Dados 2.Big Data 3.Veracidade 4.Perfil de Aplicação I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

BATISTA, André Filipe de Moraes. **Modelo de Avaliação de Conjuntos de Dados Científicos por meio da Dimensão de Veracidade dos Dados.** 2018. 167 f. Tese (Doutorado em Engenharia Elétrica) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2018.

Aprovado em 06 de novembro de 2018.

Banca Examinadora

Prof. Dr. Pedro Luiz Pizzigatti Corrêa
Escola Politécnica da Universidade de São Paulo - POLI/USP
Julgamento: Aprovado

Prof. Dr. Jorge Rady de Almeida Júnior
Escola Politécnica da Universidade de São Paulo - POLI/USP
Julgamento: Aprovado

Profa. Dra. Luciana Varanda Rizzo
Universidade Federal de São Paulo - UNIFESP
Julgamento: Aprovado

Profa. Dra. Rachel Ifanger Albrecht
Instituto de Astronomia, Geofísica e Ciências Atmosféricas - IAG/USP
Julgamento: Aprovado

Prof. Dr. Guilherme Ataíde Dias
Universidade Federal da Paraíba - UFPB
Julgamento: Aprovado

AGRADECIMENTOS

Primeiramente, agradeço a Deus, pois “Bendito é o homem cuja confiança está no Senhor. Ele será como uma árvore plantada junto às águas e que estende as suas raízes para o ribeiro. Ela não temerá quando chegar o calor, porque as suas folhas estão sempre verdes; não ficará ansiosa no ano da seca nem deixará de dar fruto.” (Jeremias 17:7-8).

Agradeço especialmente, também, a todos aqueles que contribuíram diretamente e indiretamente para a conclusão deste trabalho, incluindo:

Minha família, a rocha sólida que possibilitou a minha formação. Agradeço aos meus pais pelo esforço em criar seus filhos com o foco na educação, permitindo-me dedicar aos estudos, fornecendo apoio e incentivo nos momentos mais difíceis e árduos.

Minha esposa, companheira e parceira, Denise, que passou a compartilhar sua vida comigo, e que me apoia a todo momento, seja qual for a jornada que eu trilhe. Ao nosso filho amado, Gabriel, que é o tesouro mais precioso que Deus nos deu, que reluz cada dia mais e nos proporciona momentos incomparáveis.

Professores Giri Prakash (*Oak Ridge National Laboratory - ORNL*), Mike Frame (*United States Geological Survey - USGS*) e Suzie Allard (*The University of Tennessee - UTK*) pelas valiosas colaborações de pesquisa, revisão da tese e dos artigos, além da disponibilidade ao longo do desenvolvimento dos trabalhos.

Membros do Grupo de Estudos, Pesquisa e Extensão em Big Data da POLI-USP, pelo auxílio na condução dos eventos, colaborações de pesquisa, e discussões técnicas realizadas para a condução dos pontos abordados no estudo de caso.

Meu orientador, Prof. Dr. Pedro Luiz Pizzigatti Corrêa, pelo apoio, aconselhamento e suporte necessários para o bom andamento da pesquisa, o qual contribuiu efetivamente no meu crescimento acadêmico, profissional e pessoal.

"O conhecimento é o processo de acumular dados;
a sabedoria reside na sua simplificação."

-- Martin H. Fischer

RESUMO

BATISTA, André Filipe de Moraes. **Modelo de Avaliação de Conjuntos de Dados Científicos por meio da Dimensão de Veracidade dos Dados.** 2018. 167 f. Tese (Doutorado em Engenharia Elétrica) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2018.

A ciência é uma organização social: grupos de colaboração independentes trabalham para gerar conhecimento como um bem público. A credibilidade dos trabalhos científicos está enraizada nas evidências que os suportam, as quais incluem a metodologia aplicada, os dados adquiridos e os processos para execução dos experimentos, da análise de dados e da interpretação dos resultados obtidos. O dilúvio de dados sob o qual a atual ciência está inserida revoluciona a forma como as pesquisas são realizadas, resultando em um novo paradigma de ciência baseada em dados. Sob tal paradigma, novas atividades são inseridas no método científico de modo a organizar o processo de geração, curadoria e publicação de dados, beneficiando a comunidade científica com o reuso de conjuntos de dados científicos e a reprodutibilidade de experimentos. Nesse contexto, novas abordagens para a resolução de problemas estão sendo apresentadas, obtendo resultados que antes eram considerados de relevante dificuldade, bem como possibilitando a geração de novos conhecimentos. Diversos portais estão disponibilizando conjuntos de dados resultantes de pesquisas científicas. Todavia, tais portais pouco abordam o contexto sobre os quais os conjuntos de dados foram criados, dificultando a compreensão sobre os dados e abrindo espaço para o uso indevido ou uma interpretação errônea. Poucas são as literaturas que abordam essa problemática, deixando o foco para outros temas que lidam com o volume, a variedade e a velocidade dos dados. Essa pesquisa objetivou definir um modelo de avaliação de conjuntos de dados científicos, por meio da construção de um perfil de aplicação, o qual padroniza a descrição de conjuntos de dados científicos. Essa padronização da descrição é baseada no conceito de dimensão de Veracidade dos dados, definido ao longo da pesquisa, e permite o desenvolvimento de métricas que formam o índice de veracidade de conjuntos de dados científicos. Tal índice busca refletir o nível de detalhamento de um conjunto de dados, com base no uso dos elementos de descrição, que facilitarão o reuso dos dados e a reprodutibilidade dos experimentos científicos. O índice possui duas dimensões: a dimensão intrínseca aos dados, a qual pode ser utilizada como critério de admissão de conjunto de dados em portais de publicação de dados; e a dimensão social, mensurando a adequabilidade de um conjunto de dados para uso em uma área de pesquisa ou de aplicação, por meio da avaliação da comunidade científica. Para o modelo de avaliação proposto, um estudo de caso foi desenvolvido, descrevendo um conjunto de dados proveniente de um projeto científico internacional, o projeto GoAmazon, de modo a validar o modelo proposto entre os pares, demonstrando o potencial da solução no apoio ao reuso dos dados, podendo ser incorporado em portais de dados científicos.

Palavras-Chave – Ciência dos Dados. Veracidade. Perfil de Aplicação. Avaliação Colaborativa.

ABSTRACT

BATISTA, André Filipe de Moraes. **Scientific Datasets Evaluation Model based on The Data Veracity Dimension.** 2018. 167 f. Tese (Doutorado em Engenharia Elétrica) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2018.

Science is a social organization: independent collaboration groups work to generate knowledge as a public good. The credibility of the scientific work is entrenched in the evidence that supports it, which includes the applied methodology, the acquired data, the processes to execute the experiments, the data analysis, and the interpretation of the obtained results. The flood of data under which current science is embedded revolutionizes the way surveys are conducted, resulting in a new paradigm of data-driven science. Under such a paradigm, new activities are inserted into the scientific method to organize the process of generation, curation, and publication of data, benefiting the scientific community with the reuse and reproducibility of scientific datasets. In this context, new approaches to problem solving are being presented, obtaining results that previously were considered of relevant difficulty, as well as making possible the generation of new knowledge. Several portals are providing datasets resulting from scientific research. However, such portals do little to address the context upon which datasets are created, making it difficult to understand the data and opening up space for misuse or misinterpretation. In the Big Data area, the dimension that proposes to deal with this aspect is called Veracity. Few studies in the literature approach such a theme, focusing on other dimensions, such as volume, variety, and velocity of data. This research aimed to define a model of collaborative assessment of scientific datasets, through the establishment of an application profile, which standardizes the description of scientific datasets. This standardization of the description is based on the veracity dimension concept, which is defined throughout the research and allows the development of metrics that form the Veracity Index of scientific datasets. This index seeks to reflect the level of detail of a dataset based on the use of the descriptive elements, which will facilitate the reuse and reproducibility of the data. The index is weighted by the evaluation of the scientific community in a collaborative sense, which assess the level of description, comprehension capacity, and suitability of the dataset for a given research or application area. For the proposed collaborative evaluation model, a case study was developed that described a dataset from an international scientific project, the GoAmazon project, in order to validate the proposed model among the peers, demonstrating the potential of the solution in the reuse and reproducibility of datasets, showing that such an index can be incorporated into scientific data portals.

Keywords – Data Science. Veracity. Application Profile. Collaborative Assessment.

LISTA DE FIGURAS

Figura 1 Categorização dos desafios em Ciência dos Dados. 19

Figura 2 Áreas que compõem a Ciência dos Dados. 36

Figura 3 O processo de Ciência de Dados. 37

Figura 4 Modelo de Gestão de Dados do DataONE. 39

Figura 5 Modelo de ciclo de vida dos dados de pesquisa proposto pela USGS. 42

Figura 6 Componentes de um Perfil de Aplicação. 50

Figura 7 Representação de uma afirmação em RDF, descrita por uma tripla. 50

Figura 8 Representação de uma afirmação em RDF, descrita por uma tripla. 51

Figura 9 Descrição em RDF/XML da tripla representada na Figura 8. 52

Figura 10 Consulta em SPARQL sobre o RDF/XML definido na Figura 9. . . 52

Figura 11 Visão geral da estrutura do modelo de avaliação de conjuntos de dados científicos. 63

Figura 12 Modelo de Domínio para o Perfil de Aplicação proposto. 69

Figura 13 Modelo de Estruturação de Conjuntos de Dados. 74

Figura 14 Rede de Avaliação modelada como um grafo bipartido. 85

Figura 15 Exemplo de execução do algoritmo RBPD. 88

Figura 16 RBPD*: configurações da rede de avaliação em que a reputação dos usuários (u_i,u_j) será 0,5 (50%). 89

Figura 17 Análise da reputação dos usuários em uma rede de avaliações. . . . 90

Figura 18 Gráfico Radial para a Confiança dos Consumidores do Conjuntos de Dados d. 101

Figura 19 Gráfico Radial do Índice de Veracidade de d - Dimensal Social. Os valores foram ajustados em função da reputação dos usuários e do *aging* de cada avaliação efetuada. 102

Figura 20 Proposta de visualização do Índice de Veracidade do conjunto de dados d_1. 107

Figura 21 Análise da reputação de um consumidor em função do crescimento da rede. 108

Figura 22 Proposta de processo macro de geração do índice de veracidade e sua aplicação em um portal de dados científicos. 109

Figura 23 Localização das principais estações de amostragem do projeto GoAmazon. 113

Figura 24 Sítio T3 do Projeto GoAmazon. 113

Figura 25 Aplicação Web desenvolvida para preenchimento do Perfil de Aplicação. 115

Figura 26 Grafo interativo representando o conteúdo dos elementos dos metadados que descrevem o conjuntos de dados. 116

Figura 27 Ciclos diurnos médio da concentração de O_3 nas estações chuvosa (janeiro a junho) e seca (julho a dezembro) no sítio T3. 117

Figura 28 Medição da concentração de O_3 em T3 entre 18/08/2014 e 22/08/2014. O horário se encontra na escala de tempo UTC. O horário local em Manaus é dado por UTC - 4. 118

Figura 29 Número de focos em três períodos para o Bioma Amazônia. 119

Figura 30 Planejamento e Execução do Estudo de Caso. 120

Figura 31 Planejamento e Execução do Estudo de Caso. 121

Figura 32 Área de Formação dos Respondentes. 122

Figura 33 Formação Acadêmica dos Respondentes. 123

Figura 34 Entendimento sobre o processo de Geração de Dados 124

Figura 35 Confiança em reutilizar o conjunto de dados 126

LISTA DE TABELAS

Tabela 1 Vocabulário controlado sobre os tipos de recursos proposto pela DCMI. 48

Tabela 2 Principais definições de Veracidade encontradas na literatura. . . . 56

Tabela 3 Requisitos 5W1H para a descrição de dados científicos. 68

Tabela 4 Vocabulários considerados na proposta do Perfil de Aplicação. . . . 71

Tabela 5 Distribuição dos elementos de metadados com base no modelo 5W1H. 71

Tabela 6 Perfil de Descrição do Perfil de Aplicação para suporte à Veracidade. 73

Tabela 7 Afirmações para a verificação da confiança do consumidor de dados em um conjunto de dados científico. 94

Tabela 8 Cenário A - Métricas para cada perspectiva do índice de Veracidade para um *dataset* d. 99

Tabela 9 Cenário A - Valores médios das métricas de apoio à veracidade para a perspectiva social, já considerando o decaimento temporal das avaliações e a reputação dos consumidores. 101

Tabela 10 Avaliações sobre d_1. 106

Tabela 11 Cálculo do Índice de Veracidade de Dados, em sua perspectiva social, para d_1. 106

Tabela 12 Descrição dos Dados disponíveis no Conjunto de Dados. 114

SUMÁRIO

1 Introdução **15**

1.1 Caracterização do Problema 18

1.2 Objetivos 20

1.3 Justificativa 21

1.4 Materiais e Métodos 22

1.5 Organização do Trabalho 23

2 Referencial Teórico **25**

2.1 Informática para a Biodiversidade e Meio Ambiente 25

2.2 Ciência dos Dados 27

2.2.1 Histórico 27

2.2.2 Definições 32

2.2.2.1 O Processo de Ciência dos Dados 35

2.3 Modelos de Gestão de Dados 38

2.3.1 Modelo proposto pelo Projeto DataONE 39

2.3.2 Modelo proposto pela USGS 40

2.4 Incertezas e o Processo Analítico de Dados 44

2.5 Metadados 46

2.5.1 Vocabulários 47

2.6 Perfil de Aplicação 48

2.7 *Linked Data* 49

3 Veracidade dos Dados **53**

3.1 O conceito de Veracidade dos Dados 55

3.2 Modelo de Avaliação de Conjuntos de Dados Científicos 60
3.2.1 Perfil de Aplicação 65
3.2.1.1 Requisitos Funcionais 66
3.2.1.2 Modelo de Domínio 67
3.2.1.3 Perfil de Descrição 69
3.2.2 Métricas intrínsecas dos conjuntos de dados 74
3.2.2.1 Completude dos Metadados 75
3.2.2.2 Disponibilidade dos Recursos 77
3.2.2.3 Disponibilidade dos Agentes 78
3.2.2.4 Maturidade de Dados Abertos 79
3.2.2.5 Manutenibilidade dos Recursos 81
3.2.3 Métricas Sociais 82
3.2.3.1 Reputação em Redes de Avaliações 84
3.2.3.1.1 Algoritmo RBPD proposto Liu et al. (2017) . . . 85
3.2.3.1.2 Algoritmo RBPD* para a Rede de Avaliações de Conjuntos de Dados Científicos 88
3.2.3.2 Decaimento Temporal de Avaliações 91
3.2.3.3 Confiança do Consumidor 92
3.2.4 Índice de Veracidade de conjuntos de dados científicos 96
3.2.4.1 Cálculo do Índice de Veracidade 97
3.2.4.1.1 Cálculo da Perspectiva Intrínseca do Conjunto de Dados 98
3.2.4.1.2 Cálculo da Perspectiva Social do Conjunto de Dados 98
3.2.4.1.3 Cenário A - Um conjunto de dados, duas avaliações 99
3.2.4.1.4 Cenário B - Rede de Avaliação 103
3.2.4.2 Análise da Robustez do Índice de Veracidade 105

3.2.4.3 Análise do Comportamento do Índice de Veracidade em função do crescimento da rede de avaliações 106

3.2.5 Incorporação do Índice de Veracidade ao Ciclo de Vida dos Dados . 109

4 Estudo de Caso 112

4.1 Descrição do problema científico e dos dados 112

4.2 Compreendendo o conjunto de dados: Ciclo diurno de O_3 116

4.3 Validação do Perfil de Aplicação 118

4.3.1 Avaliando a Confiança no Uso de Conjuntos de Dados 119

4.3.1.1 Análise dos Resultados 122

4.3.1.1.1 Número de Respostas Obtidas 122

4.3.1.1.2 Área de Formação dos Respondentes 122

4.3.1.1.3 Grau de Instrução dos Respondentes 123

4.3.1.1.4 Entendimento sobre o Processo de Geração dos Dados 123

4.3.1.1.5 Confiança em reutilizar o conjunto de dados . . . 125

4.4 Considerações sobre o estudo de caso 127

5 Considerações 129

5.1 Contribuições da Tese 131

5.2 Trabalhos Futuros 132

Referências 134

Apêndice A – Conjunto de Dados GoAmazon em RDF/XML 145

Apêndice B – Questionário - Confiança do Consumidor 150

Sobre você (Opcional) 151

Confiança do Consumidor 151

Comentários Adicionais 152

Apêndice C – Documentos relacionados ao Estudo de Caso - GoAmazon 153

C.1 Relatório Técnico 153

C.2 Descrição do Conjunto de Dados com base no perfil de aplicação 159

C.3 Respostas obtidas 163

1 INTRODUÇÃO

O compartilhamento aberto de dados e publicações com o objetivo de auxiliar o uso e reuso dos dados é um sinal positivo de um ambiente científico em evolução. Nos últimos anos, o compartilhamento de dados e a sua reutilização têm se tornado características de uma prática científica moderna, permitindo que demais pesquisadores analisem as evidências de uma pesquisa, reproduzam os experimentos e verifiquem resultados previamente obtidos com o mínimo esforço possível, além de produzir novos conhecimentos a partir destes elementos. Os dados científicos são elementos essenciais de uma pesquisa baseada em evidências.

A iniciativa de preservar e reutilizar os dados científicos não está associada apenas à ação de pesquisadores isolados, mas a empreendimentos fomentados por instituições de pesquisas de vários países. Nos Estados Unidos, por exemplo, a *National Science Foundation* (NSF) está financiando uma iniciativa denominada *Data Observation Network for Earth* (DataONE), a qual objetiva prover um conjunto de ferramentas e metodologias vinculados ao tratamento, curadoria, análise e compartilhamento dos dados científicos, abordando um ciclo de vida de gestão de dados (WILSON et al., 2014).

Outra iniciativa de destaque internacional é o projeto *Green Ocean Amazon* (GoAmazon), liderado pelo Departamento de Energia dos Estados Unidos da América, o qual objetivou, em parceria com o Instituto Nacional de Pesquisas da Amazônia (INPA), obter dados sobre a composição química e física da atmosfera, na busca pela compreensão acerca dos impactos que a poluição atmosférica provoca no sistema tropical amazônico e suas possíveis influências no clima regional e global. Os dados foram coletados por meio de instrumentos instalados em três sítios experimentais no Brasil, gerando um volume em torno de 2 TBs de dados coletados diariamente e que apresentam grande potencial para a análise e descoberta de novos conhecimentos (FERRELL; KOCH; KUPERBERG, 2014).

Diante deste volume de dados, se faz necessário o uso de técnicas voltadas para a análise e a extração de conhecimento, sistemas e metodologias para a gestão dos dados, de modo a permitir a utilização deste vasto volume de dados para os mais diversos

objetivos, dentre eles, a tomada de decisão com base em uma melhor compreensão dos dados e dos processos de transformação aos quais eles foram submetidos. A integração de bases heterogêneas de dados pode ser uma fonte de novas descobertas. Por exemplo, na área ambiental, novas descobertas podem ser feitas por meio de correlações de dados meteorológicos, socioeconômicos, agropecuários, etc. (PETERSON et al., 2010).

Na busca pela construção de repositórios especializados, há uma necessidade na comunidade científica referente à disponibilização de dados ambientais e de biodiversidade integrados, reunindo um conjunto de dados relacionados à ocorrência de espécies vegetais, animais e de micro-organismos registrados em herbários, museus, coleções zoológicas e microbianas além de sistemas com dados de observação e localização. A disponibilização destes dados poderá fomentar pesquisas na área de ciência e engenharia voltadas para a preservação da biodiversidade, o desenvolvimento sustentável, assim como a construção de ferramentas de tomada de decisão para a implantação de políticas governamentais que buscam à preservação e recuperação da biodiversidade (PETERSON; SOBERóN; KRISHTALKA, 2015).

Em um estudo acerca da disponibilidade de ferramentas computacionais e métodos para a análise e descoberta de conhecimento a partir de um conjunto de dados de biodiversidade, Michener (2015) realizou uma avaliação em mais de 20 portais de dados e ferramentas para a gestão de dados, elicitando uma lista de desafios ainda não superados pela comunidade científica. Dentre os desafios elencados, tem-se:

- O desenvolvimento de padrões de dados e metadados, para uma descrição mais coerente sobre os dados, seu processo de coleta e suas características;
- O desenvolvimento de ferramentas e métodos para a qualidade dos dados, uma vez que há uma grande quantidade de dados depositados nos repositórios científicos que apresentam registros inconsistentes e/ou ausentes;
- A necessidade de uma federação de repositórios que auxilie na descoberta de dados em escala global, tratando questões como redundância de dados, proveniência, etc.;
- O desenvolvimento de métodos analíticos e de visualização que permitam a condução de avaliações acerca da integração dos dados, compreensão semântica, correlação entre dados heterogêneos, incertezas decorrentes do processo de coleta, armazenamento e análise dos dados, dentre outras atividades.

Contextualizando estes desafios ao que está sendo desenvolvido pela comunidade científica, tem-se que os desafios relacionados ao desenvolvimento de padrões de representação

de dados e metadados têm sido endereçados por iniciativas que buscam a representação de metadados pelo desenvolvimento de padrões tais como *Dublin Core*, *Darwin Core* e *Federal Geographic Data Committee Data Standard* (FGDC) (SMITH; PENEV, 2011).

Um conjunto de ferramentas para o tratamento da qualidade de dados tem sido apresentado à comunidade científica como parte do processo de curadoria dos dados. Por exemplo, a ferramenta de limpeza de dados (*data cleaning*) do projeto *speciesLink* tem por objetivo identificar possíveis erros e facilitar a padronização dos dados (CHAPMAN, 2004).

A federação dos repositórios de bases de dados é uma das metas do projeto DataONE, o qual propõe um modelo de ciclo de vida de gestão de dados e a criação de tecnologias que irão lidar com aspectos da integração, tais como redundância e proveniência (CUEVAS-VICENTTÍN et al., 2015).

Com relação a métodos e ferramentas de análise, iniciativas como a ferramenta *Vistrails* (MORISETTE et al., 2013) permitem a elaboração de *workflows* científicos, que auxiliam na compreensão do processo metodológico realizado em uma pesquisa científica, assim como contribuem para a reprodutibilidade dos dados científicos (HARDISTY et al., 2013).

O conjunto de técnicas provenientes do campo de estudo de visualização da informação, denominado Visualização Analítica (em inglês, *Visual Analytics*) (VA), vem sendo estudado nos últimos anos como um elemento no processo de descoberta de conhecimento e tomada de decisão apoiando as atividades de Ciência dos Dados (DASGUPTA et al., 2015).

Adicionalmente aos desafios apresentados, devido à variedade, à velocidade e ao volume de dados disponíveis para a comunidade científica, faz-se importante permitir a compreensão de um conjunto de dados e dos processos de transformação aplicados a este conjunto, principalmente na busca pela identificação de incertezas, vieses de coleta e tratamento dos dados, decisões realizadas durante o experimento, etc. A esta habilidade de compreensão dos dados e dos processos de transformação aos quais estes foram submetidos em um experimento científico, dá-se o nome de *Data Understandability* (BATISTA; CORRêA; PALANISAMY, 2016). Acredita-se que tal habilidade pode ser obtida quando da incorporação de uma nova dimensão para a Ciência dos Dados, denominada Veracidade, a qual foi objeto principal desta tese de doutorado.

1.1 Caracterização do Problema

Nos últimos anos a ciência tem se deparado com uma enorme quantidade de dados provenientes das mais diversas fontes como, por exemplo, a Internet, as mídias sociais, os dispositivos de coleta de dados, etc. Diversas agências de pesquisas estão financiando projetos que buscam promover a gestão, o acesso e a descoberta de conhecimento a partir de tais dados. O cenário formado por amplos conjuntos de dados em larga escala, normalmente gerados e consumidos em tempo real é conhecido como "dilúvio de dados" (em inglês, *Data Deluge*) (SELTZER; ZHANG, 2009; MATTMANN, 2013).

O termo *Big Data* tem sido utilizado para representar esta metáfora. De acordo com Mattmann (2013), existem três dimensões que classificam um conjunto de dados como *Big data*, quais sejam: o *volume* dos dados que um sistema recebe, processa e/ou dissemina; a *variedade*, isto é, o número e a complexidade dos tipos de dados; e a *velocidade* com o qual os dados são criados e/ou disponibilizados para outros usos.

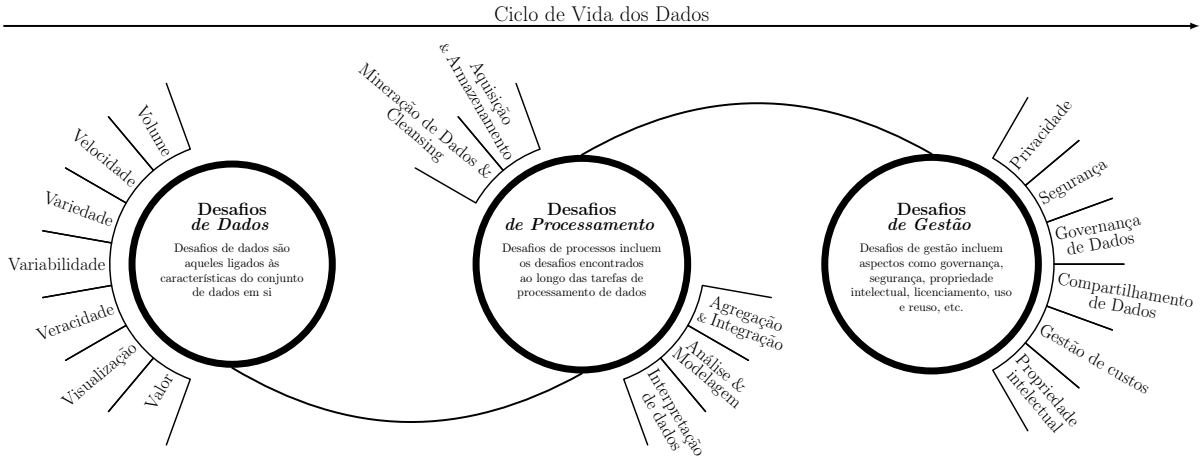
Um crescente número de pesquisas científicas envolvendo *Big Data* tem criado um novo ramo do conhecimento denominado Ciência dos Dados, mais conhecido por *Data Science* (CROSAS et al., 2014).

No trabalho de Sivarajah et al. (2017) os autores apresentam os principais desafios para a área de Ciência dos Dados em um cenário de *Big Data*. A Figura 1 ilustra os desafios apresentados no trabalho, agrupados em três grandes áreas: desafios de dados, desafios de processamento e desafios de gestão, os quais vão surgindo ao longo do ciclo de vida dos dados. Os desafios de dados são aqueles ligados às características de um conjunto de dados em si, tais como: volume dos dados, velocidade de geração e de consumo de dados, variedade nos tipos e formatos de dados, inclusive, a veracidade dos dados.

A veracidade dos dados normalmente está ligada à acurácia dos dados e a adequabilidade dos dados para uso em um modelo analítico. Quando comparada à outras dimensões de *Big Data*, como volume e velocidade, poucos trabalhos abordam a dimensão de veracidade. O desenvolvimento da dimensão de veracidade irá permitir uma maior compreensão do conjunto de dados e de sua adequabilidade aos propósitos científicos.

Os desafios de processamento, por sua vez, são inerentes às tarefas de processamento de dados para a geração de conhecimento, tais como agregação e integração de dados, análise e modelagem de dados, etc. Os desafios de gestão lidam com aspectos tais como privacidade, segurança, governança de dados, compartilhamento de dados, propriedade de dados, etc.

Figura 1: Categorização dos desafios em Ciência dos Dados.



Fonte: Adaptado de Sivarajah et al. (2017).

Para Borgman (2015b), o compartilhamento e o reuso de dados científicos é uma prática comum em poucas áreas da Ciência. Áreas como astronomia e genética, assim como pesquisas sociais, apresentam uma cultura mais desenvolvida para o compartilhamento e o reuso dos dados. A prática de lidar com dados científicos é uma característica específica de cada área do conhecimento, variando de área para área, indivíduo para indivíduo e país para país. Dessa forma, é imprescindível que outras áreas da Ciência amadureçam com relação à gestão dos dados científicos, contemplando suas próprias características e necessidades.

Os dados científicos podem ser compreendidos como representações de observações, objetos ou outras entidades utilizadas como evidências de um fenômeno o qual se busca compreender ou analisar. Raramente os dados são o único elemento de representação de tais observações. Juntamente com os dados se encontram diversos relatórios, *software* de apoio, equipamentos de análise, dentre outros elementos. Dessa forma, o uso dos dados científicos como elementos de reuso e reprodutibilidade dos resultados de uma pesquisa científica são acompanhados de um elemento contextual, que irá permitir a compreensão dos dados e dos processos sob os quais estes foram submetidos ao longo de um experimento científico.

Corrobora com esta visão o trabalho de Borgman (2015b), no qual a autora estuda os benefícios do compartilhamento dos dados científicos. Para Borgman, o compartilhamento dos dados tem se mostrado de pouca utilidade quando não acompanhando de relatórios, *software*, e informações de contato que permitam uma maior compreensão sobre a origem e os procedimentos efetuados nos dados.

O compartilhamento e o reuso dos dados normalmente pode ser compreendido sobre quatro perspectivas: política, legal, tecnológica e econômica (THANOS; RAUBER, 2015). Nas perspectiva política e legal, os estudos buscam compreender como o estabelecimento de políticas de compartilhamento de dados irão contemplar questões como prestígio, citação, propriedade intelectual, legalidade, plágio, dentre outros elementos. A perspectiva tecnológica busca identificar elementos de *hardware* e *software* que irão suportar os processos de compartilhamento e reuso de dados, abordando aspectos como interoperabilidade, compreensão semântica, segurança dos dados, dentre outros. A perspectiva econômica, por sua vez, busca compreender como se dá o suporte econômico no compartilhamento e reuso dos dados, tratando questões como "quem irá pagar pelos compartilhamento dos dados?".

O escopo desta tese permeia as perspectivas política e tecnológica, na busca por abordar como a compreensão dos dados científicos pode ser obtida com base em elementos descritores do conjunto de dados e dos processos de transformações sob os quais estes foram submetidos, que serão incorporados em um processo de gestão de dados científicos, permitindo à comunidade científica uma melhor compreensão sobre o contexto no qual os dados estão inseridos.

O problema de pesquisa desta tese é: *"como caracterizar a dimensão de veracidade de dados, e permitir sua mensuração em conjuntos de dados científicos, de modo a possibilitar uma melhor compreensão do conjunto de dados e dos processos aos quais este foi submetido ao longo de um experimento, de modo a fomentar o reuso dos dados e a reprodutibilidade da ciência?"*.

1.2 Objetivos

Esta pesquisa de doutorado teve por objetivo geral caracterizar um modelo de avaliação de conjuntos de dados científicos, por meio de um perfil de aplicação de metadados com suporte à veracidade dos dados, e de um índice de veracidade de conjuntos de dados científicos.

Para o atendimento do objetivo geral, um conjunto de objetivos específicos foram definidos, quais sejam:

- Descrever o estado da arte sobre Ciência dos Dados e seus desafios;
- Definir o conceito de Veracidade dos Dados;

- Caracterizar um perfil de aplicação de metadados que suporte a descrição de conjuntos de dados científicos, de modo a permitir a criação de métricas que irão compor o índice de veracidade dos dados proposto;
- Definir um modelo de avaliação de conjuntos de dados, gerando assim um índice de veracidade de conjuntos de dados científicos disponibilizados em portais de dados.

1.3 Justificativa

Nos últimos anos diversos portais de dados científicos foram disponibilizados para a comunidade científica, de modo que novas pesquisas estão sendo desenvolvidas com uso destes dados.

Iniciativas internacionais, como o projeto DataONE, buscam estabelecer um processo de gestão do ciclo de vida dos dados, desde sua coleta até a sua análise. O grande volume de dados em larga escala gerado por projetos como, por exemplo, o GoAmazon, fornecem à comunidade científica um desafio com relação à análise dos dados e de seu uso para estudos específicos.

Em Liu et al. (2016) os autores abordam o fato de que embora *Big Data* seja uma área promissora, esta não resolveu os problemas dos dados por si só. Dessa forma, dados ainda precisam ser analisados com critério, de modo a julgar a sua adequabilidade. Para os autores, *"aparentemente, alguns pesquisadores estão imersos no dilúvio dos dados, de modo que tendem a adotar o uso de grandes conjuntos de dados sem questioná-los ou fazer uma análise minuciosa dos mesmos"* (tradução nossa). O trabalho evoca a necessidade de dotar os pesquisadores de elementos que permitam uma compreensão profunda dos conjuntos de dados, de modo que estes possam compreender como as ações e o comportamento de um produtor de dados influenciam na adequabilidade do conjunto de dados para um novo propósito, assim como compreender os vieses dos resultados obtidos com o uso do conjunto de dados e/ou vieses estabelecidos na criação do conjunto de dados.

Berti-Équille e Ba (2016) argumentam que a próxima geração de sistemas de gerenciamento de dados deve lidar não apenas com o volume e a variedade de dados, mas também com a veracidade dos dados. Tais sistemas devem atuar na estimativa da veracidade considerando que os dados podem ser enviesados, com ruídos, desatualizados, incorretos e até mesmo inconfiáveis.

No trabalho de Presser, Weber e Norrie (2014) os autores apresentam a importância do desenvolvimento de metadados que possam ser utilizados como elementos para valida-

ção da adequabilidade de conjuntos de dados para generalização de resultados obtidos em um experimento científico. Para os autores, a confiabilidade nos dados depende de vários critérios, que vão desde o método utilizado para obter tais dados até onde, quando e como estes dados foram transformados ao longo de um experimento.Os produtores de dados devem produzir metadados que descrevem tais aspectos, de modo a permitir um julgamento mais imparcial e objetivo pelos pares, contribuindo para a avaliação da adequabilidade de conjuntos de dados.

Embora técnicas de proveniência de dados buscam mapear as origens dos dados e os processos aos quais estes foram submetidos, Munafò et al. (2017) argumentam que a avaliação de conjuntos de dados deve ser realizada considerando o aspecto social da ciência. A ciência evolui com base na revisão dos pares, na crítica construtiva de cientistas que buscam avaliar o método científico adotado e as conclusões obtidas. Com trabalhos cada vez mais interdisciplinares, a avaliação de um conjunto de dados por especialistas das mais variadas áreas do conhecimento permite com que a comunidade científica possa compreender melhor para quais aplicações e sob quais limitações um conjunto de dados pode ser utilizado em experimentos científicos.

Adicionalmente a este contexto, uma das metas para o ano de 2030 pela União Europeia (DYER; HAVER; NOWLAN, 2012) define que pesquisadores de qualquer disciplina estarão habilitados para encontrar, acessar e processar os dados que necessitam, estando confiantes sobre as suas habilidades de uso e compreensão dos dados, avaliando o grau no qual os dados podem considerados confiáveis. Tal desafio irá requerer o desenvolvimento de métodos de gestão e ferramentas computacionais que possam atender a tais expectativas e necessidades.

1.4 Materiais e Métodos

Para o detalhamento das atividades desta pesquisa, cada objetivo específico foi analisado em termos das ações a serem realizadas para seu alcance, as quais descrevem os resultados esperados.

Para o alcance do primeiro objetivo específico ("Descrever o estado da arte sobre Ciência dos Dados e seus desafios") as seguintes atividades foram consideradas, cujos resultados obtidos estão apresentados no Capítulo 2 deste documento.

A.1 Realizar levantamento bibliográfico sobre Ciência dos Dados;

A.2 Avaliação sobre as principais tecnologias para a gestão de dados, padrões de representação de dados e incertezas no processo analítico.

As seguintes atividades foram executadas para o alcance do segundo objetivo específico ("Definir o conceito de Veracidade dos Dados"), apresentadas no Capítulo 3 deste trabalho:

A.3 Revisão bibliográfica sobre veracidade dos dados;

A.4 Definir o termo veracidade de dados;

As seguintes atividades foram executadas para o alcance do terceiro objetivo ("Caracterizar um perfil de aplicação de metadados que suporte a descrição de conjuntos de dados científicos, de modo a permitir a criação de métricas que irão compor o índice de veracidade dos dados proposto"), apresentados nos Capítulos 3 e 4 deste trabalho:

A.5 Elicitação dos requisitos para um perfil de aplicação com suporte à Veracidade;

A.6 Elaboração do perfil de aplicação;

A.7 Aplicar o perfil de aplicação em um conjunto de dados proveniente do projeto GoAmazon, para verificar o aumento da confiança no uso e reuso do conjunto de dados.;

Para o alcance do quarto objetivo específico ("Definir um modelo de avaliação de conjuntos de dados, gerando assim um índice de veracidade de conjuntos de dados científicos disponibilizados em portais de dados") foram consideradas as seguintes atividades, apresentadas nos Capítulos 3 e 4 deste trabalho:

A.8 Elaboração de métricas de apoio à veracidade calculadas com base nos elementos de metadados que compõem o perfil de aplicação;

A.9 Elaboração do índice de veracidade de dados;

A.10 Elaborar um modelo de avaliação de conjuntos de dados.

1.5 Organização do Trabalho

Esta tese possui, além da Introdução, quatro capítulos, os quais são organizados conforme descrito a seguir.

O Capítulo 2 apresenta o levantamento bibliográfico realizado para o desenvolvimento da pesquisa. Inicia com uma introdução sobre a área de Informática para a Biodiversidade. Em seguida, apresenta as principais bibliografias que buscam conceituar a Ciência dos Dados. Na sequência, são apresentados dois modelos de gestão de dados científicos. O capítulo também aborda as incertezas inerentes ao processo analítico de dados. Também são apresentados os conceitos de metadados, vocabulários controlados e perfil de aplicação. Por fim, o capítulo apresenta os conceitos de formatos *linked data*.

O Capítulo 3 apresenta a proposta de definição da dimensão de Veracidade dos Dados. A partir de um conjunto de requisitos para a gestão, uso e reuso de conjuntos de dados científicos, um perfil de aplicação foi elaborado. Com base neste perfil são definidas métricas de apoio à Veracidade. Tais métricas, juntamente com a avaliação da confiança do consumidor no conjunto de dados, irão formar o índice de veracidade do conjunto de dados.

O Capítulo 4 apresenta o estudo de caso desenvolvido, com o objetivo de aplicar o perfil de aplicação proposto, na busca por verificar se a descrição do conjunto de dados por meio dos elementos de metadados do perfil de aplicação aumenta a confiança no uso e reuso do conjunto de dados.

O Capítulo 5 apresenta as considerações finais desta tese, as principais contribuições e as dificuldades encontradas durante a pesquisa. Também são descritos os trabalhos em andamento e os trabalhos futuros relacionados ao tema.

Ao término da tese, as referências utilizadas são apresentadas, seguidas pelos apêndices que compõem este trabalho.

2 REFERENCIAL TEÓRICO

2.1 Informática para a Biodiversidade e Meio Ambiente

Em 1998, o engenheiro químico e físico teórico Robert May, o qual posteriormente veio a assumir o cargo de presidente da *Royal Society of London* no período de 2000 a 2005, afirmou durante um discurso que: "O próximo século será a 'Era da Biologia', do mesmo modo em que o século XX foi a Era da Física e da Astronomia. Os países que melhor souberem como correlacionar, analisar e intercambiar informações biológicas estarão em uma posição de liderança que lhe permitirão obter avanços econômicos e científicos" (DHILLON; SIDHU, 2013, pg. 1). De fato, houve uma crescente preocupação com a biodiversidade do planeta na última década. Com o avanço da Internet e tecnologias de comunicação de dados, cada vez mais tem se buscado a integração de diversas fontes de dados de biodiversidade.

Há uma necessidade crescente na comunidade científica referente à disponibilização de informações integradas sobre a biodiversidade, reunindo um conjunto de dados relacionados à ocorrência de espécies vegetais, animais e de micro-organismos registrados em herbários, museus, coleções zoológicas e microbianas além de sistemas com dados de observação e localização. A disponibilização destas informações poderá fomentar pesquisas na área de Ciência e tecnologia voltadas para preservação da biodiversidade, desenvolvimento sustentável e também como uma ferramenta de tomada de decisão para implantação de políticas governamentais para preservação e recuperação da biodiversidade (HARDISTY et al., 2013).

Durante a última década diversas iniciativas foram criadas com o objetivo de auxiliar os processos de coleta e análise de dados sobre a biodiversidade, visando uma melhor gestão dos dados e administração dos recursos naturais. Dentre tais iniciativas, destacam-se a *Inter-American Biodiversity Information Network* (IABIN) e a *Global Biodiversity Information Facility* (GBIF). No Brasil, há destaque para iniciativas como a Plataforma de

Informação de Biodiversidade do Ministério do Meio Ambiente (MMA), desenvolvida em parceria com a Universidade de São Paulo (USP), e a plataforma online SiBBr, do Ministério da Ciência, Tecnologia e Inovação (MCTI) do Governo Brasileiro, a qual representa um nodo brasileiro da rede global de sistemas de informação de biodiversidade GBIF.

Uma iniciativa internacional de destaque é a *Data Observation Network for Earth* (DataONE), a qual pretende integrar dados de biodiversidade e ecológicos com demais dados relevantes, para que se possa melhor compreender e monitorar o sistema terrestre, com base em uma correlação de dados climáticos, socioeconômicos, agropecuários, etc. (MICHENER, 2015).

Apesar destes esforços, há ainda muito o que se avançar, no que concerne principalmente à integração semântica dos dados, para permitir que os dados provenientes de diversas fontes sejam interligados, favorecendo a descoberta de novos conhecimentos.

Os problemas mais evidentes reportados na literatura são: (i) cada pesquisador ou grupo de pesquisa gerencia os dados conforme suas necessidades locais, dificultando o processo de análise de dados provenientes de bases de diversos projetos; (ii) a colaboração entre biológos e profissionais ligados à biodiversidade com cientistas da computação e áreas correlatas ainda é pequena, (iii) a falta de um formato padrão ou comum a ser adotado para representação dos dados dificulta a integração dos mesmos e (iv) algumas instituições e alguns pesquisadores são relutantes quanto a publicar ou compartilhar os dados coletados (HARDISTY et al., 2013; DHILLON; SIDHU, 2013).

A problemática referente à gestão de informação de biodiversidade, assim como a integração de diversas bases de dados distribuídas, vem sendo abordada de maneira sistemática pela área de pesquisa em Sistema de Informação denominada *Biodiversity Informatics*. Esta área tem por objetivo enfrentar o desafio de suprir a demanda por tecnologia de apoio à conservação da biodiversidade (PETERSON et al., 2010).

As atividades atuais de pesquisa em *Biodiversity Informatics* estão focalizadas nos seguintes aspectos: a) extração de dados, que incluem a digitalização e a classificação de registros de espécies; b) na compilação de dados, que incluem tarefas como montar comunidades de proprietários de dados, organizando seus dados, e publicá-los em redes de informação; c) e na exibição e visualização desses dados, que incluem a visualização em termos geográficos da distribuição conhecida da espécie (PETERSON et al., 2010).

Em um levantamento sobre o estado da arte e os avanços tecnológicos na área de *Biodiversity Informatics*, Diepenbroek et al. (2014) destacam a necessidade de integrar diversas bases de dados de uma maneira uniforme, a qual permitirá a extração de infor-

mações que estão implicitamente escondidas nestas fontes de dados. A integração dessas fontes de dados permitirá gerar análises, indicadores e estimativas de distribuições de espécies, mecanismos essenciais para o monitoramento da biodiversidade em todo o planeta, principalmente no Brasil, que possui 15% da biodiversidade do planeta. (PETERSON et al., 2010; HARDISTY et al., 2013).

Na área ambiental são diversas as oportunidades de pesquisas suportadas por gestão de dados. Com o intuito de compreender o ciclo de vida de aerossóis em mata intocada, a influência de poluição neste processo e a importância de relâmpagos na formação de NOx, dentre outros objetivos, cita-se o projeto GoAmazon *(Green Ocean Amazon)*. Este projeto, conduzido do início de 2014 ao final de 2015 contou com recursos na ordem dos R$ 24 milhões. Em parceria com o Instituto Nacional de Pesquisas da Amazônia (INPA), o projeto obteve dados sobre a composição química e física da atmosfera para a compreensão sobre os impactos que a poluição atmosférica provoca no sistema tropical amazônico e suas possíveis influências no clima regional e global. Os dados foram coletados por meio de instrumentos instalados em três sítios experimentais principais no Brasil, gerando um volume em torno de 2 TBs de dados coletados diariamente (FERRELL; KOCH; KUPERBERG, 2014).

2.2 Ciência dos Dados

Um crescente número de pesquisas científicas envolvendo *Big Data* tem dado destaque a um ramo de pesquisa denominado Ciência dos Dados, mais conhecido por *Data Science* (CROSAS et al., 2014). A Ciência dos Dados busca elaborar modelos para a descoberta de conhecimento em um grande volume de dados, na ordem de terabytes (TBs, 10^{12} bytes), petabytes (PBs, 10^{15} bytes) e até mesmo exabytes (EBs, 10^{18} bytes), provenientes de múltiplas fontes de dados, incluindo sensores, aplicativos, instrumentos de coleta, mídias sociais e notícias (CROSAS et al., 2014; MCCLURE, 2015).

2.2.1 Histórico

A Ciência dos Dados emergiu nos últimos anos como um ramo de pesquisa, criando também uma atividade profissional denominada “cientista de dados”. Todavia, a preocupação em lidar com um volume de dados e encontrar sentido neste conjunto de dados não é algo recente. Esta seção apresenta uma visão geral dos marcos históricos referentes a criação da área de Ciência dos Dados.

No ano de **1962**, o matemático americano John Tukey publica o trabalho *"The Future of Data Analysis"* (TUKEY, 1962), como um dos artigos de um dos periódicos mais avançados sobre matemática e estatística da época, *"The Annals of Mathematical Statistics"*. Em meio a um conjunto de trabalhos matematicamente precisos, repletos de definições, teoremas e provas, Tukey apresenta um discurso controverso sobre o fato da pesquisa na área de estatística estar limitada, necessitando ser modificada e expandida. Tukey, então, define o termo ***Data Analysis*** como a verdadeira atividade realizada por estatísticos aplicados (DONOHO, 2015). A seguir apresenta-se um trecho do parágrafo inicial do trabalho de Tukey:

> "Durante muito tempo eu pensei que eu era um estatístico interessado em inferir do particular para o geral. Mas como tenho visto as estatísticas matemáticas evoluírem, tenho tido motivos para espanto e dúvida ... Ao todo, cheguei a sentir que o meu interesse principal é na análise de dados, da qual eu incluo entre outras coisas: procedimentos para análise de dados, técnicas para interpretar os resultados de tais procedimentos, formas de planejamento da coleta de dados para fazer a análise mais fácil e mais precisa, e toda a maquinaria e os resultados das estatísticas (matemáticas) que se aplicam a análise de dados."(TUKEY, 1962, pg. 1, tradução nossa).

Tukey inseria assim a estatística como um dos elementos da Análise de Dados. Para Tukey, a Análise de Dados é norteada por quatro elementos, quais sejam (DONOHO, 2015):

1. As teorias formais da estatística;
2. O desenvolvimento acelerado dos computadores e dispositivos de visualização;
3. O desafio de lidar com grandes volumes de dados nas mais diversas áreas do conhecimento;
4. A ênfase na quantificação, pelas mais diversas disciplinas, buscando compreender um conjunto de dados.

Em **1974**, o dinamarquês Peter Naur publica a obra *"Concise Survey of Computer Methods"* (NAUR, 1974), apresentando um conjunto de métodos para o processamento de dados. Para o autor, um dado é uma representação de fatos ou ideias em um formato capaz de ser comunicado ou manipulado por algum processo. Peter Naur apresenta o termo **Datalogy**, como a Ciência dos dados e dos processos relacionados aos dados. Ao longo da obra o autor define **Ciência dos Dados** como "a Ciência de lidar com dados" (PRESS, 2013).

Em **1977**, John Tukey publica a obra intitulada *"Exploratory Data Analysis"* (TUKEY, 1977), ressaltando a necessidade de se construir ferramentas simples para exploração de dados antes de se proceder na construção de um modelo ou na aplicação de alguma técnica de inferência. Tukey enfatiza a diferença entre dois tipos de análises de dados: análise exploratória e análise confirmatória. Na análise exploratória busca-se estudar os dados sob o aspecto individual até o aspecto geral do conjunto de dados, fazendo-se uso de um conjunto de técnicas estatísticas e métodos gráficos para explorar os dados em busca de padrões desconhecidos. A análise confirmatória, por sua vez, busca validar uma hipótese, provar uma tese que explica de maneira geral o comportamento dos dados analisados, por meio de métodos estatísticos gerados a partir de um modelo definido a priori (TUKEY, 1977; PRESS, 2013).

Em **1993**, John Chambers, um dos desenvolvedores da linguagem estatística S, e que trabalhou com John Tukey na *Bell Labs*, publica o artigo *"Greater or Lesser Statistics, A Choice for Future Research"* (CHAMBERS, 1993), exortando a comunidade científica para a necessidade de mudanças na área de Estatística. Em seu *abstract*, Chambers destaca:

> "A estatística enfrenta uma escolha no seu futuro de pesquisa entre a continuação da concentração em tópicos tradicionais - baseado principalmente na análise de dados apoiada por estatísticas matemáticas - e um ponto de vista mais amplo - com base no conceito inclusivo de aprender a partir de dados. O último curso apresenta severos desafios bem como excitantes oportunidades. O primeiro curso é um risco para que a estatística se torne cada vez mais de menor importância."(CHAMBERS, 1993, tradução nossa).

Semelhante ao trabalho de Tukey, Chambers direciona a estatística para um foco voltado à análise de dados e novas formas de representação. Todavia, Chambers vai além da visão Tukey, no sentido de propor a busca por novos conhecimentos a partir dos dados. Para Donoho (2015), Chambers, ao mesmo tempo em que apoia a visão de Tukey, sugere que a estatística deve se expandir para abraçar novos desafios e oportunidades provenientes da análise de dados e descoberta de conhecimento.

No ano de **1996**, realiza-se na cidade de Kobe (Japão) a conferência bi-anual da *"International Federation of Classification Societies (IFCS)"*, denominada IFCS-96, cuja coletânea recebeu como título: *"**Data science**, classification, and related methods"*. Para Press (2013), esta é a primeira conferência a apresentar o termo *Data Science* na composição de seu título.

Em **2001** o cientista da computação americano William Cleveland publica a obra

"Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." (CLEVELAND, 2001). Em sua introdução, Cleveland afirma:

> "Este documento descreve um plano para ampliar as áreas principais do trabalho técnico no campo da estatística. Uma vez que o plano é ambicioso e implica em mudança substancial, será chamado 'Ciência de dados'.
> ...[resultados em] Ciência dos dados devem ser julgados na medida em que eles permitam ao analista aprender a partir dos dados ... Os instrumentos que são utilizados pelo analista de dados são de benefício direto. As teorias que servem de base para o desenvolvimento dos instrumentos são de benefício indireto." (CLEVELAND, 2001, pg. 21, tradução nossa).

Trata-se, portanto, de uma proposta de estruturação de um currículo para o desenvolvimento da Ciência dos Dados. Cleveland confirma a necessidade de se aprender a partir dos dados, corroborando a visão de Chambers.

No currículo proposto por Cleveland são apresentadas seis linhas de atuação: (i) pesquisa multidisciplinar, (ii) modelos e métodos para dados, (iii) computação dos dados, (iv) pedagogia, (v) avaliação de ferramentas e (vi) teoria. A seguir apresenta-se um trecho extraído da obra de Cleveland no qual o autor detalha as seis linhas de atuação, juntamente com o percentual de contribuição de cada linha para a formação do currículo em Ciência dos Dados:

> " - **Pesquisa Multidisciplinar** (25%): colaborações de análise de dados em um conjunto de áreas de assunto;
> - **Modelos e Métodos para Dados** (20%): modelos estatísticos, métodos de construção de modelos, e métodos de estimativa e distribuição baseada em inferência probabilística;
> - **Computação dos Dados** (15%): *hardware*, *software* e algoritmos computacionais;
> - **Pedagogia** (15%): Planejamento de currículo e abordagens de ensino para a escola primária, escola secundária, faculdade, pós-graduação, educação continuada e treinamento corporativo;
> - **Avaliação de Ferramentas** (5%): levantamentos de ferramentas em uso na prática, levantamentos de necessidades percebidas para novas ferramentas e estudos dos processos para o desenvolvimento de novas ferramentas;
> - **Teoria** (20%): fundamentos da Ciência dos dados; abordagens gerais para modelos e métodos, computação com dados, ensino e a avaliação de ferramentas; pesquisas matemáticas de modelos e métodos de computação com dados, de ensino e de avaliação." (CLEVELAND, 2001, pgs. 21-22, tradução nossa).

Além de estabelecer um currículo para o desenvolvimento da Ciência dos Dados, Cleveland apresenta argumentos para a junção das áreas de Estatística e de Ciência da

Computação, na formação desta nova área de estudo. A seguir tem-se alguns trechos que ressaltam esta necessidade:

> "... O benefício para o analista de dados tem sido limitado, porque o conhecimento entre os cientistas da computação sobre como pensar e abordar a análise de dados é limitado, assim como o conhecimento de computação por estatísticos é limitado.
> Uma fusão das áreas de conhecimento produziria uma força poderosa para a inovação. Isto sugere que os estatísticos devem olhar a computação em busca de conhecimento, assim como a como Ciência da Computação olhou para matemática no passado." (CLEVELAND, 2001, pgs. 23-24, tradução nossa).

Com o trabalho seminal de John Tukey a comunidade científica foi exortada a compreender a necessidade de mudança de seu foco e atuação. Todavia, esta mudança começou a acontecer apenas em meados dos anos 2000 (DONOHO, 2015).

Em abril de **2002**, a organização *International Council for Science: Committe on Data for Science and Technology* (CODATA) lança o periódico *"Data Science Journal"* apresentando publicações que focam em tópicos de análise de dados (PRESS, 2013).

Em **2003**, a Universidade de Columbia, nos Estados Unidos, cria o periódico *The Journal of Data Science*, com foco em publicações que aplicam métodos estatísticos em conjuntos de dados (PRESS, 2013).

No ano de **2005**, a *National Science Foundation* (NSF), dos Estados Unidos da América, publica em seu relatório *The National Science Board* o texto denominado *"Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century"*(FOUNDATION, 2005). Os autores definem o perfil profissiográfico do **Cientista de Dados**, formado por um grupo de profissionais das áreas de Ciência da Informação, Ciência da computação, engenharia de *software*, administração de banco de dados, programação de computadores, bibliotecários, dentre outras áreas essenciais para a gestão de uma coleção digital de dados. A tarefa principal destes profissionais é conduzir uma análise e investigação criativa sobre conjuntos de dados (FOUNDATION, 2005).

A partir de **2008**, empresas como *LinkedIn* e *Facebook* já anunciavam vagas em seus *sites* para Cientistas de Dados (PRESS, 2013).

No ano de **2012**, a revista *Harvard Business Review* publicava um artigo descrevendo a profissão de cientista de dados como *"The sexiest job of the 21st century"* (DAVENPORT; PATIL, 2012).

Desde então, a área de Ciência dos Dados se expande e novas pesquisas são feitas com

base nos princípios estabelecidos em seu processo histórico de criação. Como exemplo desta expansão, em março de **2017**, a Universidade de Yale (Estados Unidos) anunciou a transformação de seu departamento de estatística em departamento de estatística e Ciência dos dados (em inglês, *Department in Statistics and Data Science* – popularmente chamado de DS^2).

2.2.2 Definições

Em uma definição comum, Ciência dos Dados é a prática de obter informações valiosas a partir dos dados, lidando com desafios relacionados ao processamento de grandes conjuntos de dados, contendo dados estruturados, não estruturados ou semiestruturados.

Com o advento da Internet e áreas de pesquisa tais como Internet das Coisas (em inglês, *Internet of Things* - IoT), a Ciência dos Dados está se expandindo para lidar com dados gerados a partir de dispositivos inteligentes em tempo real (BATISTA; CORRêA; PALANISAMY, 2016).

Conforme defendido por Cleveland (2001), a área de Ciência de Dados requer um conjunto de habilidades versáteis. Embora cientistas de dados possam se especializar em domínios específicos, tais como *marketing*, medicina, segurança, fraude, finanças, etc., sua atuação depende fortemente de conceitos e práticas das áreas de processamento de sinais, estatística, aprendizado de máquina, processamento de linguagem natural, dentre outras.

Em busca por uma definição formal sobre Ciência dos Dados, encontra-se na literatura os trabalhos de Zhu e Xiong (2015), Brodie (2015), Melorose, Perroy e Careas (2015), Donoho (2015), Dekhtyar (2016), dentre outros. Embora muito se discuta sobre a composição das atividades que permeiam a Ciência dos Dados, o seu conceito ainda não é algo fundamentalmente estabelecido.

Para Zhu e Xiong (2015), há quatro perspectivas gerais sobre a definição de Ciência dos Dados. A seguir se apresenta brevemente cada uma destas perspectivas.

♦ **Ciência dos Dados é a Ciência que estuda dados científicos**.

De acordo com o *Committe on Data for Science and Technology* (CODATA), a Ciência dos Dados é o conjunto de métodos e tecnologias utilizados na condução de pesquisas que fazem uso e gestão de dados científicos.

Sob esta perspectiva, diversas áreas de pesquisa fazem uso de dados científicos e de tecnologias da informação para auxiliar no processamento de dados. Um exemplo

desta perspectiva é a criação das áreas de pesquisa *X-informatics*. Este termo se refere ao conjunto de disciplinas específicas (X, onde X = *astro, geo, chem, urban, econ, environmental* etc.), que fazem uso de conhecimentos das áreas de Ciência da Informação, Ciência Cognitiva, Ciências Sociais, bem como da Ciência da Computação, para o estabelecimento de práticas e procedimentos de processamento de dados, com o objetivo de auxiliar nos ramos específicos de pesquisa.

Nesta ótica, os dados se referem aos dados gerados e utilizados em estudos científicos. A Ciência dos Dados é, então, o conjunto de práticas e técnicas de manuseio, processamento e uso de dados científicos para o suporte das atividades da comunidade científica.

- **Ciência dos Dados é a Ciência que estuda dados de negócio**.

  Para Loukides (2010), a Ciência dos Dados não deve se restringir à gestão, ao processamento e ao uso de dados, mas também deve possibilitar a criação de produtos de dados.

  Provost e Fawcett (2013) corroboram esta visão, afirmando que um dos conceitos fundamentais da Ciência dos Dados é a extração de conhecimento a partir dos dados. Mais especificamente, os autores afirmam que os conhecimentos extraídos dos dados são utilizados para a solução de problemas de negócio, em uma linha muito próxima ao conceito de *Business Intelligence* (BI).

  Esta perspectiva é apoiada por empresas tais como a *Amazon*, *Google*, *LinkedIn* e *Facebook*, que fazem uso de um arsenal de técnicas para processamento dos dados e extração de novos conhecimentos, com o objetivo de apoiar seus processos de negócio e de tomada de decisão empresarial.

- **Ciência dos Dados é a integração de áreas tais como Estatística, Tecnologia da Informação e Inteligência Artificial**.

  Esta perspectiva se baseia no fato de que a Ciência dos Dados pode ser definida com base nas atividades desempenhadas pelos cientistas de dados. Sob este contexto, os cientistas de dados são pessoas com habilidades em estatística, computação, inteligência artificial, dentre outras áreas do conhecimento. Normalmente estas habilidades não ficam restritas a uma única pessoa, mas sim a grupo de profissionais.

  A Ciência dos Dados, por sua vez, consiste na sobreposição de áreas do conhecimento tais como Estatística e Ciência da Computação, buscando o estabelecimento de técnicas para análise e o processamento de dados.

- **Ciência dos Dados busca resolver problemas por meio da extração de conhecimento a partir dos dados**.

  Para Dhar (2013), a Ciência dos Dados consiste no estudo da generalização da extração de conhecimento a partir dos dados. Os cientistas de dados necessitam ter habilidades em áreas tais como estatística, aprendizado de máquina, inteligência artificial, armazenamento de dados, etc. além de um profundo conhecimento da área de aplicação do problema em questão.

  De acordo com Zhu e Xiong (2015), esta perspectiva é uma junção das outras três perspectivas, generalizando a definição de Ciência dos Dados.

Na busca pela definição de Ciência dos Dados, enxerga-se um processo para adquirir conhecimento a partir dos dados, nas mais diversas áreas de pesquisa, provendo suporte para pesquisa científica assim como para processos de tomada de decisão.

No entanto, conforme ressaltado por Zhu e Xiong (2015), estes elementos não são suficientes para definir a Ciência dos Dados como um novo ramo da Ciência, uma vez que o objeto de estudo de tais perspectivas são objetos do mundo real que são naturalmente estudados por outros ramos da Ciência. Um dos objetivos da Ciência da Computação, por exemplo, é representar em forma de dados o mundo real (ZHU; XIONG, 2015). A computação surge para modelar o mundo real de uma maneira computacional, permitindo o estudo e a compreensão dos fenômenos naturais.

Todavia, com o avanço tecnológico e o desenvolvimento da Internet, um grande volume de dados passou a ser gerado de modo a representar inclusive os fatos do ciberespaço que não existem no mundo real, por exemplo, os vírus de computador, os *spammers*, etc. Ao conjunto de todas as possíveis formas de dados existentes no ciberespaço, Zhu e Xiong (2015) dão o nome ***datanature***, para o qual uma possível adaptação para a língua portuguesa, sugerida neste trabalho, é a de **Natureza dos Dados**.

Para Zhu e Xiong (2015), a natureza dos dados é formada por dois tipos macros de dados. O primeiro tipo é composto pelos dados que representam elementos do mundo real. O segundo tipo é composto pelos dados que só existem no ciberespaço, ou melhor, representam elementos virtuais, decorrentes em boa parte das ações humanas no ciberespaço.

O conceito de natureza dos dados produz novos objetos de estudo e problemas científicos, permitindo que a Ciência dos Dados se estabeleça como um novo ramo da Ciência, na busca por questões que superam as questões já existentes nas mais diversas áreas do

conhecimento, tais como (ZHU; XIONG, 2015):

- Qual o tamanho e a variedade da natureza dos dados?
- Qual é a taxa de crescimento desta natureza?
- Como é o fluxo de dados no ciberespaço?
- Como determinar a veracidade dos dados gerados no ciberespaço?

Sob este contexto, Zhu e Xiong (2015) defendem que a Ciência dos Dados não é a simples junção de áreas como Ciência da Computação e Estatística. Uma vez que seu objeto de estudo são os dados, o seu resultado é composto por novos dados, que passam a aumentar a natureza dos dados. Para o autor, a Ciência dos Dados é a Ciência do estudo da natureza dos dados.

A Ciência dos Dados é formada por atividades que buscam extrair conhecimento a partir de um conjunto de dados, por meio da determinação de regras e padrões para os dados em análise. Não há a restrição de que a compreensão obtida a partir dos dados seja mapeada para um elemento do mundo real, uma vez que o objeto de estudo (i.e., o dado) pode ser um elemento existente apenas no ciberespaço. Quando há uma representação do dado no mundo real, normalmente as atividades concentram-se em explicar os fenômenos naturais por meio da análise dos dados.

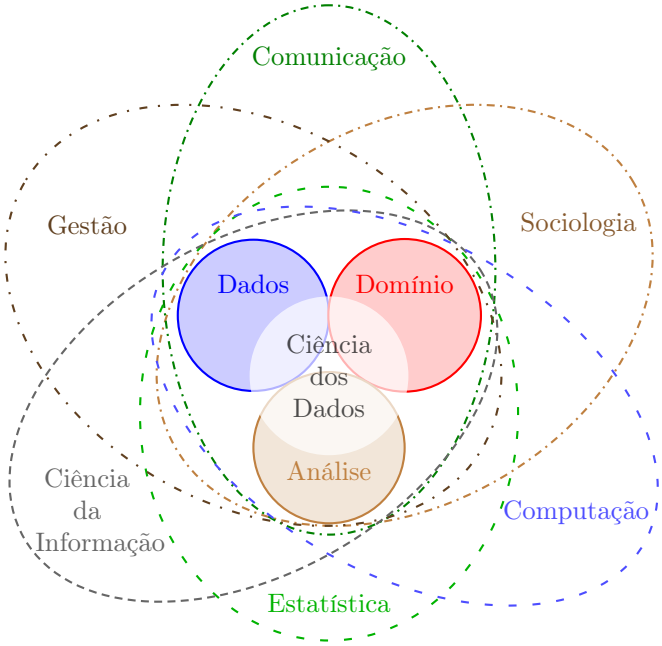
O trabalho de Cao (2017) apresenta os pilares da Ciência dos Dados. Para o autor, a Ciência dos Dados é um novo campo de estudo transdisciplinar que se baseia em uma série de disciplinas, tais como: estatística, Ciência da informação, Computação, Comunicação, Gestão de dados e Sociologia. Sob esta concepção, a Ciência dos Dados busca fazer uso de técnicas e conceitos provenientes destas múltiplas áreas provendo uma abordagem analítica, respeitando as características dos dados e dos domínios aos quais estes pertencem.

2.2.2.1 O Processo de Ciência dos Dados

Embora não haja consenso sobre a definição formal de Ciência dos Dados, surge como elemento comum em todas as definições um processo de manipulação, processamento e análise dos dados, que visa a descoberta de novos conhecimentos.

Para Dekhtyar (2016), a Ciência dos Dados pode ser definida como uma disciplina que permeia o ciclo de trabalho com os dados, lidando com atividades que compreendem

Figura 2: Áreas que compõem a Ciência dos Dados.



Fonte: Adaptado de Cao (2017).

desde a aquisição dos dados, passando pela análise dos dados, até a apresentação dos resultados e a obtenção de novos conhecimentos.

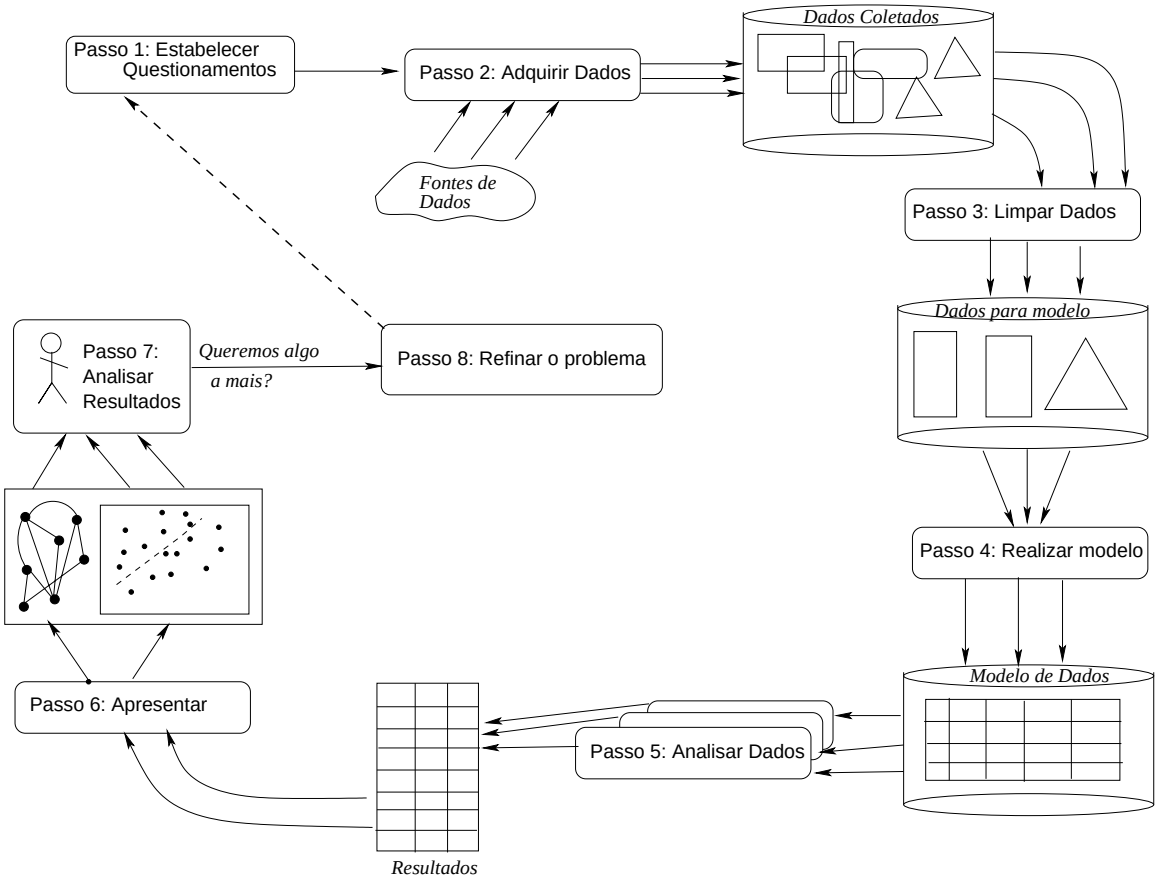
A Figura 3 apresenta o processo de trabalho com dados, na visão de Dekhtyar (2016), sob o contexto da Ciência dos Dados. O ciclo de trabalho é formado por 8 passos, quais sejam:

1. Estabelecer Questionamentos. Este é o primeiro passo no processo de Ciência dos Dados. Estabelecem-se perguntas para as quais os cientistas de dados devem buscar respostas. Normalmente a formulação dos questionamentos acontece em dois estágios. O primeiro precede o passo 2 (adquirir dados), e o segundo acontece antes do passo 4 (realizar modelo), refinando o problema de pesquisa;

2. Adquirir dados. Neste passo diversas fontes de dados são consultadas, formando a massa de dados coletados, que servirão como elementos de análise;

3. Limpar Dados. Nem todos os dados coletados são necessários para responder ao questionamento estabelecido. Da mesma forma, nem todos os dados coletados contêm informações úteis para o problema em questão. Além disso, os dados coletados podem apresentar valores duplicados ou em um formato inapropriado para a realização do modelo. Esse passo consiste na aplicação de técnicas de limpeza de dados (*data cleaning*) com o objetivo de detectar e sanar imperfeições nos dados coletados;

Figura 3: O processo de Ciência de Dados.



Fonte: Adaptado de Dekhtyar (2016).

4. Realizar Modelo. Uma vez que os dados coletados apresentam o formato adequado para a resolução do problema de pesquisa, estes são transformados em uma semântica que permita ao cientista de dados proceder com a análise dos dados. Neste estágio, os dados são agrupados, formatados e/ou transformados para os métodos analíticos. Normalmente os dados passam por procedimento de extração e seleção de características (*feature extraction, feature selection*);

5. Analisar dados. Este é considerado o passo mais fundamental do ciclo. É neste estágio que, fazendo uso de técnicas estatísticas (por exemplo, testes de hipótese, regressão linear, análise multivariada, etc.), de manipulação de dados (operações em bases dimensionais, tais como: *roll-up, slice-and-dice, pivot*, etc.) e de aprendizado de máquina (algoritmos de classificação, clusterização, regras de associação, redes neurais, similaridade, etc.), que os dados são transformados em conhecimento.

6. Apresentar Resultados. Frequentemente os resultados dos métodos analíticos são compostos por diversos elementos e a sua compreensão não é imediata. O uso de técnicas de visualização de dados auxilia na compreensão dos dados e na

consolidação dos conhecimentos adquiridos. Busca-se representar de uma maneira coerente e de fácil entendimento o conhecimento adquirido por meio da análise dos dados.

7. Analisar Resultados. Em um processo linear, este seria o último passo. Neste momento, são apresentadas as considerações sobre os resultados obtidos, e um conjunto de relatórios e artigos científicos é escrito, descrevendo o processo realizado e as descobertas que foram geradas ao longo dos passos anteriores;

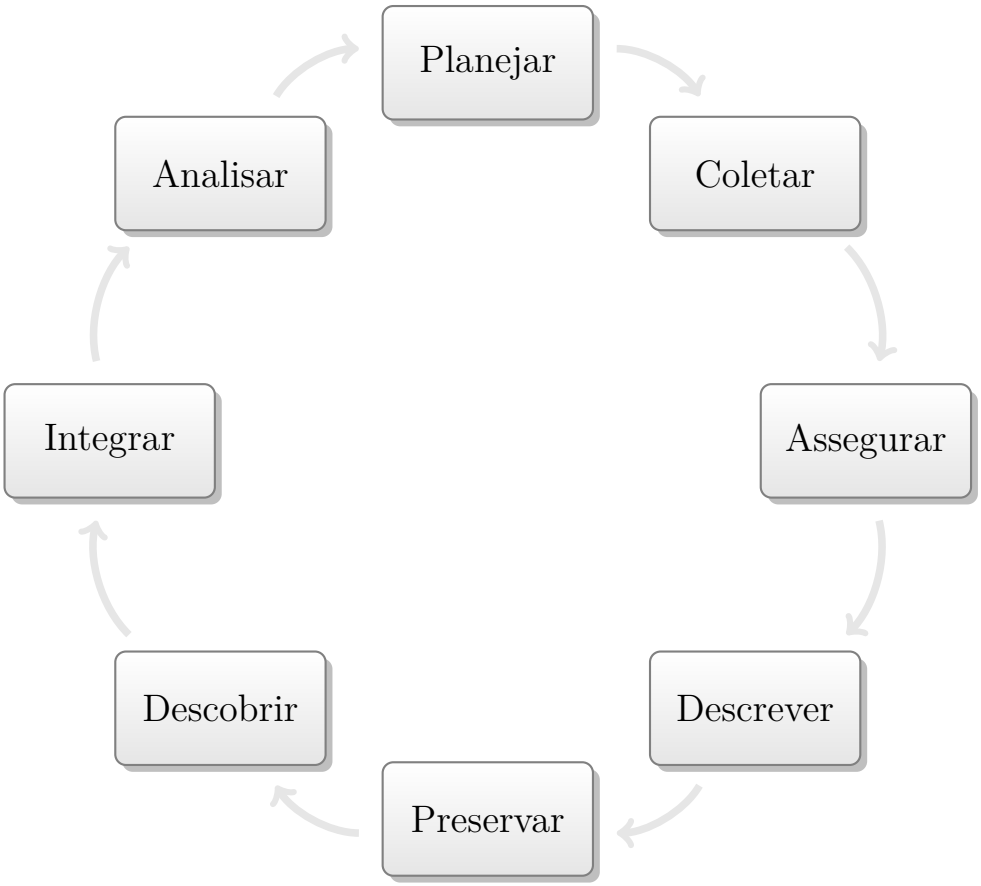
8. Refinar o problema. Se o processo for cíclico, este passo consiste no refinamento do problema de pesquisa, alimentando o passo 1, para a formulação de novos questionamentos ou para um maior detalhamento do questionamento anteriormente feito.

A Ciência dos Dados tem por objeto de estudo os dados em si. Como elemento de entrada neste processo, encontra-se conjuntos de dados, das mais variadas fontes e formatos, que são limpos e transformados, de modo a permitir que um modelo analítico fomente a análise dos dados e a geração de novos conhecimentos. Estas atividades são suportadas por um conjunto de técnicas computacionais, estatísticas e de visualização de dados, facilitando a compreensão sobre os dados que estão sendo analisados. Como resultado deste processo, além de uma série de observações e conclusões científicas, há um novo conjunto de dados, que se insere na natureza dos dados, corroborando a visão de Zhu e Xiong (2015). Normalmente, este conjunto de dados é disponibilizado à comunidade científica para um novo ciclo de análise dos dados, fomentando a busca por novos conhecimentos.

2.3 Modelos de Gestão de Dados

Modelos de Gestão de Dados buscam estabelecer as atividades necessárias para que conjuntos de dados científicos sejam gerados com um nível de qualidade e organização que permita seu armazenamento, recuperação e reuso pela comunidade científica. Nos últimos anos diversas iniciativas propuseram modelos de gestão dos dados científicos. Esta seção apresenta dois modelos de gestão: o modelo proposto pelo projeto DataONE, patrocinado pela *National Science Foundation* (NFS) e o modelo proposto pelo serviço geológico dos Estados Unidos - *United States Geological Survey* (USGS).

Figura 4: Modelo de Gestão de Dados do DataONE.



Fonte: Adaptado de Wilson et al. (2014).

2.3.1 Modelo proposto pelo Projeto DataONE

O projeto *Data Observation Network for Earth* (DataONE) é uma rede de observação de dados para a Terra. Sua estrutura é distribuída numa ciberinfraestrutura sustentável e atende as necessidades da Ciência e da sociedade, em prol do acesso aberto, persistente, robusto e seguro dos dados científicos de observação da Terra (WILSON et al., 2014). O projeto DataONE orienta pesquisadores e profissionais envolvidos com o plano de gerenciamento de dados visando a adoção de melhores decisões e práticas (OLIVEIRA, 2016).

O projeto DataONE desenvolveu um modelo de gestão de dados, denominado *Data Lifecycle Model*, composto por oito estágios apresentados na Figura 4. A seguir tem-se uma visão geral de cada estágio do modelo proposto baseando-se no trabalho de (OLIVEIRA, 2016).

- Estágio Planejar (*Plan*): este estágio está relacionado com o plano de pesquisa que

será desenvolvido pelo pesquisador. Todas as etapas e procedimentos metodológicos sobre o projeto de pesquisa devem ser pensados neste estágio;

- Estágio Coletar (*Collect*): este estágio é direcionado ao processo de coleta dos dados primários, os procedimentos e processamentos em banco de dados, software, laboratórios ou repositórios. Geralmente, os dados primários são oriundos de pesquisa de campo desenvolvida pelo pesquisador ou grupo de pesquisa. É um estágio que apoia toda a coleta e estruturação dos dados desde a sua origem;

- Estágio Assegurar (*Assure*): este estágio está relacionado com os procedimentos de garantia e qualidade dos dados desenvolvidos pelos pesquisadores que apoiam o processo de gerenciamento. São adotados critérios e procedimentos que asseguram que a qualidade e a validação dos dados seja feita por um longo período de tempo;

- Estágio Descrever (*Describe*): contempla a adoção de um padrão de metadados que provê a descrição minuciosa dos dados de pesquisa. É uma fase que requer a adoção de padrões, protocolos e repositórios visando o processamento, armazenagem, curadoria, preservação e disponibilidade dos dados em curto e longo prazos;

- Estágio Preservar (*Preserve*): este estágio trata da adoção de um plano de preservação dos dados por curto e longo prazos. A preservação dos dados em curto prazo tem a finalidade de estocar primariamente e preservar a coleta primária. A preservação de dados em longo prazo tem a finalidade de proporcionar a reutilização dos dados no futuro para outros pesquisadores, parceiros e instituições de pesquisa;

- Estágio Descobrir (*Discover*): está relacionado com a identificação de outros conjunto de dados e repositórios que podem complementar e agregar valor ao projeto de dados;

- Estágio Integrar (*Integrate*): está relacionado com a integração do projeto de dados a outros recursos de dados que possam ser capazes de gerar novas análises e investigações;

- Estágio Analisar (*Analyze*): está relacionado com a análise do projeto de dados por pesquisadores, comunidade científica, parceiros e profissionais;

2.3.2 Modelo proposto pela USGS

A *United States Geological Survey* (USGS) é uma organização científica federal dos Estados Unidos que estuda o ambiente, seus recursos e perigos naturais que ameaçam o

país. A USGS tem quatro maiores disciplinas científicas relativas à Biologia, Geografia, Geologia e Hidrologia. Pertence ao *United States Department of The Interior* como uma agência científica única e exclusiva desse departamento (OLIVEIRA, 2016).

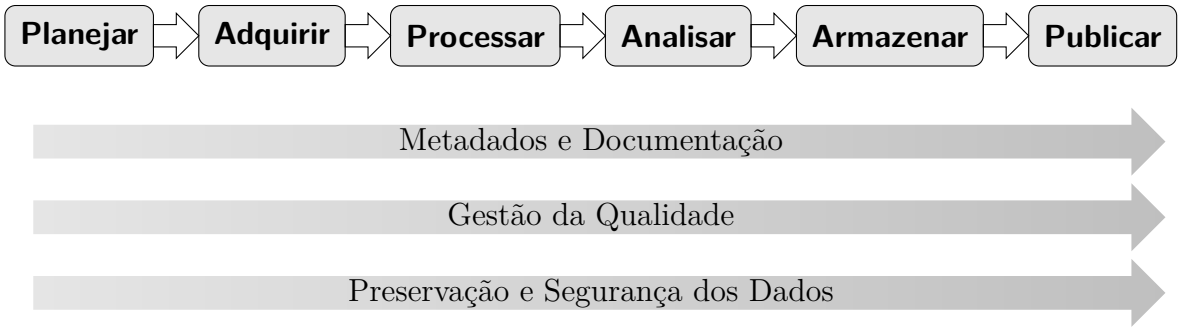
A USGS representa dados corporativos com valor potencial para além de algum uso imediato de pesquisa, e, portanto, esses precisam ser considerados e devidamente gerenciados durante todo o seu ciclo de vida. Com a devida importância relacionada ao gerenciamento de dados científicos, foi constituída no âmbito da USGS uma equipe que desenvolveu o *Scientific Data LifeCycle Model* (SDLM) com uma visão de alto nível dos dados, desde a sua concepção até a preservação e compartilhamento, para ilustrar como as atividades de gerenciamento de dados referem-se aos fluxos de trabalho do projeto e para ajudar na compreensão das expectativas adequadas de gerenciamento de dados (OLIVEIRA, 2016).

Em 2009, foi criada pela USGS a *Community for Data Integration* (CDI). A CDI representa uma comunidade dinâmica de práticas focadas no avanço de dados científicos e gerenciamento de informações e recursos transversais de integração dentro da USGS. Em 2010, a CDI estabeleceu o *Data Management Working Group* (DMWG) para desenvolver, ampliar e recomendar as melhores práticas e políticas para assistir a organização efetivamente na manipulação, documentação, preservação e promoção do acesso à agência de dados científicos. O subgrupo de trabalho investigou exaustivamente a existência de modelos de ciclo de vida de dados, que resultou na compilação de aproximadamente 50 (cinquenta) modelos advindos da academia, agências federais e organizações internacionais. Tais modelos diferenciavam-se entre si na granularidade, complexidade, apresentação e perspectiva da comunidade, bem como diferentes modelos de ciclo, por exemplo, circulares, espirais, lineares e hierárquicos (FAUNDEEN et al., 2014; OLIVEIRA, 2016).

O modelo de gestão de dados da USGS, apresentado na Figura 5 consiste em um fluxo intuitivo (da esquerda para a direita) e um ponto de partida claramente definido que se alinha com o início do projeto de pesquisa. O desenho final do modelo foi revisado e aceito em novembro de 2012 pelos patrocinadores executivos do CDI.

A finalidade primordial com a aplicação desse modelo às atividades de pesquisa era de que cientistas da USGS pudessem garantir que os produtos de dados fossem bem descritos, preservados, acessíveis e aptos para reutilização. Este modelo define também os diferentes papéis e responsabilidades para serem realizados pelo pesquisador e curador dos dados de pesquisa, que inclui dados e atividades de gestão que exigem conhecimento e habilidades especializados, bem como a educação permanente sobre os métodos e padrões.

Figura 5: Modelo de ciclo de vida dos dados de pesquisa proposto pela USGS.



Fonte: Adaptado de (FAUNDEEN et al., 2014).

Os pesquisadores ou grupo de pesquisa são responsáveis pelo planejamento, preservação e publicação/compartilhamento dos dados. Os curadores dos dados são responsáveis pelos estágios de coleta, processo, análise e, em conjunto com o pesquisador, realizam a publicação e o compartilhamento. Os elementos primários do modelo que compõem o *Data Life Cycle* da USGS são (OLIVEIRA, 2016):

1. **Planejar**: é a primeira atividade do ciclo, a qual destina assegurar e considerar todas as atividades relacionadas ao manejo de ativos de dados do projeto, desde sua concepção até o arquivamento e publicação dos dados gerados. Durante este estágio, todos os elementos do modelo devem ser avaliados, tratados e documentados. Um plano de gestão de dados (em inglês, *data management plan* - DMP) é o resultado almejado nesta atividade;

2. **Adquirir**: é a segunda atividade do ciclo e representa as tarefas pelas quais os dados são coletados ou gerados. Esta é uma etapa importante que requer dos pesquisadores habilidades e técnicas de pesquisa para a coleta de dados, considerando as questões éticas e metodológicas, a fim de manter a qualidade, proveniência e integridade dos dados;

3. **Processar**: a terceira atividade do ciclo representa as tarefas ligadas ao processamento de dados, integração de múltiplos conjuntos de dados e as transformações necessárias para a criação de um modelo. Nesta etapa do ciclo, os pesquisadores devem fazer uso de padrões e ferramentas adequados para as necessidades do projeto, registrando todas as atividades efetuadas;

4. **Analisar**: a quarta atividade representa as tarefas relacionadas com a exploração e interpretação do processamento de dados, quando as hipóteses são testadas, as descobertas são feitas e as conclusões são desenhadas. As atividades analíticas

nesse estágio do processo incluem sumarização, desenhos, análises estatísticas e espaciais, modelos que são usados para produzir informações e resultados científicos. Conforme novos dados são gerados, suas versões e os processos realizados para tal geração são devidamente documentados. O gerenciamento de dados durante a análise aprimora a eficiência das atividades, preserva a documentação que é crítica para a integridade científica e cria uma base para futuras pesquisas;

5. **Armazenar**: representa as tarefas associadas ao armazenamento de dados para o uso e acessibilidade por longo prazo. Nesse estágio a publicação e o compartilhamento dos dados reforça o compromisso de que cientistas financiados pelo governo federal devem planejar a preservação de dados, metadados, produtos auxiliares, formatos de armazenamento de aplicativos neutros e toda a documentação adicional a longo prazo, garantindo a disponibilidade e reutilização futura;

6. **Publicar**: a publicação e disseminação de dados são componentes críticos que devem receber especial atenção. Essa etapa lembra aos cientistas que tanto os dados como as publicações tradicionais são produtos de pesquisas e devem ser publicados e mantidos em portais de dados com recursos adequados que permitam a busca e recuperação dos dados e demais artefatos associados.

Além das etapas principais, o modelo da USGS estabelece elementos transversais, denominados *cross-cuting elements*. São três os elementos, quais sejam (OLIVEIRA, 2016):

- **Metadados e Documentação**: ao longo de todas as etapas do ciclo, os pesquisadores devem desenvolver metadados e documentações associadas com base em padrões bem estabelecidos. O registro das informações em cada etapa garante a precisão e compreensão dos dados científicos criados, compilados, processados e compartilhados. Com isto, busca-se permitir que outros cientistas possam validar os resultados obtidos por meio da replicação do experimento, assim como ampliar o experimento pelo reuso dos dados;

- **Gestão da Qualidade**: os pesquisadores devem adotar medidas de garantia da qualidade dos dados desde o início do projeto. Este elemento busca alinhamento com as políticas mandatórias das agências de financiamento as quais propõem a produção de artefatos científicos de alta qualidade;

- **Preservação e Segurança dos Dados**: estabelece estratégias de preservação dos dados que garantam a disponibilidade dos dados científicos, assim como a segurança

> de acesso a tais dados. Para a USGS, os "dados representam ativos corporativos com valor potencial para além de qualquer uso em investigação imediata e, portanto, precisam ser considerados e devidamente gerenciados durante todo o seu ciclo de vida" (FAUNDEEN et al., 2014).

O modelo da USGS também atua como uma estrutura para auxiliar a organização avaliar e otimizar as políticas e práticas para o gerenciamento de dados científicos, identificando as áreas que necessitam desenvolver novos instrumentos e padrões. No trabalho de Arass, Tikito e Souissi (2017) os autores analisaram 12 modelos de gestão de dados propostos nos últimos anos, sendo que o modelo de gestão da USGS foi o melhor avaliado, por cobrir um processo científico ponta-a-ponta e demonstrar um modelo de fácil implementação quando comparado aos demais modelos.

2.4 Incertezas e o Processo Analítico de Dados

Incertezas são inerentes ao método científico (SCHUNN; TRAFTON, 2012). Desde a coleta dos dados até a interpretação dos resultados provenientes de modelos, algoritmos e simulações. Muitas vezes são formadas por tarefas de transformação de dados que adicionam novas incertezas ao conjunto de dados e que precisam ser evidenciadas para uma melhor interpretação dos resultados obtidos (JAIN, 2015).

Maceachren (2015) argumenta que a incerteza, diante de um conjunto de dados heterogêneos em larga escala, leva a um cenário em que a extração de conhecimento é complexa, podendo colaborar para uma compreensão equivocada, principalmente quando a incerteza dos dados e do processo de transformação de dados não é levada em consideração. Além da necessidade de se determinar o volume adequado de dados para utilização, a complexidade dos dados pode ser melhor compreendida e mitigada por meio de técnicas de análise estatística, enquanto que os equívocos de interpretação e geração de conhecimento podem ser minimizados por meio de estratégias que buscam uma melhor interpretação dos resultados obtidos e do processo de seleção e transformação do conjunto de dados.

Muitos algoritmos e técnicas estatísticas para a manipulação dos dados envolvem conceitos matematicamente complexos que, frequentemente, são mapeados como parâmetros de um modelo. Na área de biodiversidade, mais especificamente na construção de modelos de distribuição potencial de espécies (PETERSON; SOBERóN, 2012), os modelos gerados são, em sua maioria, executados por especialistas do domínio, que conhecem o comportamento de uma espécie e como um conjunto de variáveis bioclimáticas afetam sua

distribuição espacial, sobrevivência e reprodução.

Porém, conforme ressaltado por Barry e Elith (2006), os erros encontrados em modelos de distribuição potencial de espécies possuem duas principais causas, quais sejam: (i) a deficiência nos dados, isto é, erros introduzidos no processo de coleta, digitalização e georreferenciamento, além eventuais vieses que podem ser inerentes ao processo de coleta (por exemplo, dados coletados apenas uma região de fácil acesso, mas que não ficam evidentes ao novo pesquisador que está fazendo uso de tais dados); e (ii) a parametrização do modelo, que requer um conhecimento apurado do método e que pode levar a uma interpretação equivocada, contribuindo para que novos conhecimentos sejam ocultados e ou eliminados por não se considerar as incertezas inerentes ao processo.

Com o aumento da disponibilidade de ferramentas de portais de dados (MICHENER, 2015), as quais agrupam dados das mais diversas fontes, coletados por meio de diversos métodos, em diferentes escalas e com diferentes objetivos de pesquisa, o processo de escolha de um subconjunto de dados para a análise e de parâmetros para um modelo computacional torna-se uma atividade desafiadora (QIN; LI, 2013).

No trabalho de Qin e Li (2013), é apresentada uma série de desafios relacionados à incerteza no processo de análise de dados em larga escala, tais como: (i) incerteza inerente aos dados, decorrente do processo de obtenção dos dados; (ii) incertezas geradas no processo de seleção de subconjunto de dados por meio de métodos de redução da dimensionalidade, agrupamento, etc.; (iii) incertezas provenientes de questões sigilosas (como a falta de precisão na localização de espécies ameaçadas de extinção como um mecanismo de proteção adicional, evitando que os dados sejam utilizados com propósitos que possam ameaçar a existência da espécie); (iv) valores ausentes, seja por falha em equipamentos, inconsistência de registros, etc.; e (v) integração de dados de diferentes fontes de dados.

Além da incerteza inerente aos dados e ao processo de escolha de um subconjunto de dados para um modelo computacional, Schunn e Trafton (2012) realizaram um estudo sobre a incerteza dos resultados obtidos por meio de um processo científico que faz uso de processamento de um grande volume de dados. Fazendo uso dos resultados obtidos, os autores propuseram uma taxonomia da incerteza gerada por métodos computacionais.

Nesta taxonomia, os tipos de incerteza oriundos de métodos computacionais em um conjunto de dados são classificados em (SCHUNN; TRAFTON, 2012) :

- *predição futura* - a incerteza de dados relacionados ao futuro com base em dados

obtidos em uma escala temporal anterior e que precisa ser evidenciada para evitar interpretações errôneas;

- *artefatos estatísticos* - os métodos estatísticos e computacionais normalmente atuam nos dados realizando ações de agrupamento e classificação, as quais inserem um conjunto de incertezas nos dados e que podem mascarar potenciais novos conhecimentos e correlações, assim como possibilita uma compreensão incoerente sobre o fenômeno estudado;

- *estratégias consideradas "rápidas e econômicas"* - as quais, em função do volume de dados, aplicam algumas técnicas para redução do esforço computacional e, consequentemente, do tempo necessário para a realização de análises, tais como técnicas para redução da dimensionalidade (WANG; CHIOU; MUELLER, 2015), podendo introduzir elementos de incerteza que precisam ser levados em conta no momento da análise e interpretação dos resultados.

2.5 Metadados

A definição mais comum para o conceito de metadados é a de "dados sobre dados". Na definição apresentada em NISO (2004), os metadados consistem de informações estruturadas que buscam descrever, explicar, localizar e facilitar a recuperação, o uso e a gestão de um recurso de dados. Qualquer objeto que possa ser identificado, armazenado e descrito é denominado recurso de dados.

Outra definição para metadados é apresentada por Vellucci (1998, pg. 192): "metadado é um dado que descreve atributos de um recurso, caracteriza suas relações, apoia sua descoberta e uso efetivo, e existe em um ambiente eletrônico. Usalmente consistem em um conjunto de elementos, cada qual descrevendo um atributo do recurso, seu gerenciamento, ou uso".

Os metadados são classificados em três tipos, quais sejam (NISO, 2004):

- Metadados descritivos: descrevem os recursos para garantir a descoberta e identificação. Estes metadados incluem elementos como título, resumo, autores e palavras-chave;

- Metadados estruturais: indicam como os objetos de dados são organizados e interligados. Por exemplo, um livro digital pode ser publicado como diversos arquivos

de imagens individuais, as quais representam as páginas do livro. Tais páginas são geralmente agrupadas em capítulos. A informação sobre quais páginas compõem um capítulo podem ser armazenadas em um metadado estrutural sobre o livro;

- Metadados administrativos: disponibilizam informações que auxiliam na gestão do recurso de dados. São informações como origem do recurso, detalhes do seu formato, informações técnicas, etc.

De modo a garantir interoperabilidade e compreensão semântica em múltiplos contextos, diversas iniciativas buscam padronizar as representações de metadados. Os padrões ou esquemas de metadados possuem elementos descritores organizados para descrever objetos do domínio de informação do usuário. Eles possibilitam melhor armazenamento, manipulação, recuperação e troca de informações entre aplicações e entre organizações, de forma rápida, automatizada e eficiente (SMITH; SCHIRLING, 2006).

Uma iniciativa de destaque é o padrão *Dublin Core* (WEIBEL et al., 1998). Este padrão, também denominado esquema, é constituído de 15 elementos e diversos qualificadores, e foi concebido almejando algumas características básicas: simplicidade, interoperabilidade semântica, consenso internacional, flexibilidade e modularidade na Web.

2.5.1 Vocabulários

Um vocabulário é um padrão de metadados criado para um objetivo específico. É por meio de um vocabulário que são definidos os elementos (ou termos) utilizados de forma universal para determinado objetivo, permitindo que o compartilhamento destes dados se torne mais prático e seguro.

Um vocabulário geralmente define nomes para os elementos presentes e um significado específico e bem definido. Adicionalmente, são especificadas regras relacionadas ao conteúdo destes elementos, como o tipo e o formato da informação aceito, sua cardinalidade, e a recomendação ou obrigatoriedade de sua existência.

Um vocabulário pode ser definido como *controlado* quando existe uma lista restrita de palavras ou termos que podem ser utilizados para catalogação, marcação ou indexação. Por exemplo, um elemento de metadado comum para descrição de um recurso é o seu tipo. O padrão *Dublin Core* (WEIBEL et al., 1998), por meio da *Dublin Core Metadada Initiative* (DCMI), define um conjunto de termos que podem ser utilizados para definir o tipo de recurso, apresentado na Tabela 1.

Tabela 1: Vocabulário controlado sobre os tipos de recursos proposto pela DCMI.

Termo	Definição
Collection	Uma agregação de recursos.
Dataset	Dados codificados em uma estrutura definida.
Event	Uma ocorrência baseada no tempo.
Image	Uma representação visual, diferente de texto.
Interactive Resource	Um recurso que requer interação do usuário para ser compreendido, executado ou experimentado.
Moving Image	Uma série de representações visuais que transmite uma impressão de movimento, quando mostrado em sucessão.
Physical Object	Um objeto ou substância tridimensional inanimada.
Service	Um sistema que fornece uma ou mais funções.
Software	Um programa computacional na forma de código ou compilado.
Sound	Um recurso que pode ser ouvido.
StillImage	Uma representação visual estática.
Text	Um recurso que consiste principalmente de palavras para leitura.

Fonte: Adaptado de (BOARD, 2008).

Um vocabulário controlado pode ser incorporado em processos automatizados, simplificando atividades de entrada de dados, a partir do fornecimento de listas selecionáveis para os usuários, além de suportar a validação dos dados importados, provenientes de outros sistemas. Com uso dos vocabulários controlados, a aplicação de técnicas para identificação, localização e análise de dados se tornam mais eficientes, uma vez que há padronização e organização em seu conteúdo.

2.6 Perfil de Aplicação

Um perfil de aplicação define os termos para a anotação de recursos com metadados, e relata como usá-los de forma consistente de acordo com normas específicas de uso (ALEIXO; CATARINO, 2015).

Um perfil de aplicação, conforme definido por Heery e Patel (2000), é um esquema que consiste de elementos de dados elaborados a partir de um ou mais padrões de metadados, combinados por implementadores, e otimizado para uma aplicação local em particular.

Coyle et al. (2009) descrevem que cada metadado possui necessidades diferentes e aplicações variadas, resultando numa diversidade de formatos e padrões. Criar um perfil de aplicação, permite definir quais padrões de metadados serão utilizados, além de estabelecer como e onde os termos descritivos são melhores empregados.

Com o objetivo de estabelecer um processo para a criação de perfis de aplicação, a DCMI publicou o trabalho "*Singapore Framework for Dublin Core Application Profiles*" (NILSSON; BAKER; JOHNSTON, 2008), definindo um conjunto de componentes descritivos necessários para a documentação de um perfil de aplicação. Os componentes propostos são agrupados como segue:

- Requisitos funcionais: esse é um componente obrigatório da documentação de um perfil de aplicação, e descreve as funções para as quais o perfil de aplicação foi projetado, assim como explicita o que está fora do escopo do perfil em questão. Os requisitos funcionais formam a base para a avaliação do perfil de aplicação quanto à sua consistência interna e para orientar sua correta utilização;

- Modelo de domínio: esse é um componente obrigatório da documentação de um perfil de aplicação e define as entidades básicas descritas no perfil de aplicação e seus relacionamentos funcionais. O modelo de domínio pode ser documentado em texto, diagramas ou usando uma abordagem formal, como a *Unified Modeling Language* (UML) (FOWLER, 2014);

- Perfil de Descrição (*Description Set Profile - DSP*): este componente obrigatório define o conjunto de metadados que são as instâncias válidas de um perfil de aplicação. Um perfil de descrição pode detalhar restrições nos recursos, nas propriedades e na forma que um valor pode ser registrado;

- Diretrizes de uso: este componente opcional descreve como utilizar o perfil de aplicação e como as propriedades devem ser consideradas no domínio de aplicação;

- Diretrizes de sintaxe da codificação: este termo opcional descreve regras e diretrizes sobre a codificação utilizada no perfil de aplicação.

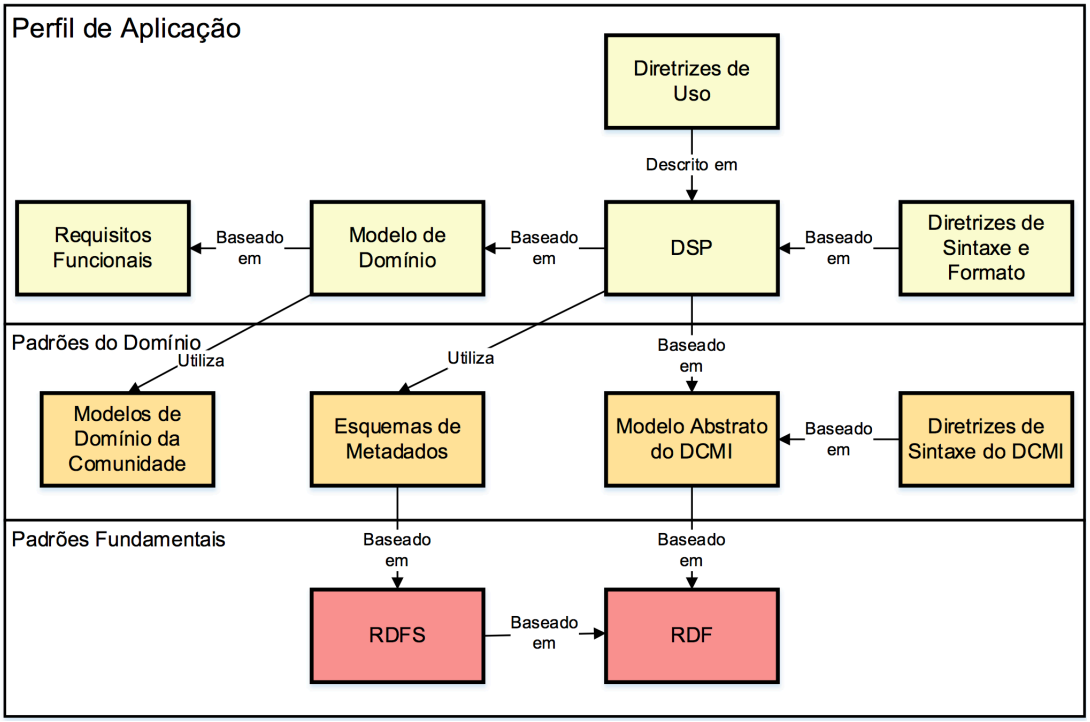
A Figura 6 apresenta as relações entre cada um dos componentes definidos por Nilsson, Baker e Johnston (2008). A representação de metadados nesta proposta se dá pelo uso de formatos *linked data*, apresentados a seguir.

2.7 *Linked Data*

Os componentes do perfil de aplicação estão voltados para uma maior compreensão semântica. Todavia, Nilsson, Baker e Johnston (2008) também definem padrões fundamentais de sintaxe.

As sintaxes mais comuns utilizadas para a codificação de metadados são textos simples (*plain texts*), *eXtensible Markup Language* (XML), *JavaScript Object Notation* (JSON) e *Resource Description Framework* (RDF). Em geral, cada uma destas sintaxes possui suas próprias ferramentas para criação e recuperação dos metadados.

Figura 6: Componentes de um Perfil de Aplicação.

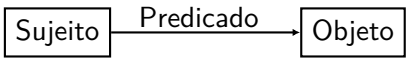


Fonte: (SILVA, 2017).

A sintaxe adotada por Nilsson, Baker e Johnston (2008) é a sintaxe do *framework* RDF. Este framework foi desenvolvido pelo *World Wide Web Consortium* (W3C), sendo um modelo padronizado para a representação e o compartilhamento de recursos de dados na Web.

RDF é formado por um conjunto de afirmações (em inglês, *asserts*) sobre um conjunto. Uma afirmação em RDF consiste de três elementos (uma tripla) e tem a seguinte estrutura: sujeito, predicado e objeto, visualmente apresentada na Figura 7.

Figura 7: Representação de uma afirmação em RDF, descrita por uma tripla.

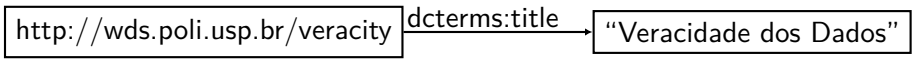


Fonte: o Autor.

Uma afirmação RDF expressa uma relação entre dois recursos: o sujeito e o objeto. O predicado representa a natureza da relação, a qual é definida de forma direcional (do sujeito para o objeto). O conjunto de triplas é chamado de grafo RDF, que pode ser representado de maneira visual como um grafo direcionado.

A identificação única e universal dos recursos e das propriedades descritas no RDF são realizadas por meio do *Uniform Resource Identifier* (URI), mais especificamente o HTTP

Figura 8: Representação de uma afirmação em RDF, descrita por uma tripla.



Fonte: o Autor.

URI (BIZER; HEATH; BERNERS-LEE, 2009). Os tipos mais comuns de URI são as *Uniform Resource Locators* (URLs), que possibilitam o acesso a uma representação visual do recurso descrito, como uma página Web, e os *Uniform Resource Names* (URNs), que são usados para identificar recursos, mas que não especificam uma localização ou forma de acesso ao recurso. Os URIs podem ser usados em qualquer uma das partes da tripla RDF (sujeito, objeto ou predicado).

Em uma tripla, o objeto também pode ter um valor textual. Nesta situação, a tripla é chamada de propriedade, pois representa uma propriedade do sujeito descrito.

Para dar um significado semântico aos predicados considerados nas triplas, são utilizados os vocabulários e as ontologias, que definem descrições formais dos significados de cada elemento do esquema, diminuindo as ambiguidades da linguagem natural. Estas descrições formais são identificadas por meio do *namespace* do vocabulário, representado também por uma URI. Por exemplo, o *Dublin Core Terms* é representado pelo *namespace "http://purl.org/dc/terms/ "* e pelo prefixo "*dcterms*". O prefixo pode ser utilizado para simplificar a representação de um *namespace*. A Figura 8 apresenta um exemplo de uma tripla em RDF. O recurso é identificado pela URL "*http://wds.poli.usp.br*" e seu título é "*Veracidade dos Dados*", o qual é definido pelo predicado "*dcterms:title*". Neste exemplo, *dcterms* é o prefixo para o *namespace* "*http://purl.org/dc/terms*"

Múltiplos *namespaces* podem ser utilizados na descrição de um mesmo recurso, possibilitando que elementos de diferentes vocabulários possam ser combinados em uma mesma descrição. Múltiplas descrições, criadas em momentos diferentes, com propósitos distintos, também podem ser ligadas umas às outras por meio do RDF.

O RDF pode ser codificado ou serializado por diferentes representações sintáticas, algumas mais adequadas para o processamento de máquina, e outras mais legíveis para as pessoas. Entre as notações mais populares estão o RDF/XML, o Turtle, o N-Triples e o JSON-LD. Tais representações são conhecidas como representações *linked data*. A Figura 9 apresenta a descrição em RDF/XML para a tripla RDF apresentada na Figura 8.

Uma vez que os metadados estão descritos em formato de *linked data*, estes podem ser armazenados em banco de dados de triplas, e terem seu conteúdo consultado por meio da

Figura 9: Descrição em RDF/XML da tripla representada na Figura 8.

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <rdf:RDF
3   xmlns:dcterms="http://purl.org/dc/terms/"
4   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
5 >
6  <rdf:Description rdf:about="http://wds.poli.usp.br/veracity">
7    <dcterms:title>Veracidade dos Dados</dcterms:title>
8  </rdf:Description>
9 </rdf:RDF>
```

Fonte: o Autor.

linguagem SPARQL (*SPARQL Protocol and RDF Query Language*) (PRUD; SEABORNE et al., 2006). Com consultas em SPARQL, é possível realizar consultas em grafos RDF. A Figura 10 apresenta um exemplo de consulta em SPARQL para obter a URL de um recurso cujo título é "Veracidade dos Dados". Com base no RDF/XML da Figura 9, a resposta para esta consulta é o recurso identificado pela URL *http://wds.poli.usp.br/veracity*.

Figura 10: Consulta em SPARQL sobre o RDF/XML definido na Figura 9.

```
1 PREFIX dcterms:<http://purl.org/dc/terms>
2 SELECT ?url
3 WHERE
4 {
5 ?url dcterms:title "Veracidade dos Dados" .
6 }
7
8 Resultado:
9 "http://wds.poli.usp.br/veracity"
```

Fonte: o Autor.

As variáveis SPARQL começam com um caractere ? e podem ser definidas em qualquer elemento da tripla (sujeito, predicado ou objeto). Tanto na cláusula SELECT quanto na cláusula WHERE é possível definir triplas, juntamente com as variáveis, para a encadeamento das regras da consulta. Além disso, o SPARQL possibilita a inclusão de filtros e operadores para possibilitar a realização de consultas complexas.

3 VERACIDADE DOS DADOS

Normalmente o termo *Big Data* é conceituado em função de suas características, as quais são costumeiramente nomeadas por substantivos que iniciam com a letra V, dentre as quais as mais predominantes são Volume, Velocidade e Variedade. Volume refere-se ao tamanho dos conjuntos de dados que são utilizados em *Big Data*. Com dados sendo produzidos a taxas cada vez mais rápidas, a característica de velocidade torna-se importante em virtude da necessidade de se encontrar métodos computacionais para processar e analisar fluxos (*streams*) de dados. Nem sempre o armazenamento destes fluxos de dados é possível, gerando a necessidade de métodos de análises *on-line* (mais conhecidos pelo termo *on the fly analysis)*. A variedade ou complexidade dos dados é outro ponto importante em *Big Data*. Uma vez que métodos para geração e consumo de fluxos de dados estão se tornando cada vez mais sofisticados, assim como os problemas se tornam mais desafiadores, a sociedade passa a produzir e consumir conjuntos de dados mais complexos, detalhados e com múltiplas dimensões.

O desafio de lidar com grandes volumes de dados não é uma simples questão de processar uma maior quantidade dados, maior necessidade de espaço de armazenamento e de meios de comunicação otimizados para transmitir os fluxos de dados em alta velocidade, embora todos estes elementos sejam parte do problema de *Big Data*. Outras questões também necessitam ser resolvidas, tais como (WONG et al., 2012):

- Como treinar pessoas para lidar com as características de *Big Data*?;
- Quais são as ferramentas necessárias para movimentar, armazenar e extrair conhecimento em grandes volumes de dados?;
- Como saber se os dados acessados em diversos repositórios são íntegros e como determinar seu nível de acurácia?;
- Especialistas na área de computação de alto desempenho (*High Performance Computing* – HPC) predizem que os componentes mais comuns destes sistemas – tais

como poder computacional, memória, armazenamento, largura de banda – terão seus desempenhos melhorados por um fator de 3 até 4 mil vezes ao longo de uma década (WONG et al., 2012), enquanto que a capacidade de cognição humana prevalecerá constante. Sob este aspecto, como encontrar mecanismos para compensar as 'fraquezas' da cognição humana?.

Um dos objetivos da União Europeia para o desenvolvimento de uma infraestrutura para a Ciência dos Dados a ser alcançado até 2030 (DYER; HAVER; NOWLAN, 2012), estabelece que pesquisadores de qualquer área de pesquisa serão capazes de encontrar, acessar e processar os dados que necessitam, com habilidades para usar e compreender tais dados, assim como para mensurar o grau no qual estes dados podem ser confiáveis.

Para alcançar este objetivo, a comunidade científica deverá responder diversas questões, não restritas apenas à tecnologia da informação. Uma das questões que merece destaque pode ser estabelecida como: "De qual maneira os usuários das mais diversas áreas do conhecimento compreenderão os dados que estão acessando e reconhecerão as circunstâncias pelas quais eles foram gerados?".

Uma das boas práticas estabelecidas no método científico consiste na avaliação e no reconhecimento dos erros inerentes a um processo de medição ou experimento. Qualquer relatório técnico, programa de monitoramento ou pesquisa científica deve incluir uma avaliação sobre a incerteza de seus resultados (SACHA et al., 2016).

Para o pesquisador David Lawson (LAWSON, 2014), quanto mais fontes de dados são acessadas, maior é a variedade dos tipos de dados disponíveis, assim como maior é a probabilidade de se encontrar erros nestes dados.

Na linha de raciocínio de Lawson, um conjunto de dados menor e com menor variedade de tipos de dados apresenta uma menor probabilidade de inconsistências nos dados. Todavia, esta é uma realidade nem sempre factível para diversos projetos de pesquisa. O contrário também é impraticável. Aceitar que só se pode trabalhar com conjuntos de dados grandes e com maior acurácia é algo que nem sempre é possível de se obter.

Diante deste cenário, a Ciência necessita "abraçar a dúvida" (*em inglês, Embrace the Doubt*). O método científico deve incorporar atividades que permitam disseminar as incertezas e dúvidas geradas, para que a comunidade científica tome suas decisões sabendo de sua existência. Lawson afirma:

> "Os dados nunca foram perfeitos e a certeza é uma ilusão temporária. *Big Data* não criou as imperfeições ou as incertezas. Ela apenas fez uma

> luz tecnológica apontar para estes problemas. Ao invés de fugir do que agora podemos ver, vamos procurar aproveitar o seu poder para nos dar o conhecimento para tomar melhores decisões e tomar as medidas mais inteligentes." (LAWSON, 2014, tradução nossa).

Pode-se agrupar aspectos como grau de confiança em um conjunto de dados, vieses de coleta, autenticidade, rastreabilidade, etc. em uma dimensão para *Big Data*, a qual denomina-se **Veracidade**. Todavia, não há um consenso na comunidade científica sobre a definição desta nova dimensão para *Big Data*.

Diante do exposto, a seguir apresenta-se uma proposta para a conceituação de Veracidade em Ciência dos Dados, como um dos objetivos específicos para a execução deste trabalho.

3.1 O conceito de Veracidade dos Dados

Em busca de uma definição sobre Veracidade, sob o contexto da Ciência dos Dados, recorre-se a uma análise de sua etimologia e de seu significado.

Na língua inglesa o termo veracidade (*Veracity*) surge em meados do século XVII, derivado do adjetivo em latim *verax*, o qual significa *verdadeiro* (MERRIAM-WEBSTER, 2017). Com relação a seu significado, as definições mais encontradas são:

1. Aderência à verdade; conformidade a um fato; acurácia (*American Heritage Dictionary*) (HERITAGE, 2016);
2. Precisão; verdade (*Collins English Dictionary*) (COLLINS, 2017);
3. Livre de falsidades; observância habitual da verdade; despretensão de dizer mentiras (*The American Heritage Roget's Thesaurus*) (ROGET, 2017).

A Tabela 2 sintetiza as definições encontradas para o termo veracidade na literatura.

Pela análise da Tabela 2, alguns pontos são de relevante destaque:

- Para Crone (2016) a veracidade é uma característica ligada à qualidade de dados, mais especificamente a adequabilidade de um conjunto de dados para um processo de tomada de decisão. O autor, todavia, não apresenta elementos práticos para o uso da veracidade nas atividades de análise de dados;

Tabela 2: Principais definições de Veracidade encontradas na literatura.

Referência	**Definição**
Crone (2016)	Habilidade de um conjunto de dados suportar processos de decisão de uma maneira apropriada, útil e com qualidade suficiente para o contexto sob o qual está sendo utilizado.
Duncan (2016)	Habilidade de verificar se os dados estão adequados para o propósito da pesquisa e habilitados para uso em um modelo analítico. Decidir se os dados podem ser testados sob um conjunto de critérios de modo a estabelecer o grau de confiança e de verdade existentes nos dados.
Marr (2016)	Refere-se a desordem ou a confiabilidade dos dados.
Owais e Hussein (2016)	Veracidade refere-se aos ruídos, vieses e anormalidades em um conjunto de dados. Busca responder a pergunta: "Os dados em questão estão sendo analisados com o objetivo de verificar se são adequados para o problema de pesquisa?".
Stantic (2015)	Veracidade relaciona-se à confiança dos dados e a predição de dados inerentemente imprecisos, com ruídos ou anormalidades,que sejam significativos ao problema em análise. As dimensões de Variedade e Velocidade trabalham contra a veracidade dos dados, uma vez que diminuem a habilidade de analisar e limpar os dados antes de se extrair informações.
Ballard et al. (2014)	Veracidade refere-se a três características: - A qualidade ou consistência/acurácia dos dados; - A proveniência ou fonte dos dados; - A descrição da intenção de uso dos dados, o qual pode afetar drasticamente o modo de como os dados podem ser utilizados e/ou serem considerados confiáveis.
Theoret (2009)	A veracidade dos dados é afetada por elementos tais como: - Acurácia dos dados (proximidade da medida relativamente ao verdadeiro valor da variável); - Precisão (se verdade, qual a margem de erro?); - Confiabilidade (isto ocorre sempre?); - Proveniência (é possível rastrear a origem dos dados?); - Integridade (Os dados foram alterados a partir da origem?) - Permissão (Pode-se fazer uso dos dados neste contexto?).

- Para Duncan (2016) e Owais e Hussein (2016), a veracidade está ligada à habilidade de se verificar se os dados estão adequados ao propósito da pesquisa, assim como habilitados para uso em um modelo analítico. Esta abordagem é normalmente apresentada pelo conceito de *fitness for use* (PÔÇAS et al., 2014), o qual busca mensurar a adequabilidade de conjuntos de dados as necessidades de aplicações dos

usuários. Ou seja, os dados são avaliados em função da aplicação para os quais serão utilizados e não apenas de um modo genérico (fazendo uso de métricas estatísticas descritivas, por exemplo);

- Para Stantic (2015) e Theoret (2009), a veracidade está relacionada a elementos de qualidade de dados, tais como acurácia, precisão, confiabilidade, integridade. Não é incomum esbarrar no conceito de qualidade de dados quando da busca pelo conceito de veracidade. Todavia, é importante ressaltar que embora um processo de "preparação de dados" (e.g., remoção de duplicidades, eliminação de valores nulos, normalização de dados, agrupamento de dados, etc.), essencial para tarefas de qualidade dos dados, seja executado corretamente, do que adianta se o dado em si é falso? Ou tendencioso? Ou ainda, não confiável (qual a procedência do dados?, por exemplo).

- Theoret (2009) e Ballard et al. (2014) citam como elementos de veracidade a proveniência dos dados, juntamente com elementos de descrição (*metadados*), que permitam a compreensão de um conjunto de dados e a análise de diversos aspectos como, por exemplo, as permissões de uso de um conjunto de dados. Para estes autores, a descrição de um conjunto de dados e a sua semântica são elementos preponderantes para a veracidade.

Esteves et al. (2018) associam o conceito de veracidade com o termo *trustworthiness*. A tradução mais adequada para este termo, a qual será considerada nesta tese, é a de confiabilidade (MOHAMMADI et al., 2014). *Trustworthiness* é uma característica de um sistema que possui um potencial para influenciar a confiança de um usuário no sistema, de uma maneira positiva ou negativa. Mohammadi et al. (2014) desenvolveram uma taxonomia de elementos que compõem a confiabilidade. Dentre os principais elementos da taxonomia, ressaltam-se: transparência / prestação de contas (*Accountability*), rastreabilidade (*Traceability*), confidencialidade (*Confidentiality*), integridade (*Integrity*), não-repúdio (*Non-repudiation*), dados abertos (*Openess*), completude (*Completeness*) e disponibilidade (*Availability*).

Ainda sobre a taxonomia proposta por Mohammadi et al. (2014), os autores defendem o fato de que para um sistema ser considerado confiável (em inglês, *trustworthy*), este sistema deve satisfazer a uma série de dimensões da qualidade, ressaltando que a confiabilidade é uma dimensão dependente do domínio e da aplicação do sistema. Dessa forma, faz-se necessário adicionar à taxonomia dimensões que irão representar características dependentes do contexto, tais como: satisfação, aprendizagem, efetividade, eficácia,

dentre outras.

Corroborando ao trabalho de Mohammadi et al. (2014), Scannapieco e Berti (2016) definem que a confiabilidade de um sistema pode ser caracterizada com base em três dimensões: reputação (o quanto a informação é considerada em termos de sua fonte ou conteúdo), credibilidade (o quanto a informação é considerada como verdadeira e verossímil) e acurácia (o quanto a informação é correta e confiável).

Embora a raiz da palavra Veracidade seja *verax* (verdade), a veracidade está mais ligada à transparência. A Ciência é uma organização social: grupos de colaboração independentes trabalham para gerar conhecimento como um bem público. A credibilidade dos trabalhos científicos está enraizada nas evidências que os suportam, as quais incluem a metodologia aplicada, os dados adquiridos e os processos para execução dos experimentos, da análise de dados e da interpretação dos resultados obtidos. Os trabalhos científicos tornam-se credíveis pela comunidade científica por meio da revisão, crítica, extensão e reprodução dos trabalhos desenvolvidos. Sem a transparência em todo este processo, os trabalhos científicos só serão credíveis com base na confiança ou na autoridade de seus pesquisadores. A transparência no método científico deve ser superior ao julgamento efetuado com base na confiança ou autoridade de um pesquisador (MUNAFÒ et al., 2017).

A veracidade dos dados não é obtida apenas por meio de uma técnica como, por exemplo, a análise estatística de um conjunto de dados, ou a execução de algoritmos de aprendizagem de máquina. Métodos estatísticos utilizados isoladamente podem não fornecer informações suficientes para a determinação da confiabilidade de conjunto de dados, assim como algoritmos de aprendizagem de máquina que, quando utilizados isoladamente, normalmente ofuscam a interpretação de um conjunto de dados. A visualização de dados, quando único elemento de análise, pode conduzir erroneamente a conclusões inverídicas. A Ciência dos Dados se faz com a junção destes três elementos, uma vez que tal junção permite aos cientistas explorar e raciocinar sobre um conjunto de dados das mais variadas maneiras (CHEN et al., 2011).

Técnicas estatísticas juntamente com técnicas de aprendizagem de máquina e de visualização de dados permitem à comunidade científica a comunicação de erros e incertezas nos dados (*credibilidade*), a comunicação das decisões na construção dos modelos e a descrição de seus parâmetros (*legitimidade*), permitindo aos usuários raciocinar e analisar um conjunto de dados em função de suas necessidades de pesquisa (*relevância*) (ROSS; DOUGLAS, 2016).

Com base na literatura pesquisada, acredita-se que a Veracidade dos dados é uma

característica dos dados e dos processos que os geram e que irá refletir na confiança de um pesquisador sobre um conjunto de dados. Propõe-se, portanto, uma definição para a Veracidade dos Dados, qual seja:

Veracidade é a dimensão que reflete a legitimidade, a transparência e a adequabilidade de um conjunto de dados para uma comunidade científica. A legitimidade busca evidenciar as decisões tomadas ao longo das atividades que produziram o conjunto de dados. A transparência reflete o acesso ao conjunto de dados e identifica partes e demais processos envolvidos na criação, publicação e reuso do conjunto de dados. A adequabilidade é obtida por meio da avaliação do conjunto de dados pela comunidade científica.

Corrobora à definição proposta para a Veracidade o trabalho de Reimer e Madigan (2018). Nele, os autores propõem que a veracidade é mensurada a partir de três elementos: a proveniência dos dados, a validação cruzada com outras fontes de dados e o contexto sob o qual os dados estão sendo utilizados. Comparando-se ao conceito proposto nesta tese, a legitimidade busca evidenciar as decisões tomadas ao longo das atividades que produziram o conjunto de dados, endereçando, portanto, a problemática da proveniência, a qual se torna viável em função da transparência de dados. A adequabilidade busca avaliar o conjunto de dados pela comunidade científica, atuando, portanto, como uma forma de validação cruzada com base no contexto no qual os dados estão sendo inseridos. Enquanto no trabalho de Reimer e Madigan (2018) a validação se dá por meio da comparação do conjunto de dados em múltiplos repositórios, a proposta aqui apresentada prima pela avaliação colaborativa dos pares como elemento de validação da adequabilidade de um conjunto de dados.

À habilidade do entendimento de um conjunto de dados e dos processos de transformação aplicados durante a análise dos dados no que concerne a incertezas, vieses, adequabilidade dos dados, auxiliadas pelo uso de descrições semânticas e de metadados, dá-se o nome de **Compreensão dos Dados** (*Data Understandability*) (BATISTA; CORRêA; PALANISAMY, 2016). A inserção da dimensão de Veracidade a um processo de gestão e análise de dados é o elemento que facilita o processo de compreensão dos dados pela comunidade científica.

As seções a seguir apresentam o desdobramento do conceito de Veracidade, de modo a explorar os aspectos da legitimidade, da transparência e da adequabilidade. A legitimidade e a transparência serão desenvolvidas por meio de um perfil de aplicação que irá atuar na descrição de um conjunto de dados científico, enquanto que a adequabilidade

será obtida por meio da avaliação do conjunto de dados pela comunidade científica.

3.2 Modelo de Avaliação de Conjuntos de Dados Científicos

A Ciência como atividade humana e social tem produzido importantes contribuições, gerando inovação e permitindo que a sociedade evolua em todos os seus setores. A rápida evolução da tecnologia nas últimas décadas impactou e mudou a forma como a Ciência é feita.

A Ciência contemporânea está definida em um novo paradigma científico chamado *"O Quarto Paradigma"* (HEY et al., 2009), o qual promove o surgimento de uma metodologia de pesquisa denominada *e-Science*. Nesse panorama, a Ciência é essencialmente aberta e inovadora, com base na possibilidade de compartilhar dados abertos para uso e reutilização.

O reuso possibilita dar continuidade a uma determinada pesquisa, gerar novas hipóteses, inferências, direcionamentos e obter novas descobertas a partir dos dados primários de pesquisa de outros pesquisadores. Esse processo de reuso é possível quando ocorre a efetiva colaboração científica (OLIVEIRA, 2016).

O termo *e-Science* é definido como a Ciência da Computação intensiva, realizada em ambientes de rede altamente distribuídos que utilizam volumosos conjuntos de dados, os quais permitem a coleta, o processamento, a preservação, a análise e o armazenamento de grandes quantidades de dados em diferentes disciplinas. Envolve um espectro vasto em relação à comunicação acadêmica, curadoria de dados, análise, representação, visualização e colaboração entre pares.

A *e-Science* é uma resposta a um fenômeno chamado "dilúvio de dados" (SELTZER; ZHANG, 2009), no qual os dados são o elemento central e o principal recurso de pesquisa. Todos esses conceitos expressam a realidade atual, onde a quantidade de dados é superior ao limite de análise e extração de conhecimento pelos seres humanos ou pelas máquinas.

Na *e-Science* "as expectativas em torno de um mundo rico em dados são imensas e incluem desde descobertas de novas drogas, passando por um melhor entendimento sobre as mudanças climáticas, até metodologias mais apuradas para examinar a história e a cultura" (BORGMAN, 2010, pg. 2, tradução nossa).

As práticas científicas da *e-Science* trazem na sua essência a mudança no *modus ope-*

randi de fazer, compartilhar, divulgar, acessar, usar, reusar e reproduzir dados de pesquisa com possibilidades de compartilhamento, interatividade, convergência e colaboração entre pesquisadores (OLIVEIRA, 2016). Assim, sua natureza colaborativa e multidisciplinar corrobora o novo fazer científico por meio do gerenciamento de dados de pesquisa.

A prática científica pressupõe uma infraestrutura que envolva os aspectos políticos, institucionais, gerenciais, metodológicos e tecnológicos. No cenário internacional, os esforços para uma Ciência orientada aos dados de pesquisa são apoiados por políticas, diretrizes, modelos, tecnologias e boas práticas. Sob o ponto de vista tecnológico, a *e-Science* é apoiada por uma infraestrutura tecnológica denominada ciberinfraestrutura (OLIVEIRA, 2016). A ciberinfraestrutura que suporta o ambiente da pesquisa científica aberta é um dos componentes de sustentabilidade da *e-Science*. Utiliza alta tecnologia, computação intensiva, ferramentas e recursos para lidar com dados. Tal ciberinfraestrutura busca apoiar a captura, armazenagem, gerenciamento, preservação, visualização e análise de dados de pesquisa.

A transição de práticas científicas do modelo tradicional para a *e-Science* está apoiada em três aspectos (BORGMAN, 2015a): o primeiro é a adoção das novas práticas a partir de um modelo de ciclo de vida dos dados que garanta o efetivo gerenciamento dos dados primários; o segundo é proporcionar uma ciberinfraestrutura sustentável para uso, reuso e reprodutibilidade dos dados de pesquisa; e o terceiro são aspectos relacionados ao compartilhamento e colaboração científica entre diferentes pesquisadores, profissionais, instituições e até países.

O modelo proposto nesta tese busca proporcionar um mecanismo de avaliação de conjuntos de dados científicos publicados em uma ciberinfraestrutura de portais de dados científicos. Por meio deste modelo, conjuntos de dados publicados em portais de dados serão avaliados pela comunidade científica, permitindo assim a geração do índice de veracidade para conjuntos de dados. Tal índice será utilizado como métrica para mensurar a compreensão sobre conjunto de dados e dos processos que o geraram, subsidiando, assim, o uso e o reuso de tal conjunto de dados.

O modelo de avaliação apoia os três aspectos da *e-Science* definidos por Borgman (2015a), da seguinte maneira:

- A adoção das novas práticas a partir de um modelo de ciclo de vida é feita pela proposta de um perfil de aplicação baseado nos requisitos estabelecidos no modelo de gestão de dados da USGS e da *Data Science Association*;

- A ciberinfraestrutura sustentável para uso, reuso e reprodutibilidade dos dados de pesquisa é contemplada quando do uso de formatos de dados interoperáveis e interpretáveis por máquina, que possibilitam uma fácil implementação em portais de dados já existentes como, por exemplo, os portais de dados da USGS e a ferramenta *ARM Archive - Data Discovery Tool*, utilizada pelo Departamento de Energia do Governo Americano (BATISTA; CORRêA; PALANISAMY, 2016);

- O último aspecto, sobre o compartilhamento e a colaboração científica, é satisfeito por meio do índice de veracidade para conjuntos de dados, o qual é composto por métricas calculadas a partir do perfil de aplicação, juntamente com a avaliação da comunidade científica sobre a efetividade da comunicação proporcionada pelo uso do perfil de aplicação, na busca por avaliar a compreensão do conjunto de dados e dos processos de transformação relacionados.

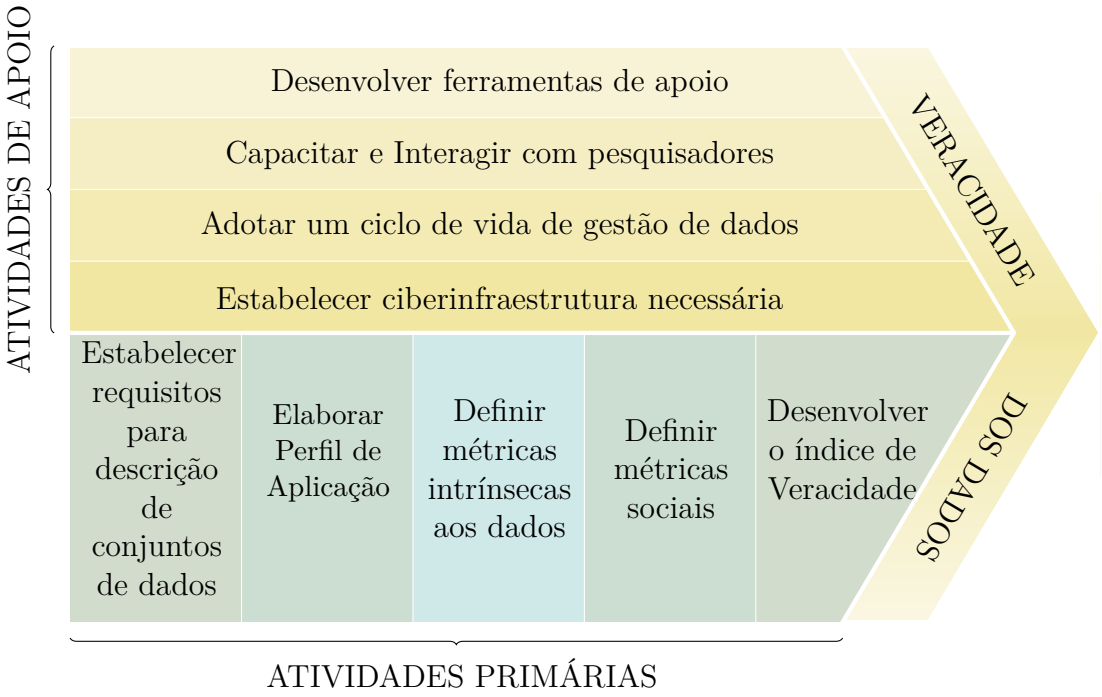
Os autores Sayão e Sales (2012) afirmam que a comunidade científica anseia por novos modelos de publicações que possam expressar a complexidade e a dinâmica da pesquisa científica contemporânea. Os modelos científicos da *e-Science* apresentam possibilidades que explicitam mais claramente a pesquisa desenvolvida, seus métodos, seu material, os dados e conhecimentos gerados, além de revelar aspectos intrínsecos da pesquisa científica, a exemplo da intenção do pesquisador e a avaliação dos resultados de forma mais eficaz e interativa.

Corrobora a ideia de mecanismos que permitam o compartilhamento de diversas avaliações sobre conjuntos de dados, o trabalho de Tenopir e King (2004), no qual os autores afirmam que "a *inovação nunca acontece em um vácuo; inovação requer comunicação e colaboração*". Os autores sugerem ainda que todos os estágios de pesquisa, desenvolvimento, desenho, e produção de dados e resultados científicos sejam o mais eficientes e efetivos possíveis, para que a informação possa ser clara e entendida.

Dessa forma, o modelo aqui apresentado propõe o uso do perfil de aplicação desenvolvido para a veracidade, como elemento necessário para a comunicação sobre o conjunto de dados e seus processos de transformação relacionados. Sob este perfil, métricas de apoio serão calculadas e, em conjunto com a percepção da comunidade científica, será possível estabelecer o índice de veracidade do conjunto de dados para uma determinada área de pesquisa ou de aplicação.

A Figura 11, inspirada no modelo de cadeia de valor de Michael Porter (PORTER; MILLAR, 1985), apresenta um diagrama dos elementos necessários para a construção do modelo. O diagrama se encontra estruturado em duas partes principais: atividades

Figura 11: Visão geral da estrutura do modelo de avaliação de conjuntos de dados científicos.



Fonte: o Autor.

primárias e atividades de apoio. As atividades primárias são as tarefas principais do processo de construção do modelo de avaliação de conjuntos de dados, sendo suportadas pelas atividades de apoio.

Dessa forma, ao longo do desenvolvimento deste modelo, buscou-se executar as seguintes atividades primárias:

- Estabelecer requisitos para descrição de conjuntos de dados: ao longo do desenvolvimento do modelo, buscou-se estabelecer um conjunto de requisitos para a descrição de conjuntos de dados científicos, que permitam uma maior compreensão do conjunto de dados e dos processos de transformação sob os quais o conjunto foi submetido ao longo de um experimento. Esta atividade é apresentada em detalhes na Seção 3.2.1;

- Elaborar perfil de aplicação: os requisitos elicitados na atividade anterior são consolidados em um perfil de aplicação, o qual é apresentado na Seção 3.2.1;

- Definir métricas intrínsecas aos dados: com base no perfil de aplicação, alguns elementos de descrição são utilizados como insumos para o cálculo de métricas que irão refletir aspectos intrínsecos ao conjunto de dados. Tais métricas são apresentadas

na Seção 3.2.2;

- Definir métricas sociais: a avaliação de um conjunto de dados pela comunidade científica se dá como uma perspectiva social do modelo de avaliação de conjuntos de dados científicos. Tal perspectiva é apresentada na Seção 3.2.3, juntamente com estratégias para a determinação de reputação de avaliadores em uma rede social de avaliação, assim como o estabelecimento de um mecanismo de decaimento temporal de tais avaliações;

- Desenvolver o índice de Veracidade: uma vez que as atividades anteriores foram concluídas, passou-se a executar a atividade de definição do índice de veracidade de conjuntos de dados científicos, o qual será calculado sob duas perspectivas: a perspectiva relacionada ao conjunto de dados e a perspectiva social. O índice é apresentado em detalhes na Seção 3.2.4.

Com relação às atividades de apoio, quatro foram elencadas, quais sejam:

- Estabelecer ciberinfraestrutura necessária: ao longo do desenvolvimento do modelo, buscou-se elencar quais elementos tecnológicos compõem a ciberinfraestrutura necessária para a publicação de conjuntos de dados científicos. Estabeleceu-se como referência de ciberinfraestrutura os portais de dados que fazem uso de formatos de dados *linked data*, compreensíveis por máquina;

- Adotar um ciclo de vida de gestão de dados: novas atividades são incorporadas ao processo de gestão de dados, de modo a suportar o conceito de veracidade. Para isto, optou-se por adotar o ciclo de gestão de dados proposto pela USGS como referência para a incorporação das novas tarefas relacionadas à veracidade. A Seção 3.2.5 apresenta uma proposta de incorporação do índice de veracidade à portais de dados científicos considerando o modelo de gestão de dados da USGS;

- Capacitar e interagir com pesquisadores: a dimensão de veracidade ainda é pouco explorada na comunidade científica. No Brasil, há ainda um *défict* na formação na área de gestão e curadoria de dados científicos. Dessa forma, como subsídio para estabelecer o modelo de avaliação de conjuntos de dados, buscou-se promover eventos de capacitação e de interação com a comunidade científica. Ao longo do desenvolvimento da tese, quatro eventos foram realizados, denominados *Workshops em Ciência dos Dados* (WDS)[1]. Os *workshops* foram promovidos pela Escola Poli-

[1]Maiores informações sobre os eventos estão disponíveis no endereço eletrônico <http://wds.poli.usp.br>.

técnica da Universidade de São Paulo nos anos de 2015 (*I WDS*), 2017 (*II WDS e III WDS*) e 2018 (IV WDS);

- Ferramentas de apoio: ao longo do desenvolvimento do modelo, diversas ferramentas foram construídas com o objetivo de facilitar o processo de descrição de conjuntos de dados científicos. Tais ferramentas serão apresentadas no Capítulo 4.

3.2.1 Perfil de Aplicação

Esta Seção apresenta os requisitos de descrição para conjuntos de dados de modo a suportar o conceito proposto para a dimensão de Veracidade dos Dados. Tais requisitos são consolidados em um perfil de aplicação, apresentado a seguir.

O cenário de *Big Data* apresenta diversas oportunidades para o desenvolvimento de pesquisas científicas que podem fazer uso de conjuntos de dados em larga escala, permitindo a busca por novas descobertas em uma miríade de áreas do conhecimento.

Embora o cenário seja promissor, a dificuldade em compreender um conjunto de dados cresce proporcionalmente à disponibilidade de recursos de dados para análise. Um dos problemas que afetam a compreensão de um conjunto de dados é o nível de descrição sobre o conjunto de dados, suas origens e as transformações as quais o conjunto foi submetido. De fato, há na literatura um apelo significativo para o uso de metadados como elementos de apoio semântico à compreensão humana, assim como pelo fato de que, quando disponíveis em um formato compreensível por máquina, tais metadados possibilitarão que algoritmos sejam desenvolvidos para auxiliar na busca, descoberta e avaliação de um conjunto de dados com base em uma necessidade específica. A qualidade na descrição dos metadados e a adoção de padrões irão contribuir para o futuro compartilhamento dos dados para uso e reuso.

Diversos conjuntos de dados estão sendo publicados em plataformas digitais denominadas *portais de dados*. Um portal de dados permite a busca e recuperação de conjuntos de dados elaborados por diversos pesquisadores, agrupados por temas.

Para o desenvolvimento da dimensão de Veracidade, esta tese propõe um perfil de aplicação para a descrição de um conjunto de dados para uso por um portal de dados, o qual irá disponibilizar diversos conjuntos de dados agrupados por área de conhecimento/aplicação.

Para a elaboração do perfil de aplicação proposto, buscou-se seguir os procedimentos metodológicos descritos em Nilsson, Baker e Johnston (2008). Inicialmente estabeleceu-se

os requisitos funcionais do perfil de aplicação. Uma vez que os requisitos estão consolidados, elaborou-se o modelo de domínio e o perfil de descrição. Estes passos são apresentados nas seções a seguir.

3.2.1.1 Requisitos Funcionais

O objetivo do perfil de aplicação para suporte à Veracidade é a definição de um conjunto de metadados que possam descrever, de forma objetiva, os conjuntos de dados e demais recursos associados, os quais serão publicados em um portal de dados, permitindo a busca e recuperação dos conjuntos de dados de interesse de um consumidor de dados.

Não é objetivo deste perfil de aplicação ser um substituto dos esquemas de metadados disponíveis na literatura, tais como o *Dublin Core* (WEIBEL et al., 1998) e o FGDC (COMMITTEE et al., 1998). Este perfil de aplicação busca, por sua vez, reunir os elementos dos metadados já existentes em um conjunto que possibilite atender a um grupo de requisitos para compreensão de conjuntos de dados.

Sobre o escopo dos requisitos para a compreensão de conjuntos de dados, fez-se uso dos requisitos estabelecidos no modelo de gestão de dados do Serviço Geológico dos Estados Unidos (*United States Geological Survey* - USGS) (FAUNDEEN et al., 2014).

A escolha pelos requisitos da USGS deu-se por duas razões principais. A primeira é proveniente do fato da USGS ser reconhecidamente uma instituição que busca estabelecer padrões de metadados para a gestão de dados científicos. A segunda é decorrente da colaboração em pesquisas entre o grupo de estudos, pesquisa e extensão em Big Data da Escola Politécnica da Universidade de São Paulo com a USGS, na figura do pesquisador Prof. *Mike Frame*, o qual atualmente exerce a função de vice-diretor da USGS.

No ciclo de vida de gestão de dados propostos pela USGS, a descrição de um conjunto de dados deve satisfazer aos seguintes itens (USGS, 2015):

1. Incluir uma descrição sobre o conjunto de dados;
2. Identificar quem é o criador/proprietário dos dados, indicando nome e dados de contato;
3. Identificar crédito a outras pessoas ou organizações pelo financiamento ou suporte ao processo de coleta e análise de dados;
4. Informar o período de tempo representado pelo conjunto de dados;

5. Informar, se houver, a cobertura espacial do conjunto de dados;
6. Indicar o plano de atualização do conjunto de dados;
7. Identificar os formatos nos quais o conjunto de dados é disponibilizado;
8. Listar as palavras-chave sobre o conjunto de dados;
9. Identificar se há elementos que fazem uso de um vocabulário controlado;
10. Listar os processos utilizados para criar o conjunto de dados, identificando todas as fontes utilizadas;
11. Identificar eventuais restrições legais sobre uso e reuso dos dados;
12. Especificar como o conjunto de dados se encontra disponível para outros pesquisadores, fornecendo uma URL;
13. Prover uma descrição breve sobre cada atributo que compõe o conjunto de dados, juntamente com as suas unidades de medida, se aplicável.

Para a determinação dos elementos de metadados que compõem o perfil proposto, atendendo aos requisitos elicitados pela USGS, buscou-se fazer uso do modelo 5W1H (*Why, Who, What, When, Where, How*) adaptado por Oliveira (2016) para o estabelecimento e validação de metadados voltados para a descrição de conjuntos de dados. Este modelo é apresentado na Tabela 3.

3.2.1.2 Modelo de Domínio

Após a definição dos requisitos funcionais do perfil de aplicação proposto, criou-se o modelo de domínio. O modelo de domínio é a descrição de quais elementos o perfil de aplicação irá descrever e os relacionamentos entre estes elementos (COYLE et al., 2009).

A Figura 12 apresenta o modelo de domínio para o perfil de aplicação proposto. Neste modelo, um **catálogo** representa uma coletânea de conjuntos de dados de sobre um determinado tema. O modelo de catálogo de dados para o agrupamento de conjuntos de dados foi inspirado na proposta do *Federal Geographic Data Committee* (FGDC) do governo dos Estados Unidos, para a estruturação de conjuntos de dados sob um conjunto de temas (FGDC, 2008). A FGDC mantém um conjunto de temas para o agrupamento de conjunto de dados, e este conjunto é compartilhado com a comunidade científica que faz uso de dados geográficos. Por meio da definição dos temas, a FGDC instruí aos

Tabela 3: Requisitos 5W1H para a descrição de dados científicos.

Perguntas	**Orientações**
Why (Porquê)	Refere-se ao entendimento dos dados e o contexto que eles foram criados. Os detalhes devem incluir as questões científicas ou proposições para a coleção de dados, razões que envolvem contribuidores voluntários e uma visão geral sobre o conjunto de dados.
Who (Quem)	Trata-se da descrição dos metadados relacionados a quem coletou os dados ou a quem patrocinou a coleta. Deve especificar quem deverá ser contactado com perguntas sobre os dados e como deve ser citado o conjunto de dados para garantir que todos os envolvidos obtenham reconhecimento na pesquisa.
What (O quê)	Requer uma descrição apurada dos dados. Foram desenvolvidas algumas categorias que ajudam a nortear a descrição: contexto digital (nomes de arquivos, formatos de arquivos, modificações de datas, arquivos de exemplo, conjuntos de dados relacionados e procedimentos de processamento de dados); Parâmetro/detalhes variáveis (unidades de medida, formatos de dados, precisão e exatidão); Informações sobre dados (taxonomias, codificação, procedimentos); Conteúdo de arquivos de dados (definições de parâmetros e explicação de seus formatos, avaliação de qualidade).
When (Quando)	Relaciona-se ao detalhamento dos aspectos temporais, como período (ano, mês, dia), tempo, regularidade e intervalos da coleta de campo.
Where (Onde)	Direciona-se ao detalhamento dos aspectos de cobertura geográfica e espacial que estão relacionadas com o país, a cidade, local e coordenadas espaciais.
How (Como)	Estabelece as condições de criação dos dados. O detalhamento de protocolos, técnicas, métodos, diagramas, software, modelos, esquemas e estatísticas utilizadas nos procedimentos metodológicos e de coleta de campo que originou os dados são relevantes para uma descrição qualitativa dos metadados. Documentos adicionais aos dados também devem ser informados e se possível vinculados ao dado ou conjunto de dados.

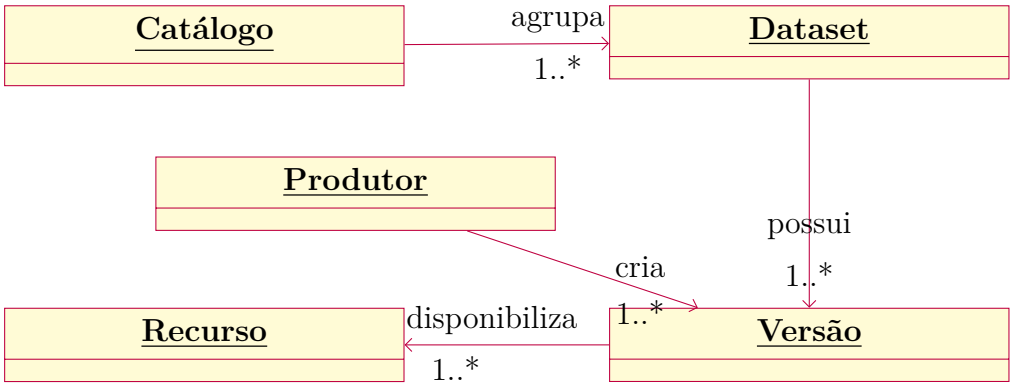
Fonte: (OLIVEIRA, 2016).

especialistas que definam seus conjuntos de dados inserindo-os sob um determinado tema, facilitando assim a catalogação e a busca por conjuntos de dados.

Um **catálogo** agrupa múltiplos ***datasets***. Define-se como ***dataset*** um conjunto de dados disponível para acesso ou *download* em um ou mais formatos de arquivo, disponibilizado em um portal de dados. Um ***dataset*** apresenta uma ou mais **versões**. Conforme exposto por Dumontier et al. (2016), diversos portais de dados carecem de manter todas as versões de um conjunto de dados, e este fato inviabiliza a recuperação de uma versão específica de um conjunto de dados utilizado em um experimento. Busca-se, portanto, fomentar o reuso dos conjuntos de dados por meio do armazenamento e garantia da disponibilidade das múltiplas versões de um conjunto de dados.

Cada **versão** é constituída por diversos **recursos** ou **distribuições**. Um **recurso** é

Figura 12: Modelo de Domínio para o Perfil de Aplicação proposto.



Fonte: o Autor.

o nome dado para a forma de possibilitar o acesso aos dados existentes em um dataset. O formato mais comum para um recurso é o de arquivos de dados, tais como arquivos nos formatos *csv* e *xls*.

Complementando o modelo de domínio, há ainda a figura do **produtor** de dados, o qual irá criar **versões** de um dataset. Nesta tese, os termos produtor e provedor de dados são utilizados como sinônimos. Em um ambiente de pesquisa colaborativa, há múltiplos produtores para um *dataset*, permitindo o desenvolvimento de múltiplas versões de um conjunto de dados, em que cada versão é criada diferentes produtores de dados.

3.2.1.3 Perfil de Descrição

Após o refinamento do modelo de domínio, buscou-se analisar quais vocabulários e esquemas de metadados podem ser utilizados para a descrição de um conjunto de dados e seus recursos associados.

Embora vários vocabulários sejam relevantes na descrição de conjuntos de dados, nenhum deles é suficiente para fornecer completamente a abrangência dos requisitos identificados na estruturação do perfil de aplicação. Os termos de metadados da *Dublin Core Metadata Initiative* (DCMI) (WEIBEL et al., 1998) oferecem um amplo conjunto de tipos e relações para a captura de metadados do documento. O vocabulário de catálogo de dados (DCAT) (DCAT-AP, 2015) é usado para descrever conjuntos de dados em catálogos, mas não aborda a questão da evolução do conjunto de dados e do controle de versão. A ontologia de proveniência (PROV) (LEBO et al., 2013) pode ser usada para capturar informações sobre entidades, atividades e pessoas envolvidas na produção ou modificação de dados. O vocabulário de conjuntos de dados interligados (VoID) (ALEXANDER et

al., 2009) é um vocabulário para expressar metadados descritos sobre o modelo *Resource Description Framework* (RDF).

Ao longo da escolha dos elementos de metadados que serão utilizados no perfil de aplicação, um conjunto de princípios de publicação eletrônica de conjuntos de dados, denominado FAIR (acrônimo para *Findable, Accessible, Interoperable, Reusable*), foi utilizado. Os princípios da iniciativa FAIR são (WILKINSON et al., 2016):

- Achável (*Findable*): Um objeto de dados deve ser identificável unicamente e de maneira persistente, permitindo a sua descoberta a qualquer tempo. Os dados são descritos por meio de um conjunto de elementos de metadados. Os metadados contém indicadores para o conjunto de dados que estão descrevendo.;
- Acessível (*Accessible*): Um objeto de dados deve ser sempre acessível por usuários e/ou máquinas, fazendo uso de formatos processáveis por máquina, vocabulários e ontologias que permitam a sua estruturação e compreensão. Além disso, a política de acesso do objeto deve estar explicitada, garantindo assim compreensão sobre os direitos de uso e reuso. Uma vez que os metadados estão separados do conjuntos de dados, é altamente recomendado que um metadado esteja sempre disponível e acessível, mesmo quando a licença do conjunto de dados não permite o livre acesso a seu conteúdo;
- Interoperável (*Interoperable*): metadados são descritos sob o formato RDF, fazendo uso de vocabulários bem definidos e consolidados (*Dublin core*, DCAT, etc.).
- Reutilizável (*Reusable*): metadados irão possuir um conjunto de elementos que fomentarão a reutilização do conjunto de dados, descrevendo questões de licenciamento, proveniência e conformidade com vocabulários e taxonomias de uma área específica.

Dessa forma, esta tese fez uso dos vocabulários, e seus respectivos termos, apresentados na Tabela 4.

Após a definição dos vocabulários a serem considerados no perfil de aplicação, buscou-se elencar quais elementos de metadados serão utilizados para satisfazer aos requisitos propostos. Os elementos foram escolhidos com base nos requisitos da USGS, fazendo-se uso da estruturação 5W1H proposta por Oliveira (2016).

A Tabela 5 apresenta a distribuição dos elementos de metadados propostas no perfil de aplicação nas dimensões do modelo 5W1H. Há de se observar que alguns elementos

Tabela 4: Vocabulários considerados na proposta do Perfil de Aplicação.

Vocabulário	**Prefixo**	**Referência**
Data Catalog Vocabulary	dcat	(DCAT-AP, 2015)
Dublin Core Terms	dct	(WEIBEL et al., 1998)
Federal Geographic Data Committee Metadata	fgdc	(COMMITTEE et al., 1998)
Provenance, Authoring and Versioning Vocabulary	pav	(CICCARESE et al., 2013)
Resource Description Framework	rdf	(CYGANIAK; WOOD; LANTHALER, 2014)
Vocabulary of Interlinked Datasets	void	(ALEXANDER et al., 2009)

Fonte: o Autor.

podem satisfazer os requisitos de mais de uma dimensão.

Tabela 5: Distribuição dos elementos de metadados com base no modelo 5W1H.

***Why* (Porquê)**	***Who* (Quem)**	**What (O quê)**	***When* (Quando)**	***Where* (Onde)**	***How* (Como)**
dct:description	pav:createdBy	dct:title	pav:createdOn	dct:spatial	fgdc:complete
dct:source	pav:providedBy	dct:keyword	dct:issued	fgdc:overview	dcat:provenance
dcat:provenance	dcat:contactPoint	dcat:theme	dct:temporal	fgdc:eaover	
fgdc:complete	pav:contributedBy	void:vocabulary	dct:accuralPeriodicity	fgdc:eainfo	
dct:license	foaf:page	dcat:distribution			
dct:accessRights	dct:language	dcat:downloadURL			
		dcat:conformsTo			
		dcat:byteSize			
		dct:format			
		dct:hasCurrentVersion			
		dct:isVersionOf			
		pav:previousVersion			
		pav:version			
		fgdc:eainfo			
		fgdc:eaover			
		fgdc:attrlabl			
		fgdc:attrtype			
		fgdc:attrdef			
		fgdc:atprecis			
		fgdc:attscale			

Fonte: o Autor.

Para a consolidação do perfil de aplicação, o conjunto de elementos foi expandido para o modelo de domínio, de modo a estabelecer quais elementos serão utilizados para a descrição de um conjunto de dados, de suas versões e de seus recursos associados. A Tabela 6 apresenta o perfil de descrição proposto. Os elementos foram classificados em recomendados e opcionais.

A classificação dos elementos em recomendados e opcionais é baseada na RFC 2119 (BRADNER, 1997), a qual conceitua os termos que devem ser utilizados na elaboração de uma especificação, quais sejam:

- Elementos Recomendados: o produtor dos dados deve fornecer as informações necessárias, sempre que estas estiverem disponíveis. Este é um requisito similar ao conceito de *should*, definido na RFC 2119, o qual indica que podem existir razões

válidas em circunstâncias particulares para ignorar um item específico, mas todas as implicações devem ser compreendidas e cuidadosamente ponderadas;

- Elementos Opcionais: o produtor de dados pode fornecer as informações necessárias, mas não é obrigado a fazê-lo, e a ausência de tais informações não irá prejudicar a compreensão pelo consumidor dos dados. Este é um requisito similar ao conceito de *may*, definido na RFC 2119, indicando que um produtor pode optar por incluir um item, em razão particular de uma área ou porque o produtor sente que isso melhora a descrição enquanto outro produtor pode omitir o mesmo item.

A classificação dos elementos de metadado foi feita com base na revisão da literatura, mais especificamente tomando-se por base o perfil de aplicação proposto pela União Europeia (DCAT-AP, 2015) e o perfil de aplicação colaborativo para Ciências da vida proposto por Dumontier et al. (2016). A consolidação do perfil de descrição foi feita nos eventos realizados pelo grupo de estudos, pesquisa e extensão em Big Data da Escola Politécnica da Universidade de São Paulo ao longo do desenvolvimento desta tese.

Tabela 6: Perfil de Descrição do Perfil de Aplicação para suporte à Veracidade.

			Objetos (Recomendado, Opcional)		
Requisito	**Identificador**	**Objetivo**	**Dataset**	**Versão**	**Recurso**
Incluir uma descrição sobre o conjunto de dados.	rdf:type	Identificar o tipo do objeto	rdf:type [dcat:dataset]	rdf:type [vera:datasetVersion]	rdf:type [dcat:distribution]
	dct:identifier	Identificar unicamente o objeto	dct:identifier [rdfs:literal]	dct:identifier [rdfs:literal]	dct:identifier [rdfs:literal]
	dct:title	Prover uma descrição textual curta sobre o objeto	dct:title [rdfs:literal]	dct:title [rdfs:literal]	dct:title [rdfs:literal]
	dct:description	Prover uma descrição textual sobre o objeto	dct:description [rdfs:literal]	dct:description [rdfs:literal]	dct:description [rdfs:literal]
	foaf:page	Informar uma página Web que provê maiores informações sobre o objeto	foaf:page [rdfs:resource]		
	dct:language	Especificar o idioma utilizado	dct:language [rfds:literal]	dct:language [rfds:literal]	dct:language [rfds:literal]
	dct:source	Identificar um recurso que originou o objeto, possibilitando obter maiores informações sobre a fonte dos dados	dct:source [rfds:resource]		
	pav:hasCurrentVersion	Identificar qual é a última versão do objeto	pav:hasCurrentVersion [rdfs:resource]		
	pav:createdOn	Especificar a data na qual o objeto foi criado	pav:createdOn [rfds:literal]	pav:createdOn [rfds:literal]	pav:createdOn [rfds:literal]
	dct:issued	Identificar a data na qual o objeto tornou-se disponível	dct:issued [rfds:literal]	dct:issued [rfds:literal]	dct:issued [rfds:literal]
Identificar quem é o criador/proprietário dos dados, indicando nome e dados de contato.	pav:createdBy	Especificar o(s) agente(s) responsável(eis) pela criação do objeto	pav:createdBy [rdfs:resource]	pav:createdBy [rdfs:resource]	pav:createdBy [rdfs:resource]
	pav:providedBy	Especificar o(s) agente(s) responsável(eis) pela publicação do objeto	pav:providedBy [rfds:resource]	pav:providedBy [rfds:resource]	pav:providedBy [rfds:resource]
	dcat:contactPoint	Especificar um ponto de contato para o objeto	dcat:contactPoint [rfds:Resource]	dcat:contactPoint [rfds:resource]	dcat:contactPoint [rfds:resource]
	dct:isVersionOf	Especificar o objeto original		dct:isVersionOf [rfds:resource]	
	pav:previousVersion	Identificar a versão anterior do objeto		pav:previousVersion [rfds:resource]	
	pav:version	Identificar o código da versão do objeto	pav:version [rfds:literal]		
Identificar crédito a outras pessoas ou organizações pelo financiamento ou suporte ao processo de coleta e análise de dados.	pav:contributedBy	Especificar o(s) agente(s) que contribuíram para a criação do objeto	pav:contributedBy [rdfs:resource]	pav:contributedBy [rdfs:resouce]	pav:contributedBy [rdfs:resouce]
Informar o período de tempo representado pelo conjunto de dados.	dct:temporal	Especificar a temporabilidade dos dados contidos no objeto	dct:temporal [rfds:resource]	dct:temporal [rfds:resource]	dct:temporal [rfds:resource]
Informar, se houver, a cobertura espacial do conjunto de dados.	dct:spatial	Delimitar espacialmente o objeto	dct:spatial [rdfs:resource]	dct:spatial [rdfs:resource]	dct:spatial [rdfs:resource]
Indicar o plano de atualização do conjunto de dados.	dct:accuralPeriodicity	Especificar a frequência de atualização do objeto		dct:accuralPeriodicty [rdfs:resource]	dct:accuralPeriodicty [rdfs:resource]
Identificar os formatos nos quais o conjunto de dados é disponibilizado.	dct:format	Identificar os formatos disponíveis			dct:format [rfds:literal]
Listar as palavras-chave sobre o conjunto de dados.	dcat:keyword	Identificar palavra(s)-chave sobre o objeto	dcat:keyword [rdfs:literal]	dcat:keyword [rdfs:literal]	dcat:keyword [rdfs:literal]
Identificar se há elementos que fazem uso de um vocabulário controlado.	dcat:theme	Identificar o tema associado ao objeto	dcat:theme [rdfs:resource]		
	void:vocabulary	Identificar o vocabulário utilizado para descrever/mapear o objeto	void:vocabulary	void:vocabulary	void:vocabulary
Listar os processos utilizados para criar o conjunto de dados.	fgdc:complete	Detalhar eventuais omissões, critérios de seleção, generalização e premissas	fgdc:complete [rdfs:literal]	fgdc:complete [rdfs:literal]	fgdc:complete [rdfs:literal]
	dcat:provenance	Descrição textual da proveniência do dado, ou machine-readable version	dcat:provenance	dcat:provenance	
Identificar eventuais restrições legais sobre uso e reuso dos dados.	dct:accessRights	Declarar os direitos de acesso e de uso do objeto	dct:accessRights [rdfs:literal]	dct:accessRights [rdfs:literal]	
	dct:license	Declarar a licença de acesso e de (re)uso do objeto	dct:license [rdfs:literal]	dct:license [rdfs:literal]	
Especificar como o conjunto de dados se encontra disponível para outros pesquisadores, fornecendo uma URL.	dcat:distribution	Especificar a distribuição de um objeto		dcat:distribution [rdfs:resource]	
	dcat:downloadURL	Especificar a URL para download do objeto			dcat:downloadURL [rdfs:literal]
	dcat:conformsTo	Especificar que um objeto é distribuição de outro			dcat:conformsTo [rdfs:resource]
	dcat:byteSize	Especificar o tamanho (em bytes) do objeto			dcat:byteSize [literal]
Prover uma descrição breve sobre cada atributo que compõe o conjunto de dados, juntamente com as suas unidades de medida, se aplicável.	fgdc:eainfo	Delimitar o bloco de descrição das entidades e dos atributos		fgdc:eainfo [rdfs:literal]	fgdc:eainfo [rdfs:literal]
	fgdc:eaover	Descrever, de maneira sumarizada, os atributos disponíveis		fgdc:eaover [rdfs:literal]	fgdc:eaover [rdfs:literal]
	fgdc:attrlabl	Especificar o nome do atributo		fgdc:attrlabl [rdfs:literal]	fgdc:attrlabl [rdfs:literal]
	fgdc:attrtype	Especificar o tipo do atributo		fgdc:attrtype [rdf:literal]	fgdc:attrtype [rdf:literal]
	fgdc:attrdef	Descrever o conceito do atributo (definição)		fgdc:attrdef [rdfs:literal]	fgdc:attrdef [rdfs:literal]
	fgdc:atprecis	Especificar a precisão do atributo, se aplicável		fgdc:atprecis [rdfs:literal]	fgdc:atprecis [rdfs:literal]
	fgdc:attscale	Especificar a escala do atributo, se aplicável		fgdc:attscale	fgdc:attscale [rdfs:literal]

Fonte: o Autor.

3.2.2 Métricas intrínsecas dos conjuntos de dados

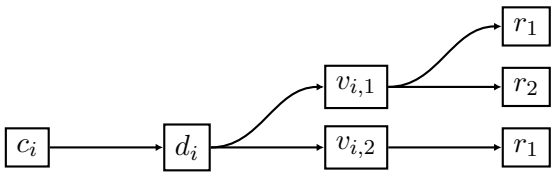
O modelo de avaliação de conjuntos de dados possui uma perspectiva de aspectos intrínsecos de um conjunto de dados. Tais aspectos representam características inerentes a um conjunto e que não são dependentes de uma área de aplicação. Por exemplo, um dos aspectos intrínsecos é a disponibilidade dos agentes, a qual busca mensurar o quanto os provedores dos dados e demais envolvidos na criação e publicação de dados forneceram informações de contato ao longo da descrição dos conjuntos. Esta característica, por sua vez, pode ser mensurada independentemente da área de pesquisa para o qual o conjunto de dados será utilizado. Os aspectos específicos de uma área são apresentados como métricas sociais, os quais são apresentados na Seção 3.2.3.

Para a apresentação das métricas utilizadas para os aspectos intrínsecos, buscou-se uma formalização matemática, a qual busca promover níveis de clareza e objetividade adequados. Tal formalização é apresentada a seguir.

Seja p um portal de dados. $C(p)$ contém os catálogos de conjuntos de dados disponíveis em p. $D(c_i) = \{d_1, d_2, ..., d_n\}$ representa o conjunto de *datasets* existentes em c_i, sendo $c_i \in C(p)$.

Considere o *dataset* d_i, $d_i \in D(c_i)$. d_i possui j versões, representadas por $V(d_i) = \{v_{i,1}, v_{i,2}, \ldots, v_{i,j}\}$, onde $v_{i,j}$ é a versão j do *dataset* i. Cada versão $v_{i,j}$ possui um conjunto de recursos, dado por $R(v_{i,j}) = \{r_1, r_2, \ldots, r_n\}$. A Figura 13 apresenta visualmente como este modelo se encontra estruturado.

Figura 13: Modelo de Estruturação de Conjuntos de Dados.



Fonte: o Autor.

Um *perfil de aplicação* é um conjunto de elementos de metadados, regras e recomendações definidas para uma aplicação específica. Seja PA um perfil de aplicação composto por um conjunto de elementos de metadados, $keys_{pa}$, onde $keys_{pa} = \{K_r \cup K_o\}$. Neste conjunto, K_r representa o subconjunto de elementos recomendados e K_o, por sua vez, o subconjunto de elementos opcionais do perfil de aplicação PA. O conjunto de elementos utilizados pelo produtor dos dados para descrever o *dataset* d_i, é dado por $Keys(d_i)$.

A seguir são apresentadas as métricas para os aspectos intrínsecos dos conjuntos de

dados, as quais irão avaliar os metadados que descrevem um *dataset* e seus recursos. Tais métricas são aplicadas em uma determinada versão $v_{i,j}$ de um *dataset* d_i, o qual pertence ao catálogo c_i.

As métricas propostas são inspiradas nos trabalhos de Dumontier et al. (2016), Kučera, Chlapek e Nečaskỳ (2013), Mohammadi et al. (2014), Umbrich, Neumaier e Polleres (2015), Vines et al. (2014), Zaveri et al. (2016) e Munafò et al. (2017).

A seguir, apresenta-se a definição das seguintes métricas:

- Completude dos Metadados;
- Disponibilidade dos Recursos;
- Disponibilidade dos Agentes;
- Maturidade de Dados Abertos;
- Manutenibilidade dos Recursos.

De modo a facilitar a compreensão das métricas, adotar-se-á o termo *dataset* como equivalente a uma determinada versão do *dataset* d_i, sob a qual uma métrica pode ser aplicada. Adicionalmente, o *dataset* d_i será referenciado apenas como d, representando assim a versão sob a qual as métricas apresentadas serão aplicadas.

Nas fórmulas a serem apresentadas, a notação $|x|$ indica a cardinalidade de x.

3.2.2.1 Completude dos Metadados

A métrica de *completude dos metadados* mensura o grau de descrição de um conjunto de dados pelos elementos recomendados por um perfil de aplicação. Esta métrica é inspirada no trabalho de Kučera, Chlapek e Nečaskỳ (2013), o qual versou sobre o estabelecimento de elementos recomendados para a descrição de conjuntos de dados.

A *completude dos metadados* verifica adicionalmente se os valores armazenados em cada elemento de descrição não são nulos ou, fazendo uso de um vocabulário definido, garante que o vocabulário esteja sendo respeitado, passando então a validar o conteúdo existente em cada elemento de metadado.

Completude dos Metadados

Seja k um elemento de metadado utilizado para descrição de um objeto (*dataset* ou recurso). Se k deve ter seu conteúdo definido por um vocabulário controlado, *vocab*, a completude de k em relação a *vocab*, $compl(k, d, vocab)$ é mensurada da seguinte forma:

$$compl(k, d, vocab) = \begin{cases} 1, \ se \ k \in |K_r^d| \ \wedge \ valor(k) \in vocab \\ 0, \ caso \ contrário \end{cases} \tag{3.1}$$

K_r^d representa o conjunto de elementos de metadados recomendados para a descrição do conjunto de dados d. No caso de não haver o uso de um vocabulário controlado para k, a completude é dada por:

$$compl(k, d) = \begin{cases} 1, \ se \ k \in |K_r^d| \ \wedge \ valor(k) \ \neq null \\ 0, \ caso \ contrário \end{cases} \tag{3.2}$$

De modo a consolidar as equações (3.1) e (3.2), a equação (3.2) pode ser reescrita como:

$$compl(k, d) = \begin{cases} 1, \ se \ k \in |K_r^d| \ \wedge \ (valor(k) \ \in vocab \ \vee \ valor(k) \neq null) \\ 0, \ caso \ contrário \end{cases} \tag{3.3}$$

A equação (3.3) trata, portanto, os casos em que k está sujeita a um vocabulário controlado ($valor(k) \in vocab$) ou, na hipótese de não haver vocabulário controlado para k, verifica se o valor do elemento de metadado k não é nulo.

Fazendo uso da equação (3.3), a completude dos metadados para um *dataset* d, $Q_c(d)$, é dada por:

$$Q_c(d) = \frac{\sum_{k \in K_r^d} compl(k, d) + (\sum_{r \in R(d)} \sum_{k \in K_r^r} compl(k, r))}{|K_r^d| + |K_r^r| \cdot |R(d)|} \tag{3.4}$$

onde K_r indica o conjunto dos elementos recomendados. $R(d)$ indica o conjunto dos recursos do *dataset* d. K_r^r representa o conjunto de elementos de metadados recomendados para a descrição do recurso r.

3.2.2.2 Disponibilidade dos Recursos

No trabalho de Vines et al. (2014), os autores realizaram uma análise em 516 artigos científicos com data de publicação entre 2 e 22 anos. O resultado demonstrou que, dentre os artigos que disponibilizaram os dados em algum repositório digital, a probabilidade de se encontrar tais dados decaiu a uma taxa de 17% ao ano. Os autores evidenciaram a necessidade de que os recursos de um *dataset* estejam disponíveis e, de preferência, publicados em um portal de dados que facilite a busca e recuperação de tais recursos.

A métrica de *disponibilidade dos recursos* refere-se ao grau no qual os dados de um *dataset* e seus recursos associados podem ser capturados por um humano ou *software* específico.

Disponibilidade dos Recursos

A disponibilidade de recursos de um *dataset d* é obtida por meio da validação da URL (*Uniform Resource Locator*) de cada recurso. Tal validação se dá por uma requisição GET do protocolo HTTP (*Hypertext Transfer Protocol*), a qual tenta obter um determinado recurso. Se o valor retornado por GET for 200, o recurso foi obtido com sucesso. Um valor diferente de 200 indica erros na requisição. Por exemplo, o retorno 204 representa que não houve resposta do servidor, enquanto o retorno 403 indica que o acesso ao recurso foi negado. É importante ressaltar que o método GET é apenas uma das possíveis formas de se obter recursos de dados por meio do protocolo HTTP. Outros métodos podem ser utilizados, modificando apenas os aspectos técnicos de implementação da métrica.

Seja x_{r_i} a URL de acesso do recurso r_i, a disponibilidade $disp(r_i)$ é dada por:

$$disp(r_i) = \begin{cases} 1, se\ HTTP_{GET}(x_{r_i}) = 200 \\ 0,\ caso\ contrário \end{cases} \tag{3.5}$$

Com base na equação (3.5), a disponibilidade dos recursos do *dataset* d, $Q_d(d)$, é dada por:

$$Q_d(d) = \frac{\sum_{r \in R(d)} disp(r)}{|R(d)|} \tag{3.6}$$

3.2.2.3 Disponibilidade dos Agentes

A Ciência é um organismo social e a interação entre seus agentes é essencial para o processo de evolução do conhecimento. É fundamental que os agentes envolvidos (produtor dos dados, colaboradores de pesquisa, instituições envolvidas, etc.) sejam acessíveis. No trabalho de Vines et al. (2014), entre 516 artigos avaliados, com data de publicação entre 2 e 22 anos, a probabilidade de se encontrar um e-mail válido decaiu a uma taxa de 7% ao ano. Constatou-se, portanto, um aumento da indisponibilidade dos recursos e da forma de contato especificada nas publicação, em função do tempo. Quanto mais antigo um artigo, por exemplo, maior será a probabilidade de que o canal de comunicação definido pelos autores esteja indisponível ou, até mesmo, inexistente.

A métrica de disponibilidade dos agentes busca verificar se informações de contato estão disponíveis no conjunto de dados e se seu conteúdo é válido.

Como exemplo de dados de contato destacam-se o e-mail do produtor dos dados ou a URL do *website* onde o consumidor pode obter mais informações. Embora a validação dos e-mails possa ser feita por meio de um mecanismo de confirmação de e-mail válido, muito utilizado nos *websites* contemporâneos, há um conjunto de recomendações na RFC 2822 (RESNICK, 2001) que deve ser levado em consideração, principalmente quanto à sintaxe de um endereço eletrônico de e-mail.

A verificação da disponibilidade do e-mail de contato tem como premissa que, no momento do cadastro do conjunto de dados no portal, um e-mail de verificação é enviado, o qual deve ser confirmado pelo produtor. Uma vez que o e-mail é confirmado, a métrica de disponibilidade dos agentes verificará, em determinado intervalo de tempo, se o servidor de e-mails indicado continua em funcionamento.

Por exemplo, para o e-mail *andrefmb@usp.br*, há um registro no DNS (*Domain Name Server*), que indica quais são os servidores de e-mail para o domínio *usp.br*. A verificação (*lookup*) da entrada MX (*mail exchanger*) no DNS atua, portanto, como uma forma para verificar se o domínio possui servidores responsáveis por contas de e-mail. Como exemplo oposto, embora o e-mail *andrefmb@usp123.br* possua sintaxe válida em relação à RFC 2822, uma busca no DNS não encontra resultados para MXs do domínio *usp123.br*. Dessa forma, o e-mail *andrefmb@usp123.br* é considerado indisponível.

Sobre as URLs de contatos, a validação das mesmas pode ser feita de maneira análoga a função auxiliar definida na equação (3.5).

Disponibilidade dos Agentes

Seja d um *dataset* no qual diversos agentes podem ser relacionados, tais como: produtor dos dados, instituição de pesquisa, publicador, dentre outros.

Seja $ag(d) = \{a_1, a_2, \ldots, a_n\}$, os agentes relacionados no dataset d. Um agente a_i pode conter duas informações de contato: $email_{a_i}$ (e-mail de contato) e/ou $URLa_i$ (URL de contato). A validação destes elementos se dá pelas seguintes equações:

$$válido(email_{a_i}) = \begin{cases} 1, se\ Sintaxe_{RFC2822}(email_{a_i})\ \wedge\ Lookup_{MX}(email_{a_i}) \\ 0,\ caso\ contrário \end{cases} \tag{3.7}$$

onde $Sintaxe_{RFC2822}(email_{a_i})$ possui valor verdadeiro quando $email_{a_i}$ possui sintaxe validada de acordo com as regras da RFC 2822, e $Lookup_{MX}(email_{a_i})$ possui valor verdadeiro quando a busca por registros MX para o domínio em questão são encontrados no DNS.

$$válido(url_{a_i}) = \begin{cases} 1, se\ HTTP_{GET}(url_{a_i}) = 200 \\ 0,\ caso\ contrário \end{cases} \tag{3.8}$$

Consolidando as equações (3.7) e (3.8), tem-se:

$$válido(a_i) = \begin{cases} 1, se\ válido(email_{a_i})\ \vee\ válido(url_{a_i}) \\ 0,\ caso\ contrário \end{cases} \tag{3.9}$$

A *disponibilidade dos agentes* de d, Q_{da}, é dada por:

$$Q_{da}(d) = \frac{\sum_{a_i \in ag(d)} válido(a_i)}{|ag(d)|} \tag{3.10}$$

3.2.2.4 Maturidade de Dados Abertos

De acordo com Oliveira (2016), dados científicos abertos são dados de pesquisa disponíveis livremente para reuso e republicação sem restrições de *copyright*, patentes ou controle de propriedade intelectual. Devem ser disponíveis gratuitamente na Internet, permitindo a qualquer usuário baixar, copiar, analisar, reprocessar, capturá-los via *software*, utilizá-los, reusar e reproduzir para qualquer finalidade, sem barreiras financeiras,

legais ou técnicas adicionais.

Esta métrica está relacionada ao conceito de maturidade de dados abertos estabelecido por Tim Berners-Lee (BIZER; HEATH; BERNERS-LEE, 2009), o qual define 5 níveis de maturidade de dados abertos, classificados em número de estrelas. De uma maneira breve, os 5 níveis de maturidade propostos são definidos como:

- ($\star$) 1 estrela: os dados estão disponíveis na Web, independente de formato, sob uma licença aberta;
- ($\star\star$) 2 estrelas: este nível contempla o nível anterior, além do fato de que os dados estão publicados em um formato compreensível por máquina;
- ($\star\star\star$) 3 estrelas: contempla os níveis anteriores, além do fato de que o formato do arquivo é não-proprietário;
- ($\star\star\star\star$) 4 estrelas: contempla os níveis anteriores, além de fazer uso de URIs bem definidas para a identificação de todos os elementos associados;
- ($\star\star\star\star\star$) 5 estrelas: é o último nível de maturidade proposto. Contempla os requisitos dos níveis anteriores, considerando também o uso de *linked data*.

Um perfil de aplicação pode ser descrito por metadados em um formato *linked data*, tais como os formatos RDF/XML e JSON-LD. A aplicação desta métrica em metadados descritos em formatos *linked-data* tem por objetivo verificar se o provedor de dados atende a três requisitos:

1. O *dataset* se encontra com uma licença de uso definida;
2. O formato dos recursos do *dataset* é um formato aberto e compreensível por máquina;

A verificação da licença de um conjunto de dados é realizada pela validação quanto ao preenchimento do elemento `<dct:accessRights>` no *dataset*. É por meio deste elemento que o produtor de dados especifica a licença de uso do conjunto de dados.

Para verificar se o formato de dados do conjunto de dados é aberto, não proprietário e compreensível por máquinas, faz-se o uso do elemento `<dct:format>`, a qual contém o tipo de arquivo disponibilizado por um recurso do *dataset*. Como exemplos de formatos que atendem tais requisitos, pode-se citar: csv, geojson, html, jpeg, netcdf e xml.

Maturidade de Dados Abertos

Inicialmente, verifica-se a completude do elemento `<dct:accessRights>` especificada para o *dataset d*. Por meio da verificação da completude do elemento de metadado, estabelece-se um deflator denominado *mrd* (*machine readable deflactor*), o qual penaliza a métrica no caso de ausência do elemento de metadado sobre a licença.

Para cada recurso r do dataset d, verifica-se a URL disponível no elemento `<dcat:downloadURL>` e determina-se o formato do arquivo referenciado pela URL, de modo a validar se o formato é aberto e compreensível por máquina. A função $R_{open}(r), r \in R(d)$, verifica se o recurso r se encontra em um formato aberto e compreensível por máquina, é dada por:

$$R_{open}(r) = \begin{cases} 0, se\ r\ é\ um\ formato\ aberto\ \wedge compreensível\ por\ máquina \\ 1,\ caso\ contrário \end{cases} \tag{3.11}$$

O deflator *mrd* é calculado com base na seguinte função:

$$mrd = \begin{cases} 0,5, se\ compl(k_{<dct:accessRights>}, d)\ = 0 \\ 0,5 \cdot \sum_{r \in R(d)} \frac{R_{open}(r)}{|R(d)|},\ caso\ contrário \end{cases} \tag{3.12}$$

Uma vez que o deflator foi calculado, a *maturidade de dados abertos* do *dataset* d, $Q_m(d)$, é dada por:

$$Q_m(d) = max(0, (1 - mrd)) \tag{3.13}$$

3.2.2.5 Manutenibilidade dos Recursos

Conhecer a frequência de atualizações ou rotinas de ajustes em conjuntos de dados é um elemento importante para subsidiar a escolha de um conjunto de dados, auxiliando a tomada de decisão sobre o uso do conjunto de dados em questão para atendimento a uma necessidade específica.

Esta métrica busca mensurar o grau de comunicação do provedor de dados sobre a frequência de atualização do conjunto de dados, possibilitando assim que o consumidor possa efetuar consultas específicas como, por exemplo, “obter conjuntos de dados que são atualizados semanalmente”.

A *manutenibilidade dos recursos* pode ser insumo para outros mecanismos de avaliação, proporcionando o desenvolvimento de novas métricas calculadas automaticamente. Por exemplo, se o provedor de dados informa que um conjunto de dados recebe atualizações semanais, uma métrica para avaliar se as atualizações de fato estão ocorrendo pode ser desenvolvida, mensurando a diferença entre manutenibilidade planejada e a manutenibilidade realizada.

Manutenibilidade dos Recursos

O elemento `<dct:accuralPeriodicity>` indica a frequência de atualização do recurso (diária, mensal, anual, etc.). Em situações em que um recurso não possui frequência de atualização definido, este elemento pode ser utilizado para explicitar que não há previsão de atualização do recurso.

A *manutenibilidade dos recursos* do *dataset* d, denominada $Q_p(d)$, é calculada com base na equação de completude dos metadados (3.2).

$$Q_p(d) = \frac{\sum_{r \in R(d)} compl(< dct : accuralPeriodicity >, r)}{|R(d)|} \tag{3.14}$$

3.2.3 Métricas Sociais

O modelo de avaliação de conjuntos de dados se baseia em um mecanismo de avaliação social, no qual um grupo de usuários, denominados consumidores de dados, agrupados por uma área de pesquisa ou de aplicação, irão avaliar um ou mais conjuntos de dados.

A avaliação de conjuntos de dados pela comunidade científica é uma exigência cada vez mais requisitada por instituições financiadoras de pesquisas mundiais. Por exemplo, no ano de 2014, o Conselho de Pesquisas em Engenharia e Ciências Físicas (EPSRC), principal agência de financiamento de pesquisas na área de Engenharia e Ciências Físicas do Reino Unido, publicou os requisitos para a publicação de conjuntos de dados gerados pelas pesquisas financiadas, estabelecendo a necessidade de que os autores busquem métricas de avaliação da comunidade científica (não apenas o número de citações ou de *downloads*), para verificar a efetividade e o impacto do conjunto de dados para o uso pela comunidade científica (EPSRC, 2014).

Ainda sobre a necessidade de mecanismos de avaliação de conjuntos de dados, no trabalho de Ball e Duke (2015) os autores destacam a importância de que o *feedback* da comunidade científica pode fomentar o reuso dos conjuntos dos dados, assim como permitir

ao provedor de dados verificar se os métodos adotados para a produção, publicação e disponibilização do conjunto de dados estão produzindo resultados satisfatórios para a comunidade científica.

A adoção um modelo que considere a avaliação da comunidade científica para o cálculo da Veracidade de um conjunto de dados possibilita a compreensão sobre o valor e o impacto de um conjunto de dados, assim como o quanto os processos utilizados para sua produção e publicação estão aderentes às necessidades da comunidade científica. Tal conceito vai ao encontro do exposto por Oliveira (2016), a qual afirma que os pesquisadores estão focados, além do fato de obter a recompensa científica por seus trabalhos, em terem seus trabalhos reutilizados e reproduzidos por experimentos de pesquisa semelhantes.

O *feedback* da comunidade científica sobre um conjunto de dados é mensurado pela métrica de aspecto social denominada *confiança do consumidor*. A métrica consiste em um questionário avaliado pelo consumidor de dados. Com base nas respostas, calcula-se a avaliação do consumidor de dados em relação ao conjunto de dados. Tal avaliação, juntamente com as métricas de aspecto intrínseco, irão possibilitar a elaboração do índice de veracidade dos dados.

Ao longo do desenvolvimento dessa métrica, deparou-se com a necessidade de lidar com eventuais vieses de avaliação. Por exemplo, caso um conjunto de dados seja avaliado por um consumidor que costuma atribuir o menor valor da escala aos conjuntos de dados avaliados, e este valor não concordar com as demais avaliações, tal avaliação deverá receber um peso menor no cálculo do índice. Outro exemplo ocorre quando um conjunto de dados é avaliado por um único consumidor, recebendo nota máxima, e tal nota passa a ser o índice de veracidade do conjunto. Para esta situação, o modelo deve favorecer conjuntos de dados que possuam um maior número de avaliações, as quais atuarão subsidiando a determinação do consenso que está se formando na rede de avaliações sobre o conjunto de dados em questão. Dessa forma, um conjunto que foi avaliado por apenas uma única nota não pode ter o índice de veracidade representado pelo mesmo valor.

Essa problemática requer o desenvolvimento de estratégias que determinam a reputação de um usuário em uma rede de avaliações, na qual um grupo de usuários avaliam um conjunto de objetos. A reputação do usuário irá definir o peso de suas avaliações na referida rede. Com este peso, ponderar-se-ão as avaliações do usuário efetuadas na rede.

Ao longo do desenvolvimento da pesquisa, os algoritmos propostos por Allen e Seaman (2007), Liao et al. (2014), Gao et al. (2015) e Liu et al. (2017) foram avaliados e implementados, com o objetivo de escolher qual seria a estratégia mais adequada para

tratar tais situações. Ao término do processo de implementação e análise, optou-se por desenvolver uma versão modificada[2] do algoritmo *User Reputation via the Beta Probability Distribution* (RBPD) proposto por Liu et al. (2017), a qual será apresentada na Seção 3.2.3.1. A escolha por fazer do algoritmo RBPD a estratégia para modelar o comportamento da rede de avaliações se deu pelo fato de sua simplicidade de implementação, além de não fazer uso de parâmetros de inicialização, o que requer um estudo mais aprofundado para a determinação dos parâmetros necessários utilizados pelos outros modelos, indo de encontro ao escopo delimitado para o desenvolvimento desta pesquisa.

Como as avaliações sobre um conjunto de dados podem ser sensíveis ao tempo, faz-se necessário usar um fator de decaimento com base no tempo de avaliação (denominado fator de *aging*). Tal mecanismo será apresentado na Seção 3.2.3.2.

A composição de todas as avaliações para um conjunto de dados forma o valor da perspectiva social do *índice de veracidade* deste conjunto de dados. Considera-se que, para cada área de pesquisa ou de aplicação, indo ao encontro do conceito de *fitness for use* (TAYI; BALLOU, 1998), uma rede de avaliações será formada por consumidores e conjuntos de dados pertencentes à respectiva área. Dessa forma, o modelo pressupõe que cada área de pesquisa pode adaptar o método do cálculo do índice para a sua realidade, assim como customizar a mensuração da confiança do consumidor em função das características específicas que um conjunto de dados possa possuir.

As Seções a seguir apresentam o algoritmo de reputação em redes de avaliação, RBPD, originalmente proposto por Liu et al. (2017), assim como a proposta de modificação do RBPD para a rede de avaliações utilizada nesta pesquisa. Também é apresenta a estratégia de *aging* das avaliações. O algoritmo de reputação e a estratégia de *aging* serão elementos necessários para o cálculo do índice de veracidade, em sua perspectiva social, o qual será apresentado na Seção 3.2.4.

3.2.3.1 Reputação em Redes de Avaliações

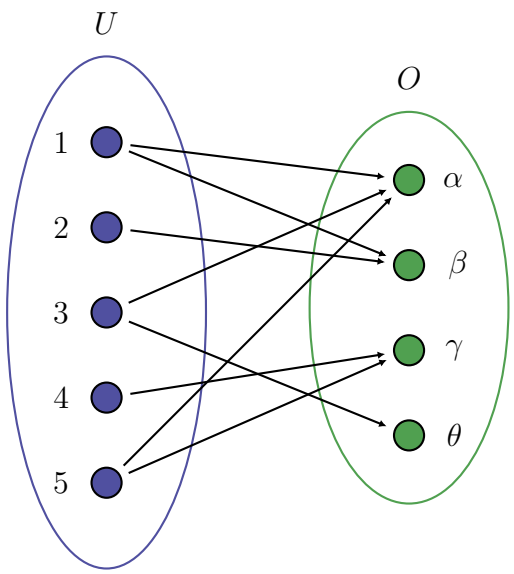
A seguir apresenta-se a versão original do algoritmo RBPD, proposta por Liu et al. (2017), juntamente com a proposta de modificação do RBPD para suportar o conceito de uma rede de avaliações de conjuntos de dados. Para fins de simplificação, a versão customizada do RBPD, a qual será utilizada nesta tese, será denominada RBPD*.

[2] A customização do algoritmo RBPD, utilizada nesta tese, encontra-se disponível no endereço eletrônico <https://goo.gl/QJNvBn>

3.2.3.1.1 Algoritmo RBPD proposto Liu et al. (2017)

O algoritmo *User Reputation via the Beta Probability Distribution* (RBPD) proposto por (LIU et al., 2017) consiste em determinar a reputação de usuários em uma rede de avaliação sobre a qualidade de objetos, por meio de uma análise Bayesiana. Tal rede de avaliação é estruturada como um grafo bipartido com pesos, possuindo dois tipos de nós: usuários (U) e objetos (O). Os números de usuários, objetos e avaliações existentes no grafo (arestas) é denotado por $|U|, |O|$ e $|E|$, respectivamente. A avaliação $r_{i\gamma}$ dada pelo usuário u_i para o objeto o_γ é o peso da aresta no grafo bipartido. A Figura 14 apresenta a modelagem da rede de avaliação como um grafo bipartido.

Figura 14: Rede de Avaliação modelada como um grafo bipartido.



Fonte: o Autor.

Todas as avaliações são armazenadas em uma matriz $\mathbf{A}$. O conjunto U_γ é formado por todos os usuários que avaliaram o objeto o_γ. Por sua vez, o conjunto O_i é constituído por todos os objetos avaliados pelo usuário u_i. Adicionalmente, o grau de um usuário u_i e de um objeto o_γ é dado por k_i e k_γ, respectivamente.

Na rede de avaliação, a reputação de um usuário u_i é dada por R_i. A análise Bayesiana adota para cada avaliação efetuada por um usuário na rede: avaliação justa (*fair*, denotada por 1) e avaliação injusta (do inglês *unfair*, denotada por 0). A noção de uma avaliação justa efetuada na rede é dada pela seguinte definição: o usuário u_i provê uma avaliação $r_{i\gamma}$ sobre um objeto o_γ e tal avaliação será julgada de modo a determinar a sua consistência quando comparada a outras avaliações do objeto o_γ efetuadas por outros usuários da rede. A determinação da consistência de uma avaliação se dá quando esta mesma avaliação é

feita por pelo menos 50% dos usuários que avaliaram o objeto o_γ. Para esta contabilização, faz-se necessário inicialmente normalizar todas as avaliações existentes na rede, fazendo uso da seguinte função:

$$r'_{i\gamma} = \begin{cases} \frac{2(r_{i\gamma}-r_i^{min})}{r_i^{max}-r_i^{min}} - 1, \ se \ r_i^{max} \neq r_i^{min} \\ 0, \ se \ \ r_i^{max} = r_i^{min} \end{cases} \tag{3.15}$$

onde r_i^{max} e r_i^{min} denotam a maior e menor avaliação dada pelo usuário i na rede, respectivamente.

Dessa maneira, todas as avaliações são normalizadas e vão assumir valores entre [-1, 1], onde a maior avaliação que um usuário forneceu na rede será mapeada para 1, enquanto a menor avaliação será mapeada para -1. Para os casos em que a maior e menor avaliação de um usuário possuem valores iguais, suas avaliações serão normalizadas para o valor 0. A matriz de avaliações normalizadas é dada por $\mathbf{A}'$. Com base nessa normalização, os valores da matriz $\mathbf{A}'$ indicam se uma avaliação do usuário é positiva ($r'_{i\gamma} > 0$) ou negativa ($r'_{i\gamma} <= 0$).

$$\mathbf{A}' = \begin{array}{cc} o_\gamma & \\ \begin{bmatrix} -0.5 \\ 0.5 \\ 0.8 \end{bmatrix} & \begin{array}{c} u_1 \\ u_2 \\ u_3 \end{array} \end{array}$$

A partir das avaliações normalizadas em $\mathbf{A}'$, constrói-se a matriz $\mathbf{B}$, proveniente da consistência das avaliações dos usuários para cada objeto. Por exemplo, supondo que em uma rede de avaliações o objeto o_γ foi avaliado pelos usuários u_1, u_2 e u_3. Como em $\mathbf{A}'$, mais do que 50% das avaliações normalizadas são positivas, considera-se que as avaliações normalizadas positivas ($r'_{i\gamma} > 0$) serão classificadas como justas (*fair*, denotada por Y), enquanto que as demais avaliações normalizadas (no caso, $r'_{i\gamma} <= 0$) serão classificadas como injustas (*unfair*, denotada por N), populando assim a matriz B.

$$\mathbf{B} = \begin{array}{cc} o_\gamma & \\ \begin{bmatrix} N \\ Y \\ Y \end{bmatrix} & \begin{array}{c} u_1 \\ u_2 \\ u_3 \end{array} \end{array}$$

No trabalho de Liu et al. (2017) os autores demonstram que, sob este modelo, a reputação de um usuário i é definida pela probabilidade θ_i de que o usuário i irá fornecer uma nota justa para os objetos da rede, a qual será um valor entre [0,1]. Os autores demonstram que tal probabilidade pode ser obtida por meio do cálculo do valor esperado, e este valor pode ser determinado fazendo uso da distribuição de probabilidade beta. Não é escopo desta tese demonstrar esta prova, podendo a mesma ser verificada no referido trabalho. Sendo assim, os autores propõem que a reputação R_i do usuário u_i é dada por:

$$R_i = \frac{s_i + 1}{k_i + 2} \tag{3.16}$$

Onde s_i denota o número de avaliações efetuadas por u_i que foram classificadas como *fair* (Y), e que estão armazenadas na matriz **B**. k_i é o grau do usuário u_i.

Uma vez determinada a reputação dos usuários, pode-se determinar a qualidade do objeto avaliado. A qualidade Q_γ do objeto γ é dada por:

$$Q_\gamma = max_{i \in U_\gamma}\{R_i\} \frac{\sum_{i \in U_\gamma} R_i r_{i\gamma}}{\sum_{i \in U_\gamma} R_i} \tag{3.17}$$

onde $max_{i \in U_\gamma}\{R_i\}$ é definido como um fator de penalização baseado na seguinte hipótese: se um objeto é avaliado por usuários de baixa reputação, apesar de as avaliações recebidas serem altas, tal objeto não poderá ser classificado como de alta qualidade.

- **Exemplo de execução do algoritmo RBPD**

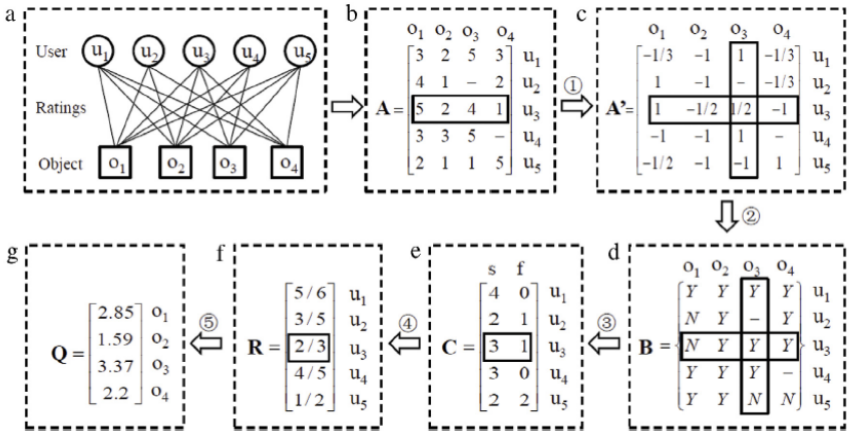
  A Figura 15 apresenta um exemplo de execução do algoritmo RBDP, extraído do artigo original de Liu et al. (2017).

  Considere-se a rede de avaliações apresentada na Figura 15(a), formada por cinco usuários $U = \{u_1, u_2, u_3, u_4, u_5\}$ e quatro objetos $O = \{o_1, o_2, o_3, o_4\}$. As avaliações dos usuários sobre os objetos são armazenadas na matriz **A**. Na matriz **A** [Figura 15(b)], as linhas representam os usuários, enquanto as colunas representam os objetos que compõem a rede. Se $r_{i\gamma} = -$, então o usuário u_i não avaliou o objeto o_γ. O próximo passo do algoritmo consiste em normalizar as avaliações, armazenando-as na matriz $\mathbf{A}'$ [Figura 15(c)]. Tomando o usuário u_3 como exemplo, tem-se que

  $$r_{32} = 2, r'_{32} = 2 \cdot \frac{2-1}{5-1} - 1 = -\frac{1}{2}$$

  Uma vez normalizadas, as avaliações são armazenadas na matriz **B** [Figura 15(d)]. Tomando novamente como exemplo o objeto o_3, em $\mathbf{A}'$ temos que $r'_{13} > 0, r'_{33} >$

Figura 15: Exemplo de execução do algoritmo RBPD.



Fonte: (LIU et al., 2017).

$0, r'_{43} > 0$ *e* $r'_{53} < 0$. Portanto, as avaliações dadas por u_1, u_3 *e* u_4 serão consideradas justas (Y), enquanto que a avaliação u_5 será considerada injusta (N).

Seja s_i e f_i o número de avaliações realizadas por u_i consideradas justas e injustas, respectivamente. Tais valores são armazenados na matriz **C** [Figura 15(e)]. Tomando como exemplo o usuário u_3, que atribuiu 3 avaliações consideradas justas e 1 avaliação considerada injusta, de modo que $s_3 = 3, f_3 = 1$. Em [Figura 15(f)], tem-se a matriz de reputação **R**, calculada a partir da seguinte regra:

$$R_3 = \frac{1 + s_3}{2 + s_3 + f_3} = \frac{2}{3}$$

Em [Figura 15(g)] tem-se a matriz da qualidade dos objetos. Considerando o_3 como um exemplo, tem-se que:

$$Q_3 = \frac{5}{6} \cdot \frac{5 \cdot \frac{5}{6} + 4 \cdot \frac{2}{3} + 5 \cdot \frac{4}{5} + 1 \cdot \frac{1}{2}}{\frac{5}{6} + \frac{2}{3} + \frac{4}{5} + \frac{1}{2}} = 3,37$$

3.2.3.1.2 Algoritmo RBPD* para a Rede de Avaliações de Conjuntos de Dados Científicos

A rede de avaliações modelada neste trabalho fará uso de uma versão modificada do RBPD, a qual será denominada **RBPD***.

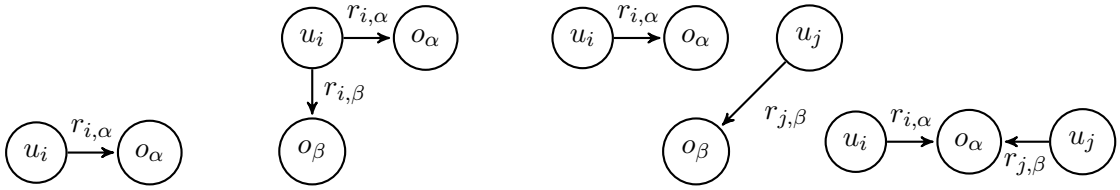
Tal modificação se deu pelo fato de que o algoritmo RBPD original atua na determinação da reputação em redes de avaliações pre-existentes, não abordando situações em que a rede está em formação. No trabalho em questão, os autores fazem uso do algoritmo para mensurar a reputação em redes com mais de 1 milhão de usuários, cada usuário

possuindo no mínimo 20 avaliações[3].

Para uma rede com um único usuário e um único objeto de avaliação, o algoritmo RBPD original irá calcular uma reputação, conforme equação 3.16, de 0.33 (33%) para o usuário. Baseando-se em uma cultura de avaliação científica justa, faz-se necessário modificar o RBDP para que, nessa situação, o avaliador receba uma reputação de 50%,em virtude de que não é possível classificar a sua reputação apenas com uma avaliação.

As Figuras 16a, 16b, 16c e 16d ilustram possíveis configurações na rede de avaliação, em que a reputação inicial dos usuários será calculada como 0,5 (50%).

Figura 16: RBPD*: configurações da rede de avaliação em que a reputação dos usuários (u_i,u_j) será 0,5 (50%).



(a) Um único usuário avaliando um único conjunto de dados

(b) Um único usuário na rede, avaliando múltiplos conjuntos de dados.

(c) Dois usuários na rede, avaliando dois conjuntos de dados distintos.

(d) Dois usuários na rede, avaliando um único conjunto de dados.

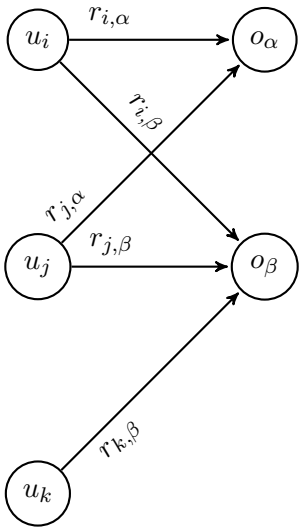
A partir do momento em que o usuário passar a avaliar outros conjuntos de dados, a sua reputação será calculada conforme a coerência de suas notas em relação ao consenso que está sendo formado na rede. Para que a reputação de um usuário possa ser afetada, faz-se necessário que na rede haja ao menos dois objetos distintos, o_i e o_j, avaliados pelo usuário, onde $grau(o_i) > 2$, $grau(o_j) > 2$.

Para fins de exemplificação, considere-se a rede de avaliações apresentada na Figura 17a. Ela representa o tempo zero de uma rede de avaliações, na qual os usuários u_i, u_j e u_k possuem uma reputação de 50%. Ao passo em que, no tempo 1 (Figura 17b), o usuário u_k avalia o objeto o_α. Com isto, os usuários u_i, u_j e u_k terão suas reputações afetadas, em virtude de que agora é possível determinar o consenso que está sendo formado na rede. Quanto maior a rede de avaliações, e maior o seu grau de conectividade, mais apurado será o cálculo da reputação dos usuários.

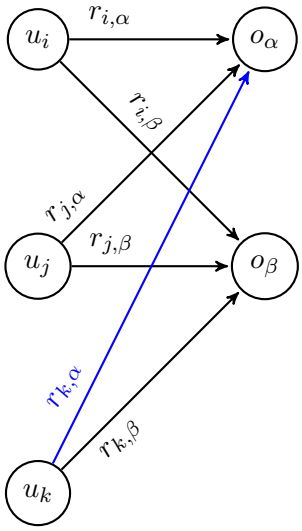
Algoritmo 1 apresenta o algoritmo RBPD* modificado para suportar estas configurações, recebendo um grafo $G(V, A)$, onde V representa o conjunto de vértices do grafo bipartido, e A as respectivas arestas. O valor de uma avaliação do usuário u sobre um objeto α é denotado por $a(u, \alpha)$. As funções auxiliares $maxA(v)$ e $minA(v)$, obtém, res-

[3]Rede de avaliações de filmes disponíveis na plataforma digital Netflix.

Figura 17: Análise da reputação dos usuários em uma rede de avaliações.



(a) Tempo 0: Reputação dos usuários em 50%.



(b) Tempo 1: Reputação dos usuários será afetada, pelo fato de que agora é possível determinar o consenso (maioria positiva ou maioria negativa) sobre os objetos avaliados.

pectivamente, a aresta de maior e a aresta de menor valor, que possuem como uma de suas extremidades o vértice v. A função auxiliar $moda(v, +)$ verifica se a maior parte das arestas que possuem v como um de seus vértices apresenta valores positivos, enquanto $moda(v, -)$ verifica se a maior partes das arestas possuem valores negativos. Já a função auxiliar $empate(v)$ retorna verdadeiro quando o número de arestas de v com valor positivo (> 0) é igual ao número de arestas com valor negativo (< 0).

Algoritmo 1: Algoritmo RBPD* para redes de avaliações de conjuntos de dados científicos.

```
   Entrada: Grafo G(V, A), onde V são os vértices e A são as arestas
   Saída: Grafo atualizado com a reputação dos usuários na rede de avaliação
1  início
2      para cada u ∈ usuarios(G) faça
3          para cada α ∈ vizinhos(u) faça
4              se maxA(u) = minA(u) então
5                  a(u, α) ← 0
6                  empate(u) ← verdadeiro
7              fim
8              senão
9                  a(u, α) ← (2·(a(u,α) − minA(u)) / (maxA(u) − minA(u))) − 1
10             fim
11         fim
12     fim
13     para cada α ∈ objetos(G) faça
14         para cada u ∈ vizinhos(α) faça
15             se moda(α,+) e a(u, α) > 0 então
16                 u.fair += 1
17             fim
18             se moda(α,-) e a(u, α) < 0 então
19                 u.fair += 1
20             fim
21             se empate(u) ou empate(α) então
22                 u.fair += 0,5
23             fim
24         fim
25     fim
26     para cada u ∈ usuarios(G) faça
27         u.reputacao ← (u.fair + 1) / (grau(u) + 2)
28     fim
29 fim
30 retorna G(V, A)
```

3.2.3.2 Decaimento Temporal de Avaliações

Diversas são as propostas para a implementação de mecanismos de *aging* em redes de avaliações (DING; LI, 2005; WU et al., 2012; LI et al., 2013; KOWALD et al., 2014). Dentre as propostas encontradas, é comum o uso do decaimento exponencial para refletir o efeito do tempo na avaliação de usuários em uma rede de avaliações (LARRAIN et al., 2015).

O mecanismo de *aging* com decaimento exponencial inicia-se com a determinação do tempo de meia-vida, T_m, de uma avaliação. Seja $a_{i\gamma}(0)$ uma avaliação feita pelo usuário u_i sobre o objeto o_γ no tempo $T = 0$. Quando $T = T_m$, esta avaliação terá 50% de seu valor inicial. Dessa forma:

$$a_{i\gamma}(T_m) = \frac{a_{i\gamma}(0)}{2} \tag{3.18}$$

A função de *aging* com decaimento exponencial, $aging(t)$, no tempo t, é dada por:

$$aging(t) = e^{-\lambda(t-t_0)} \tag{3.19}$$

onde λ é dado em função do tempo de meia-vida T_m, como segue:

$$\lambda = \frac{ln\ 2}{T_m} \tag{3.20}$$

O cálculo do novo valor da avaliação $a_{i\gamma}$ no tempo t, $t > t_0$, é dado por:

$$a_{i\gamma}(t) = a_{i\gamma}(0) \cdot aging(t) \tag{3.21}$$

A determinação do tempo de meia-vida se dá em função da natureza do conjunto de dados e de sua área de aplicação. Em um estudo sobre a determinação do tempo de meia-vida para o decaimento exponencial aplicado em redes de avaliação, Larrain et al. (2015) apresentam a proposta de que, quando não conhecido a priori, o tempo de meia-vida pode ser determinado como a mediana do *aging* das avaliações existentes na rede.

3.2.3.3 Confiança do Consumidor

Embora os aspectos intrínsecos sejam passíveis de verificação automática por máquinas, a confiança do consumidor no conjunto de dados, e consequentemente nos metadados que descrevem tal conjunto de dados, é elemento vital para verificar a adequabilidade do conjunto de dados para o uso e o reuso pela comunidade científica.

A métrica aqui proposta busca mensurar o grau de confiança do consumidor em relação ao conjunto de dados. A determinação da confiança é baseada em um questionário, o qual o consumidor irá julgar um conjunto de afirmações em uma escala de 1 a 5.

Na elaboração das afirmações a serem julgadas, fez-se uso do método *Goal-Question-Metric* (GQM) definido por Basili e Rombach (1988). Neste método, estabelece-se um conjunto de objetivos e, para cada objetivo, elencam-se afirmações que serão julgadas pelos consumidores com base em uma métrica. Para a métrica de avaliação, adotou-se a escala *Likert* (ALLEN; SEAMAN, 2007). Ao contrário das perguntas sim/não, a escala *Likert* permite mensurar o grau de conformidade do entrevistado com uma afirmação proposta.

Para a construção dos objetivos a serem validados pelo questionário, fez-se uso de

reuniões do grupo de estudos, pesquisa e extensão em Big Data da Escola Politécnica da Universidade de São Paulo, nas quais diversos pesquisadores colaboraram para a escolha dos objetivos a serem validados. Ressalta-se a valiosa contribuição da Profa. Dra. *Suzie Allard*[4], da *The University of Tennesseee* (EUA), especialista reconhecida na área de Gestão de Dados e Ciência da Informação, além de ser uma das pesquisadoras principais do projeto DataONE. A proposta do questionário, juntamente com seus elementos, foi objeto de avaliação da pesquisadora, de modo a possibilitar o alcance dos objetivos estabelecidos.

Buscou-se, também, como referência, o *checklist* de revisão de conjuntos de dados proposto pela USGS (USGS, 2014), assim como se buscou pela aderência ao código de conduta dos cientistas de dados, estabelecido pela *Data Science Association* (WALKER, 2015), o qual define um guia de conduta para profissionais que atuem na área de Ciência dos Dados. Dentre os principais elementos do código de conduta, destacam-se:

- Um cientista de dados deve informar sobre todos os resultados e fatos relevantes conhecidos que permitirão ao consumidor tomar decisões informadas, independentemente de as evidências serem ou não adversas à hipótese de pesquisa;
- Um cientista de dados deve avaliar a qualidade dos dados e divulgar tal classificação para permitir que o consumidor tome decisões que considere tal aspecto;
- Um cientista de dados deve conscientemente:
    1. não utilizar métodos científicos na realização de seu processo de análise de dados;
    2. não classificar a qualidade dos resultados de maneira razoável e compreensível para o consumidor de dados;
    3. alegar algum resultado fraco ou incerto como um resultado forte;
    4. usar resultados fracos ou incertos para comunicar uma realidade falsa ou promover uma ilusão de compreensão;
    5. não classificar a qualidade dos dados utilizados de forma razoável e compreensível para o consumidor de dados;
    6. inibir a replicação dos resultados obtidos;
    7. desvirtuar os resultados para comunicar uma realidade falsa ou promover uma ilusão de compreensão;

[4] http://orcid.org/0000-0001-9421-3848

Tabela 7: Afirmações para a verificação da confiança do consumidor de dados em um conjunto de dados científico.

Objetivo	Afirmação
Uso: Avaliar a capacidade de julgar como o conjunto de dados foi utilizado pelo provedor de dados.	A documentação sobre a metodologia adotada pelo provedor de dados pode ser facilmente obtida.
	Os elementos de metadados e suas descrições permitem compreender o contexto em que o conjunto de dados foi utilizado.
Criação: Avaliar como o conjunto de dados foi criado.	Compreendo as premissas do provedor de dados para a criação do conjunto de dados.
	Compreendo as transformações ocorridas nos dados ao longo do experimento.
	Os procedimentos adotados pelo provedor de dados para a produção do conjunto de dados estão adequados ao propósito de seu experimento.
Descrição: Avaliar a descrição efetuada sobre o conjunto de dados.	As unidades de medida, símbolos e definições usadas no conjunto de dados se encontram sob um padrão conhecido e estabelecido.
	Consigo compreender os atributos do conjunto de dados.
	O provedor de dados fornece comentários que explicam eventuais valores adversos nos dados.
	Compreendo os direitos de acesso e uso do conjunto de dados.
Reputação: Julgar a reputação do provedor de dados.	O conjunto de dados é proveniente de uma instituição ou pesquisador reconhecido, com o suporte de uma reconhecida instituição de pesquisa.
	Identifico o(s) provedor(es) de dados e demais envolvidos na criação do conjunto de dados.
Impacto: Avaliar a adequabilidade de reutilização do conjunto de dados.	É possível reutilizar o conjunto de dados em novos experimentos.

Fonte: o Autor.

A Tabela 7 apresenta os objetivos elencados para verificar a confiança do consumidor, com as respectivas afirmações a serem avaliadas. A Tabela possui cinco (5) objetivos de validação, os quais estão agrupados em cinco caraterísticas que serão julgadas: (i) uso do conjunto de dados pelo provedor, (ii) criação do conjunto de dados, (iii) descrição do conjunto de dados, (iv) reputação do provedor de dados e (v) a adequabilidade de reutilização do conjunto de dados em novos experimentos. Estas características foram

determinadas após discussão com o grupo de estudos, e teve por referência o trabalho de Ball e Duke (2015), o qual apresenta critérios para determinação do impacto de um conjunto de dados científico para a comunidade científica. No trabalho em questão, os autores apresentam diversos impactos que um conjunto de dados pode trazer à comunidade científica. Dentre os impactos apresentados, encontra-se o reuso do conjunto de dados pela comunidade científica. Adotou-se, nesta avaliação, que o reuso será o impacto mensurado.

É importante ressaltar que o modelo permite que um portal de dados ou instituição de pesquisa estabeleça quais características serão mensuradas, sendo este, portanto, um método que poderá ser adaptado para atender as mais diversas necessidades de avaliação, as quais variam conforme a área de pesquisa ou de aplicação do conjunto de dados.

As assertivas a serem julgadas são, portanto, classificadas em grupos. Os grupos irão avaliar o uso do conjunto de dados pelo provedor de dados (grupo *uso*), como o conjunto de dados foi criado (grupo *cria*), como o conjunto de dados foi descrito (grupo *desc*), a reputação que o provedor de dados possui perante a comunidade científica (grupo *rep*) e o impacto que o conjunto de dados está gerando, no caso a possibilidade de reutilizar o conjunto de dados (grupo *reuso*).

Confiança do Consumidor

Seja $Q_g = \{q_1, q_2, ..., q_n\}$ um conjunto de afirmações que são julgadas por um determinado usuário u_i sobre um *dataset* d, e g é o grupo de características que estão sendo julgadas. No caso, $g = uso, cria, desc, rep, reuso$, representando os grupos de afirmações apresentados na Tabela 7.

a_{ij}^d é a avaliação dada pelo usuário u_i a questão q_j com relação ao dataset d. a_{ij}^d é um número inteiro entre [1,5], baseado na escala *likert* de 5 pontos.

A confiança do usuário u_i sobre o *dataset* d, em relação a um grupo g, denominada $Q_t^g(u_i, d)$, é dada por:

$$Q_t^g(u_i, d) = -0.25 + 0.25 \cdot min_{u_i,d}(Q_g) \tag{3.22}$$

onde $min_{u_i,d}(Q_g)$ é a menor nota dada para um grupo de questões g.

A confiança do consumidor pode ser sumarizada em uma única métrica que consolide as avaliações de cada grupo, denominada $Q_t(u_i, d)$, dada pela moda das notas obtidas para cada Q_t^g.

$$\begin{aligned} Q_t(u_i, d) = moda\big(&Q_t^{uso}(u_i, d), Q_t^{cria}(u_i, d), Q_t^{desc}(u_i, d), \\ &Q_t^{rep}(u_i, d), Q_t^{reuso}(u_i, d)\big) \end{aligned} \tag{3.23}$$

A escolha da moda como medida de tendência central para as avaliações realizadas por um usuário, deu-se pelo fato de que as cinco opções para o usuário (desde discordo totalmente até concordo totalmente) são tipos de dados ordinais que, embora estejam em uma certa ordem, a distância entre seus elementos não é conhecida. Por exemplo, é difícil especificar qual o significado da média entre *neutro* e *concordo totalmente*. A moda, por sua vez, permite obter o valor avaliação que foi escolhido com maior frequência pelo consumidor (BOONE; BOONE, 2012).

3.2.4 Índice de Veracidade de conjuntos de dados científicos

Esta seção apresenta a proposta para a criação do *índice de veracidade* para conjuntos de dados científicos, fazendo uso das métricas de aspectos intrínsecos ao conjunto de dados, juntamente com a métrica de aspecto social. O índice de veracidade será utilizado como mecanismo de mensuração da incorporação da dimensão de Veracidade a um ciclo de

gestão de dados, o qual irá gerar, dentre outros artefatos, um ou mais conjuntos de dados.

Ao longo da elaboração do índice de veracidade, buscou-se atender as características fundamentais de uma métrica, estabelecidas por Frank e Walker (2016), os quais afirmam que uma métrica deve ser **válida** (fortemente correlacionada com o atributo de interesse), **confiável** (fornece valores consistentes ao longo do tempo e entre múltiplos observadores), **eficiente** (não requer uma quantidade considerável de tempo e recursos para seu cálculo), **transferível** (a mesma métrica pode ser utilizada em uma variedade de contextos) e **comparável** (uma métrica é comparável quando permite uma comparação padronizada sob múltiplos contextos, com o potencial de se tornar um padrão para a mensuração do atributo de interesse).

As métricas dos aspectos intrínsecos aos conjuntos de dados são calculadas automaticamente por meio de um processo de *parsing* de um arquivo de metadados no formato *linked data* (RDF/XML ou JSON-LD, por exemplo). Este formato de dados já se encontra sendo utilizado por portais de dados, o que permitirá uma adesão ao modelo proposto sem muitas dificuldades técnicas. Já a métrica de *confiança do consumidor*, de aspecto social, é calculada por meio de uma questionário respondido pela comunidade científica.

3.2.4.1 Cálculo do Índice de Veracidade

O índice de Veracidade será composto por duas perspectivas, quais sejam:

1. Perspectiva intrínseca do conjunto de dados - representa os valores das métricas dos aspectos intrínsecos do conjunto de dados;
2. Perspectiva social do conjunto de dados - representa os valores das métricas sociais obtidos pela avaliação da comunidade científica.

Optou-se por estruturar o índice de veracidade em duas perspectivas, uma vez que as métricas de avaliação intrínseca do conjunto de dados sofrem menos atualizações quando comparadas à métrica de aspecto social. Além disso, tais métricas são independentes da área de pesquisa sob a qual o conjunto de dados está sendo avaliado. A métrica social, por sua vez, será calculada por área de pesquisa, sendo atualizada a cada nova avaliação sobre o conjunto de dados. A atualização dos valores intrínsecos do conjunto de dados pode ser feita de maneira periódica, para garantir principalmente que o conjunto de dados se mantenha disponível. Se, porventura, o provedor de dados atualizar algum elemento de descrição associado a tais métricas, estas também podem ser atualizadas, refletindo assim o estado atual do conjunto de dados.

3.2.4.1.1 Cálculo da Perspectiva Intrínseca do Conjunto de Dados

Sejam $Q_c, Q_d, Q_{da}, Q_m, Q_p$ as métricas intrínsecas para um conjunto de dados, apresentadas na Seção 3.2.2.

A Avaliação Intrínseca do Conjuntos de Dados, denominada por $V_{dc}(d)$ é dada por:

$$V_{dc}(d) = min(Q_c, Q_d, Q_{da}, Q_m, Q_p) \tag{3.24}$$

V_{dc} apresenta um valor numérico entre 0 e 1. A adoção pelo menor valor das métricas se deu pelo fato de que a ausência de elementos de descrição necessários para uma das métricas compromete a capacidade de reuso dos dados. Por exemplo, no caso de indisponibilidade dos recursos de um conjunto de dados, Q_d receberia valor zero. De fato, se um conjunto de dados estiver indisponível, de nenhum valor este será para a Veracidade, impossibilitando o reuso do mesmo. Dessa forma, a avaliação intrínseca do conjunto de dados em relação a veracidade receberá valor zero.

3.2.4.1.2 Cálculo da Perspectiva Social do Conjunto de Dados

Seja I_d o conjunto de avaliações sociais sobre um conjunto de dados d, de modo que $I_d = \{Q_t(u_i, d), Q_t(u_j, d), \ldots, Q_t(u_z, d)\}$, onde u_i, u_j *e* u_z são consumidores que avaliaram d e, consequentemente, $Q_t(u_i, d)$ é, no caso, a confiança do consumidor u_i sobre d, sumarizada conforme a Equação 3.23.

Embora o algoritmo de reputação RBPD denomine os avaliadores como usuários, adotar-se-á o termo consumidor, para simbolizar o papel de um pesquisador que avalia um conjunto de dados. Todavia, os termos usuário e consumidor possuem ambos o mesmo sentido e devem ser entendidos como sinônimos.

$fu(a_{id})$ é o fator de decaimento temporal (*aging*) sobre a avaliação $Q_t(u_i, d)$, a ser calculado conforme equação (3.19). Neste contexto, R_i é a reputação do consumidor i na rede de avaliações.

Calcula-se o valor médio $avg(g)$ para cada grupo g, para todos $a_{id} \in I_d$, da seguinte forma:

$$avg(g) = \frac{\sum_{a_i \in I_d} Q_t^g \cdot R_i \cdot fu(a_{id})}{|I_d|} \tag{3.25}$$

onde $g = uso, cria, desc, rep, reuso$.

A equação (3.25) calcula o valor médio de cada grupo de questões sobre a confiança do consumidor, ponderando-as pela reputação do consumidor e pelo fator de decaimento temporal.

$Avg_d = \{avg(uso), avg(cria), avg(desc), avg(rep), avg(reuso)\}$ irá armazenar os valores médios para cada grupo. Dessa forma, a perspectiva social do índice de veracidade de d, $V_s(d)$, é calculada conforme segue:

$$V_s(d) = \frac{\sum_{m \in Avg_d}(m)}{|Avg_d|}, \tag{3.26}$$

$V_s(d)$ apresenta um valor numérico entre 0 e 1.De modo a exemplificar e analisar o índice proposto, apresenta-se dois cenários de cálculo do índice de veracidade.

O primeiro cenário, cenário A, é formado por um único conjunto de dados sendo avaliado por dois consumidores. O índice de veracidade será influenciado pelo decaimento exponencial em relação ao tempo de cada avaliação e pela baixa reputação dos consumidores, em virtude do baixo número de avaliações disponíveis na rede de avaliação.

O segundo cenário, cenário B, apresenta uma rede composta por cinco consumidores e cinco conjuntos de dados, de modo a verificar o efeito da rede de avaliações na determinação da reputação dos consumidores.

3.2.4.1.3 Cenário A - Um conjunto de dados, duas avaliações

A Tabela 8 apresenta as métricas de avaliação sobre o conjunto de dados d juntamente com a avaliação de confiança de dois consumidores u_1 e u_2. Nessa tabela, a coluna *Age* indica o tempo de avaliação em relação ao tempo atual de cálculo do índice de veracidade. No exemplo em questão, o consumidor u_1 avaliou d há 50 unidades de tempo (dias ou semanas, por exemplo), enquanto o consumidor u_2 avaliou há 60 unidades de tempo.

Tabela 8: Cenário A - Métricas para cada perspectiva do índice de Veracidade para um *dataset d*.

Perspectiva Intrínseca	Q_c	Q_d	Q_{da}	Q_m	Q_p		
	0,8	0,7	0,65	0,8	1,0		
Perspectiva Social	**Age**	**Consumidor**	Q_t^{uso}	Q_t^{cria}	Q_t^{desc}	Q_t^{rep}	Q_t^{reuso}
	50	u_1	0,5	0,75	0,75	0,25	0,75
	60	u_2	0,75	0,5	0,75	0,5	0,5

Fonte: o Autor.

Perspectiva Intrínseca do Conjunto dos Dados

A perspectiva intrínseca do conjunto de dados é calculada conforme a Equação 3.24. O seu valor é dado por:

$$V_{dc}(d) = min(Q_c, Q_d, Q_{da}, Q_m, Q_p)$$
$$V_{dc}(d) = min(0.8, 0.7, 0.65, 0.8, 1.0)$$
$$V_{dc}(d) = 0,65$$

Nesta situação, o menor valor foi obtido pela métrica de *disponibilidade dos agentes* (0.65). Assim, a perspectiva intrínseca da Veracidade do conjunto de dados possui o valor 0,65.

Perspectiva Social

A avaliação dos consumidores u_1 e u_2 sobre d pode ser visualizada com o auxílio de um gráfico radial, apresentado na Figura 18.

O cálculo do índice de veracidade se inicia pela determinação do fator de *aging* a ser aplicado nas avaliações, juntamente com a reputação de cada consumidor.

Para a determinação do fator de *aging*, deve-se inicialmente estabelecer o tempo de meia-vida das avaliações na rede de avaliação. Para o cálculo do tempo de meia-vida será adotada, conforme proposto por Larrain et al. (2015), a mediana de *Age*. No exemplo em questão $T_m = mediana(50, 60) = 55$.

Uma vez determinado o tempo de meia-vida, calcula-se o valor de λ, conforme a equação (3.20):

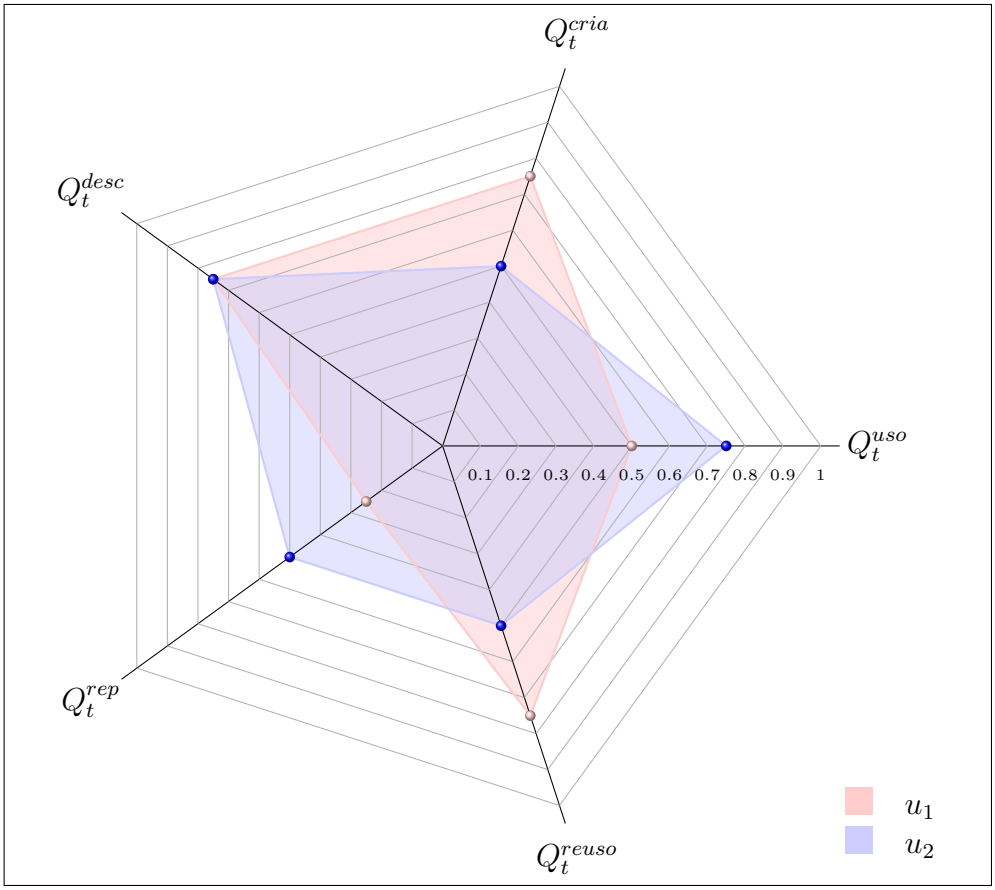
$$\lambda = \frac{ln\ 2}{T_m} = \frac{ln\ 2}{55} = 0,0126$$

Com o valor de λ determinado, será calculado o fator de *aging* de cada avaliação, com base na equação (3.19):

$$f_{u1}(t = 50) = e^{-0.01226 \cdot (50)} = 0,5325$$
$$f_{u2}(t = 60) = e^{-0.01226 \cdot (60)} = 0,4695$$

Conforme era esperado, a avaliação do consumidor u_2 teve um peso menor quando comparada à avaliação mais recente do consumidor u_1.

Figura 18: Gráfico Radial para a Confiança dos Consumidores do Conjuntos de Dados d.



Fonte: o Autor.

Considera-se que o algoritmo de reputação RBPD* será executando com base na rede de avaliações formada pelas avaliações apresentadas na Tabela 8 (coluna Q_t). Para fins práticos, não será demonstrada aqui a execução passo-a-passo do algoritmo. A reputação calculada pelo algoritmo será de $0,5$ para ambos os consumidores.

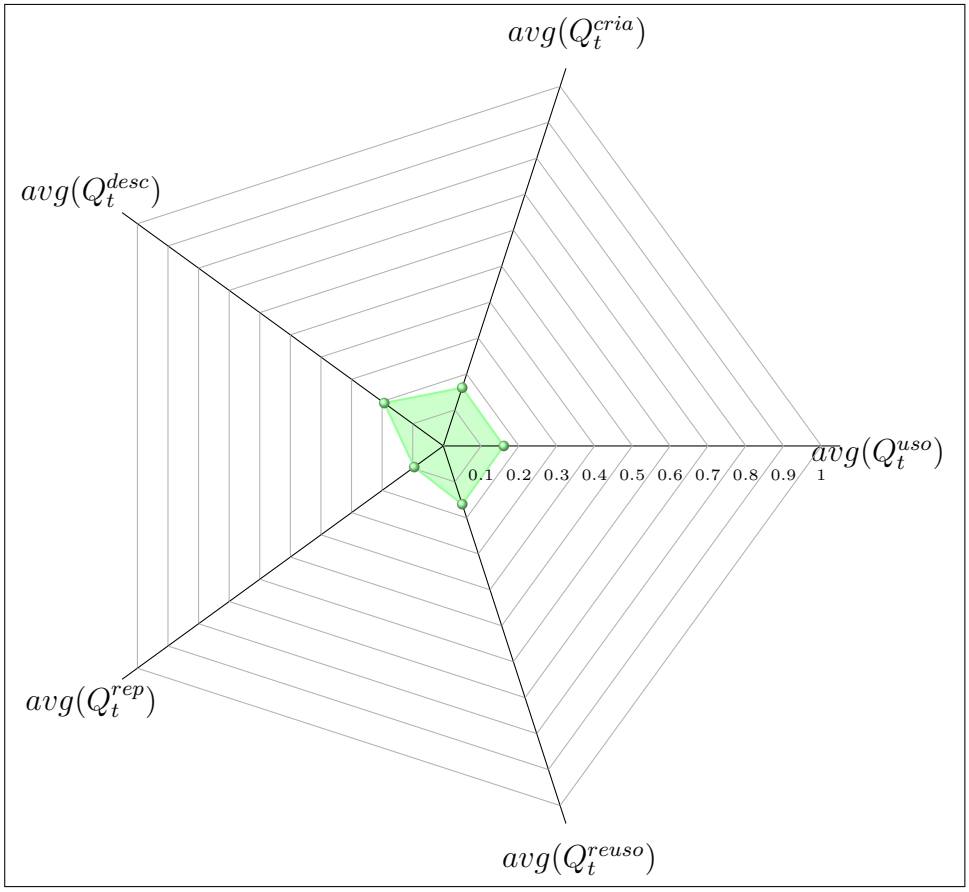
Uma vez determinados o fator de *aging* e a reputação de cada consumidor, calculam-se os valores médios de cada grupo de avaliação social, conforme equação (3.25). O resultado é apresentado na Tabela 9.

Tabela 9: Cenário A - Valores médios das métricas de apoio à veracidade para a perspectiva social, já considerando o decaimento temporal das avaliações e a reputação dos consumidores.

$avg(Q_t^{uso})$	$avg(Q_t^{cria})$	$avg(Q_t^{desc})$	$avg(Q_t^{rep})$	$avg(Q_t^{reuso})$
0,160	0,162	0,193	0,095	0,162

A Tabela 9 contém os componentes que irão permitir o cálculo do índice de veracidade de d. Propõe-se o uso do gráfico radial para uma apresentação visual do índice. O gráfico radial que representa a Tabela 9 se encontra apresentado na Figura 19.

Figura 19: Gráfico Radial do Índice de Veracidade de d - Dimensal Social. Os valores foram ajustados em função da reputação dos usuários e do *aging* de cada avaliação efetuada.



Fonte: o Autor.

Com base na Tabela 9, o índice de veracidade de d, em sua perspectiva social, é calculado conforme equação (3.26).

$$V_s(d) = \frac{\sum_{m \in Avg_d}(m)}{|Avg_d|}$$

$$V_s(d) = \frac{avg(Q_t^{uso}) + avg(Q_t^{cria}) + avg(Q_t^{desc}) + avg(Q_t^{rep}) + avg(Q_t^{reuso})}{5}$$

$$V_s(d) = \frac{0,160 + 0,162 + 0,193 + 0,095 + 0,162}{5}$$

$$V_s(d) = 0,154$$

O valor do índice de veracidade é de 0,154 pelo fato de que as duas avaliações existentes já ultrapassaram o tempo de meia-vida determinado, além do fato de que no exemplo em questão, os consumidor u_1 e u_2 avaliam apenas um único conjunto de dados, de modo que a sua reputação ficou restrita a 50%. É preciso, portanto, que novas avaliações sejam inseridas na rede, para que a veracidade do conjunto de dados possa aumentar. As novas avaliações serão menos influenciadas pelo fator de *aging*, assim como permitirão que a aferição das reputações se torne mais apurada.

3.2.4.1.4 Cenário B - Rede de Avaliação

Para efeitos de simplificação, considerar-se-á que as métricas Q_c, Q_d, Q_{da}, Q_m e Q_p (perspectiva intrínseca) possuem valor igual a 1, além disso, não será considerado o efeito de *aging*, simulando uma rede de avaliações sendo calculada pela primeira vez, com todas as avaliações criadas no mesmo instante. Dessa forma, busca-se observar o efeito da reputação do consumidor em uma rede de avaliações, na determinação do índice de veracidade de um conjunto de *datasets*, em sua perspectiva social.

A matriz $\mathbf{Q_t}$ apresenta as avaliações de confiança de cinco consumidores $(u_1, u_2, u_3, u_4, u_5)$ sobre cinco conjuntos de dados $(d_1, d_2, d_3, d_4, d_5)$.

$$\mathbf{Q_t} = \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} \\ \begin{bmatrix} 0.7 & 0.5 & 0.5 & - & - \\ 0.4 & - & - & 0.3 & - \\ 0.4 & - & - & - & 0.3 \\ - & 0.5 & 0.5 & 0.6 & 0.6 \\ - & 0.5 & 0.5 & 0.6 & 0.6 \end{bmatrix} \end{array} \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array}$$

A execução do algoritmo RBPD* para o conjunto de avaliações de confiança, resulta nas seguintes reputações sobre cada consumidor:

$$\mathbf{R_{u_i}} = \begin{array}{cc} \begin{bmatrix} 0.8 \\ 0.5 \\ 0.5 \\ 0.83 \\ 0.83 \end{bmatrix} & \begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array} \end{array}$$

É de se observar que quanto mais avaliações um consumidor realizar na rede, e estas avaliações estiverem de acordo com a maioria das outras avaliações sobre o mesmo objeto, maior será a reputação do consumidor. No caso, os consumidores que menos avaliaram os conjuntos de dados foram u_2 e u_3 recebendo, consequentemente, a menor reputação da rede.

No cenário proposto, uma vez que não há efeito de *aging* a ser considerado, o item de avaliação para cada avaliação existente será calculado como o produto da avaliação $a_{i\gamma}$ efetuada pelo consumidor u_i e a sua respectiva reputação R_{u_i}. O resultado obtido encontra-se apresentado na matriz **I**.

$$\mathbf{I} = \begin{array}{cc} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} & \\ \begin{bmatrix} 0.56 & 0.4 & 0.4 & - & - \\ 0.2 & - & - & 0.15 & - \\ 0.2 & - & - & - & 0.15 \\ - & 0.415 & 0.415 & 0.498 & 0.498 \\ - & 0.415 & 0.415 & 0.498 & 0.498 \end{bmatrix} & \begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array} \end{array}$$

O índice de veracidade para cada *dataset* será calculado como a média dos itens de avaliação recebidos por um *dataset*, obtendo os seguintes valores:

$$\mathbf{V_{d_i}} = \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} \\ \begin{bmatrix} 0.32 & 0.41 & 0.41 & 0.382 & 0.382 \end{bmatrix} \end{array}$$

Nesta situação, observa-se que dentre os conjuntos de dados, d_2 e d_3 apresentam a maior veracidade. Isto se deve ao fato que os dois consumidores com maior reputação na rede avaliaram tais conjuntos de dados. O efeito esperado para o índice de veracidade é que ele saiba lidar com avaliações tendenciosas, como consumidores que sempre avaliam com as maiores notas, ou o contrário, fornecendo sempre a menor nota possível. O fato

de que quanto maior a participação na rede do consumidor, e suas avaliações estarem alinhadas com as avaliações de outros consumidores permite evitar cenários em que um conjunto de dados é avaliado por poucos consumidores com notas muito altas, e sua veracidade seja considerada alta, mesmo embora com poucas avaliações para confirmar este valor.

3.2.4.2 Análise da Robustez do Índice de Veracidade

De acordo com Hubbard (2010.) uma métrica é considerada robusta quando é capaz de minimizar o impacto de valores extremos em seus parâmetros. Para testar este cenário, será considerada uma rede de avaliações em que há um consumidor, u_1, o qual avaliou todos os conjuntos de dados com a maior nota possível da escala adotada (no caso, 1.0). A avaliação da confiança (Q_t) dos consumidores na rede com relação a cinco conjuntos de dados é apresentada na matriz Q_t apresentada a seguir. Para efeitos de simplificação, será demonstrando o efeito da reputação sob a métrica sumarizada Q_t, uma vez que o comportamento será análogo para as métricas individuais.

$$\mathbf{Q_t} = \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} \\ \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 0.5 & - & - & 0.75 & - \\ 0.5 & - & - & - & 0.5 \\ - & 0.5 & 0.5 & 0.75 & 0.75 \\ - & 0.5 & 0.5 & 0.75 & 0.75 \end{bmatrix} \end{array} \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array}$$

A determinação da reputação dos consumidores nesta rede, calculada pelo algoritmo RBPD*, deverá refletir a robustez esperada, caracterizada pelo consumidor u_1 recebendo uma reputação baixa, uma vez que suas avaliações não concordam com a maioria das outras avaliações. A reputação calculada encontra-se apresentada na matriz R_{u_i}, conforme abaixo.

$$\mathbf{R_{u_i}} = \begin{bmatrix} 0.43 \\ 0.62 \\ 0.50 \\ 0.58 \\ 0.58 \end{bmatrix} \begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array}$$

Observa-se, portanto, que o consumidor u_1 recebeu a menor reputação, conforme o esperado. Como exemplo, considere-se apenas o conjunto de dados d_1. Este conjunto recebeu as seguintes avaliações:

Tabela 10: Avaliações sobre d_1.

Consumidor	Q_t^{uso}	Q_t^{cria}	Q_t^{desc}	Q_t^{rep}	Q_t^{reuso}	Q_t
u_1	1,0	1,0	1,0	1,0	1,0	1,0
u_2	0,25	0,5	0,5	0,75	0,5	0,5
u_3	0,5	0,25	0,5	0,5	0,75	0,5

De posse da reputação de cada usuário, calcula-se o valor do índice de veracidade do conjunto de dados, em sua perspectiva social, conforme apresentado na Tabela 11.

Observa-se que as avaliações do consumidor u_1 não influenciaram consideravelmente a veracidade de d_1. Por exemplo, efetuando-se uma média simples nas avaliações recebidas por d_1 na matriz Q_t, obtém-se o valor 0.6. A sua veracidade é, no entanto, 0.330, pelo fato de que os consumidor que avaliaram tal conjunto de dados apresentam os menores valores de reputação da rede.

A Figura 20 apresenta proposta para visualização do índice de veracidade de um conjunto de dados. Na Figura, considera-se que o índice de Veracidade de d_1, na perspectiva intrínseca ao conjunto de dados, será 1.0. Nela é possível observar as avaliações efetuadas por cada pesquisador membro da rede de avaliações, o fator de *aging* de cada avaliação e a reputação do pesquisador, a qual será considerada na determinação do índice de veracidade do conjunto de dados.

3.2.4.3 Análise do Comportamento do Índice de Veracidade em função do crescimento da rede de avaliações

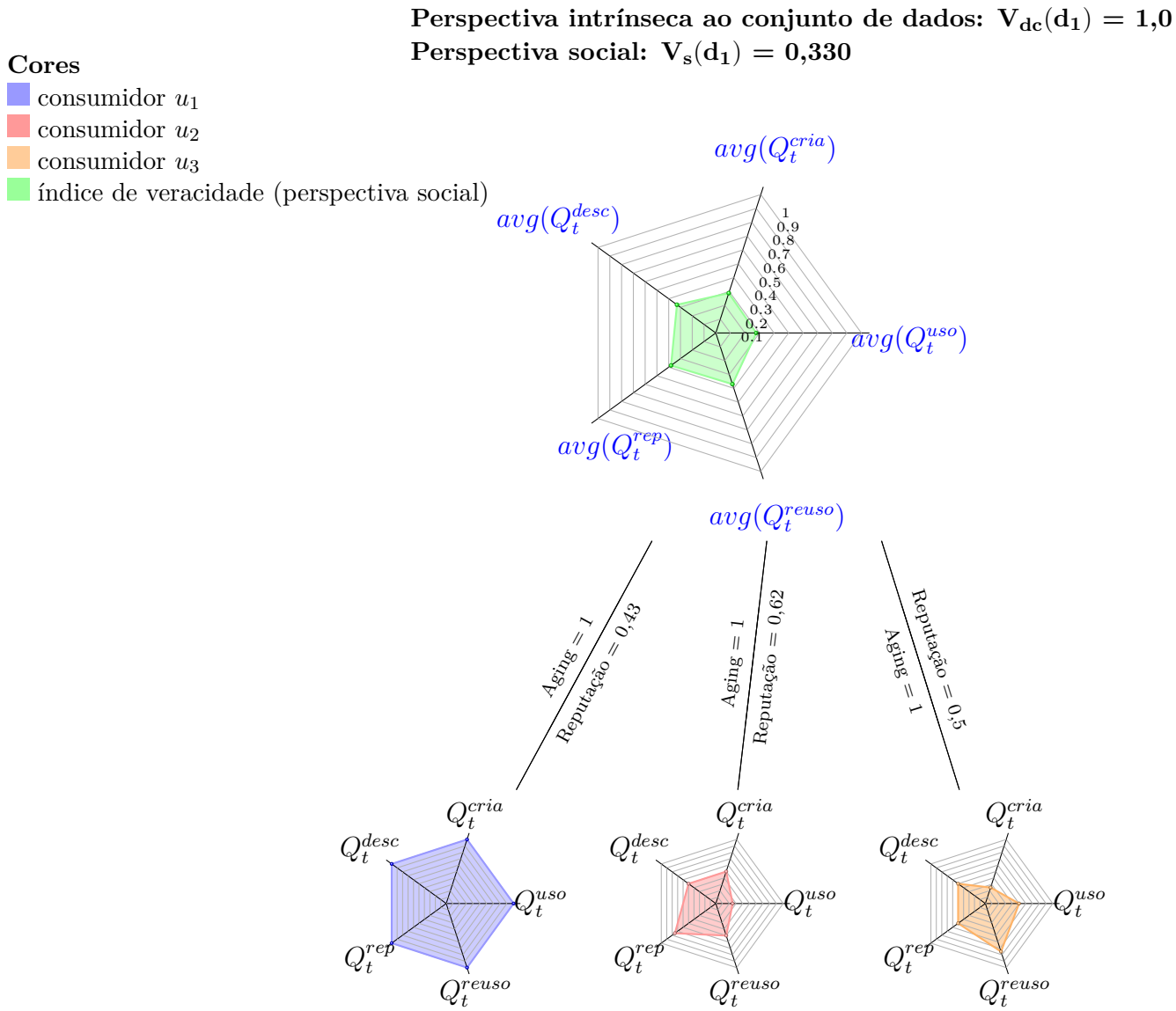
Ao longo da elaboração do índice de veracidade, buscou-se determinar como se dá a sua variação ao longo do crescimento da rede de avaliações.

Para fins de simplificação da análise, desconsiderou-se o decaimento temporal, uma

Tabela 11: Cálculo do Índice de Veracidade de Dados, em sua perspectiva social, para d_1.

Consumidor	$Q_t^{uso} \cdot R$	$Q_t^{cria} \cdot R$	$Q_t^{desc} \cdot R$	$Q_t^{rep} \cdot R$	$Q_t^{reuso} \cdot R$	Reputação (R)
u_1	0,430	0,430	0,430	0,430	0,430	0,43
u_2	0,155	0,310	0,310	0,465	0,310	0,62
u_3	0,250	0,125	0,250	0,250	0,375	0,50
Média	0,278	0,288	0,330	0,382	0,372	
$V_s(d)$	0,330					

Figura 20: Proposta de visualização do Índice de Veracidade do conjunto de dados d_1.



Fonte: o Autor.

vez que seu comportamento já se encontra descrito na literatura (LARRAIN et al., 2015). As métricas de apoio à veracidade foram também consideradas apresentando seu valor máximo, na premissa de que todos os requisitos do perfil de aplicação se encontram satisfeitos.

Em um primeiro momento, projeta-se uma rede de avaliações formada por um único consumidor e um único conjunto de dados. O cálculo da veracidade, em sua perspectiva social, de tal conjunto se dá, conforme apresentado na Equação 3.26, como segue:

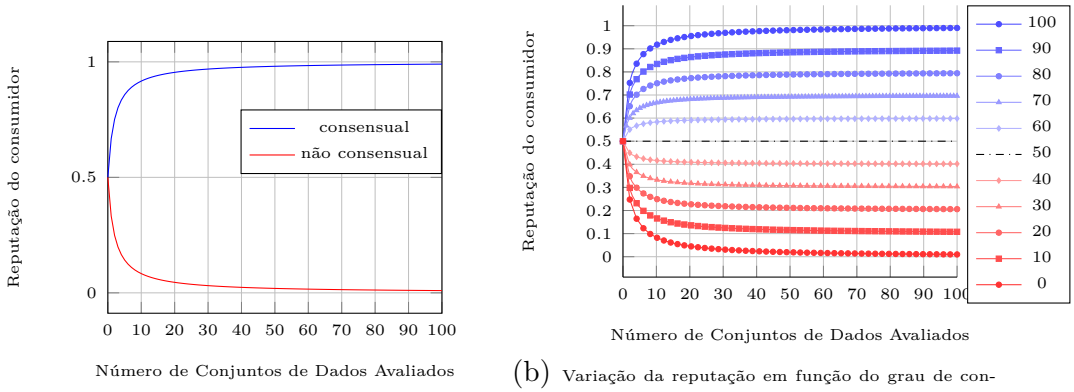
$$V_s(d) = \frac{\sum_{m \in Avg_d}(m)}{|Avg_d|} \tag{3.27}$$

Para o cenário de um único consumidor, o valor da veracidade será ponderado pelo valor da reputação do consumidor. Como visto no Cenário A, em uma rede de avaliações em que um consumidor avalia apenas um único conjunto de dados, o valor da reputação calculado pelo algoritmo RBPD* será 0,5.

Com a expansão da rede, a veracidade de um conjunto de dados se dará em função da média da reputação dos consumidores que avaliaram tal objeto. Dessa forma, buscou-se analisar o comportamento da reputação de um avaliador conforme aumentam-se o número de objetos por este avaliados.

Para tal análise, projetou-se dois comportamentos para os consumidores: consumidor consensual e consumidor não consensual. O consumidor consensual é aquele que apresenta todas as suas avaliações classificadas como justas. Já o não consensual, por sua vez, é aquele em que todas as suas avaliações são classificadas como injustas. Justa e injusta, conforme apresentado na Seção 3.2.3.1, são os termos utilizados pelo algoritmo RBDP para determinar o quanto uma avaliação sobre um objeto concorda com a maioria das demais avaliações que o objeto recebeu.

Figura 21: Análise da reputação de um consumidor em função do crescimento da rede.



(a) Reputação para o caso extremo: consensual e não consensual.

(b) Variação da reputação em função do grau de consenso de um consumidor em relação às demais avaliações. Legenda: percentual de avaliações consensuais.

A Figura 21a ilustra a simulação do crescimento da rede de avaliações, na qual o número de conjuntos avaliados por um consumidor varia entre 0 e 100. Quanto maior o número de objetos avaliados por um consumidor consensual, maior será a sua reputação, e consequentemente, maior será o fator de ponderação para a determinação da veracidade dos conjuntos de dados.

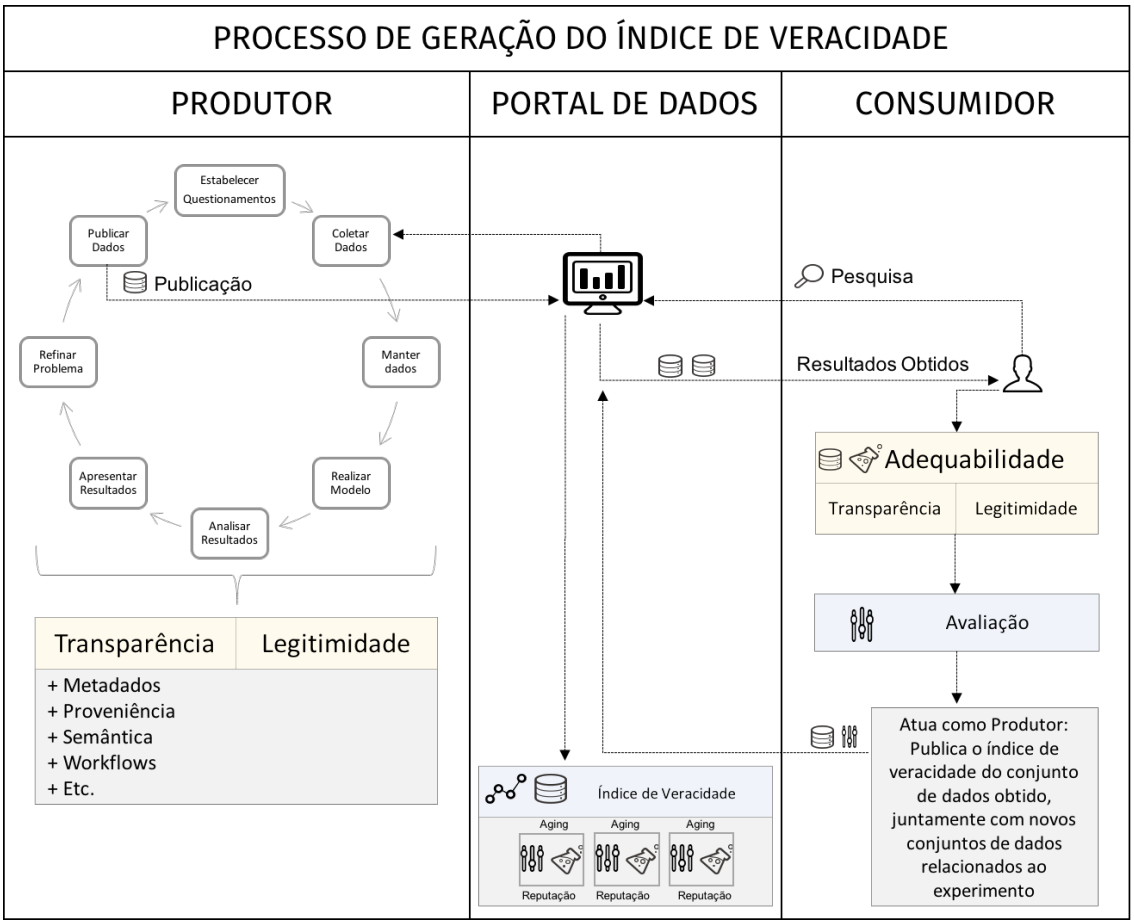
No caso de um conjunto de dados que foi avaliado por múltiplos consumidores, o comportamento será análogo, uma vez que, conforme a Equação 3.25, a veracidade será dada pela média das reputações dos consumidores. Dessa forma, extrapolando a análise

para um cenário em que múltiplos consumidores avaliam diversos conjuntos de dados em uma rede de avaliações, a veracidade de um conjunto de dados crescerá em função de quanto mais consenso houver com relação à métrica de *confiança do consumidor* (Q_t). A Figura 21b apresenta a variação da reputação de um consumidor na rede de avaliações, em função do percentual de consenso de suas avaliações em relação as demais avaliações existentes na rede.

3.2.5 Incorporação do Índice de Veracidade ao Ciclo de Vida dos Dados

A Figura 22 apresenta uma visão macro do processo de geração do índice de veracidade. Nela há três papeis sendo apresentados: PRODUTOR, CONSUMIDOR e PORTAL DE DADOS.

Figura 22: Proposta de processo macro de geração do índice de veracidade e sua aplicação em um portal de dados científicos.



Fonte: o Autor.

Na abstração proposta, inicialmente há um processo de produção de um conjunto

de dados pelo PRODUTOR para publicação em um PORTAL DE DADOS. Considera-se que, ao longo do processo científico, o PRODUTOR irá fazer uso de elementos tais como metadados, proveniência, anotações semânticas, *workflows*, dentre outros, possibilitando assim a captura dos elementos estabelecidos no perfil de aplicação para a veracidade, favorecendo, portanto, os aspectos de TRANSPARÊNCIA e de LEGIMITIDADE da Veracidade. A legitimidade e transparência serão elementos capturados pela perspectiva intrínseca ao conjunto de dados do índice de veracidade. Dessa forma, o PORTAL DE DADOS pode fazer uso dessa dimensão para determinar critérios de aceitação da publicação de conjuntos de dados.

Uma vez que o conjunto de dados se encontra publicado no PORTAL DE DADOS um outro pesquisador, denominado CONSUMIDOR, poderá encontrá-lo. A busca pelo conjunto de dados será otimizada, uma vez que para efetuar as descrições sobre tal conjunto, o PRODUTOR fez uso de formatos de dados compreensíveis por máquina.

Ao obter o conjunto de dados e os elementos de representação da TRANSPARÊNCIA e LEGITIMIDADE, o CONSUMIDOR irá realizar o processo de análise da ADEQUABILIDADE do conjunto de dados para o experimento que pretende desenvolver, alimentando assim a perspectiva social do conjunto do índice de veracidade.

A análise de adequabilidade se dá pela métrica de Confiança do Consumidor (Q_t). Uma vez realizada a avaliação do conjunto de dados pelo CONSUMIDOR, tal avaliação irá compor a rede de avaliações do conjunto de dados, momento no qual será calculada a reputação do CONSUMIDOR. Com base na reputação, assim como no *aging* da avaliação, determinar-se-á o índice de veracidade do conjunto de dados, em sua perspectiva social. O índice de veracidade será armazenado no PORTAL DE DADOS juntamente com a área de pesquisa ou de aplicação a qual este índice se refere.

Como mecanismo de atualização do índice de veracidade do conjunto de dados, o PORTAL DE DADOS atualizará a perspectiva social do índice a cada inclusão de avaliação por um CONSUMIDOR, para a área de pesquisa em questão, momento no qual os valores de reputação e *aging* de toda a rede de avaliações serão atualizados. O PORTAL pode, também, estabelecer uma periodicidade de revisão do índice, atualizando periodicamente a perspectiva intrínseca ao conjunto de dados, de modo que o *aging* possa ser aplicado nas avaliações anteriores.

É importante ressaltar que ambos os papeis de PRODUTOR e CONSUMIDOR podem ser exercidos por um mesmo pesquisador e efetuados ao mesmo tempo. Neste momento, exercendo a função de PRODUTOR, o pesquisador irá executar a sua pesquisa, juntamente

com a publicação do conjunto de dados.

Como exemplo, seja d um conjunto de dados de ocorrência geográfica de uma determinada espécie produzido por um instituto de pesquisa, e dois experimentos e_1 e e_2, ambos fazendo uso de d, executados, respectivamente, pelos pesquisadores p_1 e p_2. Ambas as avaliações dos pesquisadores estarão disponíveis no portal de dados, as quais irão compor o índice de veracidade de d para a área de pesquisa de *modelagem de distribuição de espécies*. Quando um outro pesquisador, p_3, for pesquisar por conjuntos de dados no portal de dados ligados à área de *modelagem de distribuição de espécies*, atuando então como CONSUMIDOR, poderá conhecer o índice de veracidade do conjunto de dados para esta área.

O pesquisador p_3 poderá, portanto, fazer uso de funções de análise disponíveis no portal, para conhecer cada avaliação realizada por outros pesquisadores, possibilitando assim o estabelecimento de critérios para seleção dos dados. Considerando que p_3 fez uso do conjunto de dados para seu experimento, ele também irá fornecer uma avaliação sobre o conjunto, passando a fazer parte da rede de avaliações do conjunto de dados, atualizando, portanto, o valor do índice de veracidade para d.

4 ESTUDO DE CASO

Este capítulo apresenta e discute os resultados do estudo de caso desenvolvido para verificação do perfil de aplicação proposto para suportar a dimensão de Veracidade.

O estudo de caso foi desenvolvido com base nas experiências e colaborações desenvolvidas no *II Workshop em Ciência dos Dados*, realizado na Escola Politécnica da Universidade de São Paulo. Neste evento, os pesquisadores Prof. Dr. Paulo Eduardo Artaxo Neto e a Profa. Dra. Luciana Varanda Rizzo disponibilizaram conjuntos de dados relacionados ao projeto GoAmazon (MARTIN et al., 2017).

4.1 Descrição do problema científico e dos dados

O projeto GoAmazon coletou dados entre os anos de 2014 e 2015 na busca pela compreensão sobre como o ciclo de vida de aerossóis e nuvens é influenciado pelos poluentes da cidade de Manaus em meio a floresta tropical (MARTIN et al., 2017).

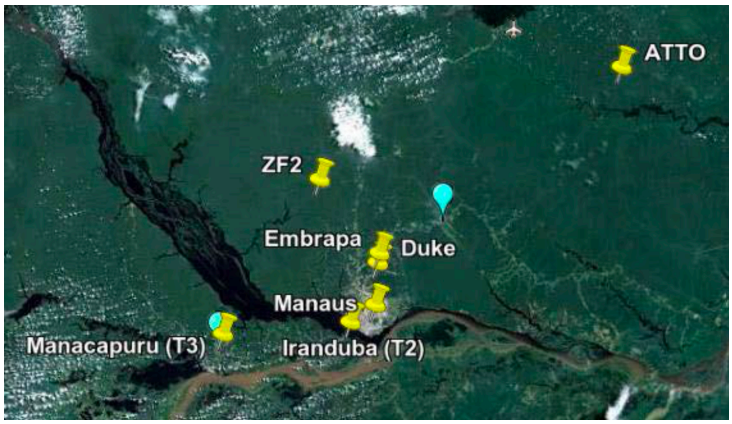
Manaus é uma cidade que possui 2 milhões de habitantes e uma frota veicular de aproximadamente 700 mil veículos, cercada por milhares de quilômetros de floresta, criando uma área urbana isolada dentro da Bacia Amazônica (HOLANDA, 2015).

No ecossistema Amazônico ocorrem interações entre a biosfera e a atmosfera, com impactos no ciclo hidrológico, nos ciclos biogeoquímicos, e na composição química da atmosfera. Por receber altas taxas de incidência solar, e sua atmosfera ser rica em vapor de água, a atmosfera da Amazônia pode ser considerada um "grande reator químico" (RODRIGUES, 2015).

O estudo dos processos atmosféricos nessa região é importante para avaliações regionais e globais de mudanças climáticas. Em todos os locais de coleta do GoAmazon, um conjunto de medidas das propried ades dos aerossóis foi coletado, tais como: espalhamento e absorção espectral da luz; distribuição de tamanho; composição química; componentes orgânicos e inorgânicos; profundidade ótica, entre outras medidas (HOLANDA, 2015).

A Figura 23 apresenta a localização das principais estações de amostragem do projeto GoAmazon. Os dados utilizados neste estudo de caso foram coletados na região de Manacapuru, Amazonas, local do sítio T3.

Figura 23: Localização das principais estações de amostragem do projeto GoAmazon.



Fonte: (SILVA et al., 2015).

Ao longo do *II Workshop* em Ciência dos Dados, foi disponibilizado um conjunto de dados processado pelo Laboratório de Física Aplicada do IFUSP, com dados referentes a concentração de Ozônio (O_3) no sítio T3 do projeto GoAmazon. T3 foi escolhido estrategicamente vento abaixo de Manaus, no curso preferencial da pluma urbana de aerossóis e gases emitidos pela cidade. A Figura 24 apresenta os equipamentos utilizados nas instalações do sítio de monitoramento T3.

Figura 24: Sítio T3 do Projeto GoAmazon.

Fonte: <https://www.arm.gov/images/cms/ARM-Mobile-Facility-in-Brazil-e1449472255314.jpg>.

A camada de O_3 estratosférica é benéfica para a vida na Terra, uma vez que impede parte da radiação ultravioleta (UV) de atingir a superfície o planeta. Em contrapartida, o O_3 presente na troposfera é considerado um poluente de alta periculosidade, devido a seus impactos na saúde humana e no ambiente. A inalação de O_3 pode causar sérios impactos

Tabela 12: Descrição dos Dados disponíveis no Conjunto de Dados.

Coluna	**Descrição**
Time (UTC)	data e hora no formato UTC
Unix Time (UTC)	data e hora no formato UTC, segundos desde 01/01/1970
O3	Concentração de O_3

Fonte: o Autor.

à saúde, sobretudo no sistema respiratório. Ele pode oxidar diretamente as células das paredes pulmonares e vias aéreas ou reagir com antioxidantes que fazem parte de defesa do organismo, o que pode acarretar em um processo inflamatório (RODRIGUES, 2015). A quantidade de O_3 que causa danos à saúde varia para cada organismo, mas atmosferas com razões de mistura a partir de 50 ppb (partes por bilhão) já são consideradas de alta periculosidade (WHO, 2005). A União Europeia, por meio do programa *Clear Air for Europe*, definiu que níveis de O_3 acima de 35 ppb são considerados inseguros para a saúde humana (FOWLER et al., 2008).

O O_3 é considerado um poluente secundário, uma vez que sua formação depende de óxidos de nitrogênio (NO_x), monóxido de carbono (CO), compostos orgânicos voláteis (COVs) e da incidência de radiação solar (RODRIGUES, 2015).

Os dados da concentração de O_3 foram coletados por meio do Analisador Fotométrico UV de Ozônio, modelo *49i* da *Thermo Fisher Scientific*.

Em T3, o aparelho foi instalado a 5 metros de altura. Seu limite inferior de detecção é de 1 ppb (partes por bilhão). As observações foram realizadas em intervalos de 5 minutos. Dois arquivos de séries temporais foram criados: o primeiro deles apresenta a medição a cada 5 minutos, enquanto o segundo apresenta a média da concentração a cada 30 minutos (RODRIGUES, 2015).

O conjunto de dados foi disponibilizado no endereço eletrônico <http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/>. A Tabela 12 apresenta os dados disponíveis, juntamente com uma breve descrição de cada coluna do conjunto de dados.

Juntamente com os dados, os pesquisadores disponibilizaram arquivos de apoio, de modo a subsidiar o entendimento do conjunto de dados e dos procedimentos que geraram tal conjunto. Com base em tais documentos, os metadados estabelecidos no perfil de aplicação proposto foram preenchidos. Para tanto, desenvolveu-se uma aplicação *Web*, a qual permite o mapeamento de um conjunto de dados com base no perfil de aplicação proposto. Tal aplicação encontra-se ilustrada na Figura 25.

Figura 25: Aplicação Web desenvolvida para preenchimento do Perfil de Aplicação.

Fonte: o Autor.

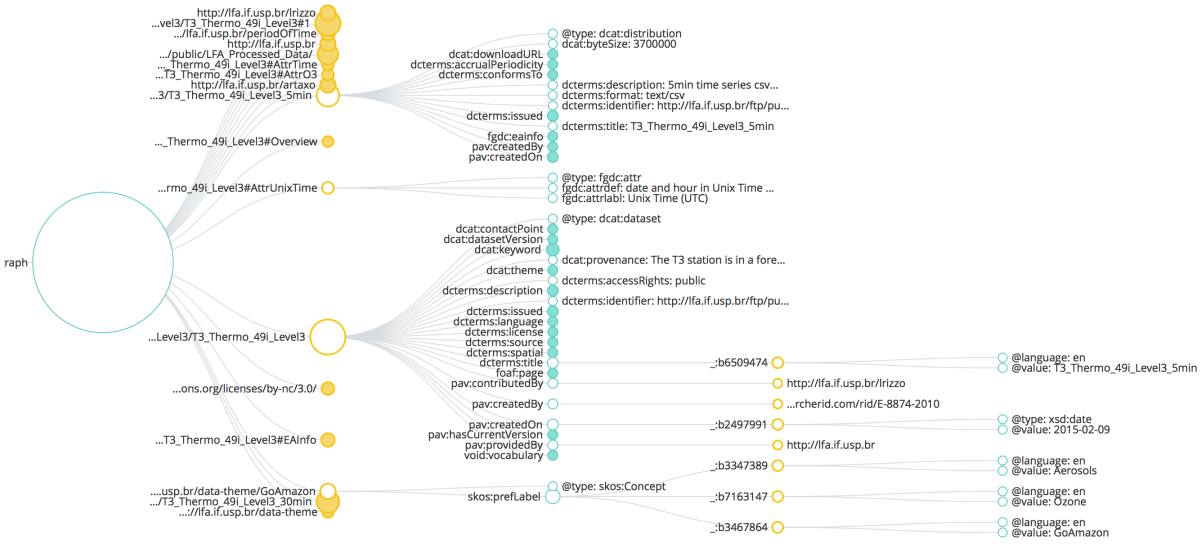
Uma vez que os metadados sobre o conjunto de dados são preenchidos na ferramenta Web, um arquivo em formato RDF/XML é disponibilizado, permitindo que as métricas de apoio à Veracidade sejam calculadas.

O arquivo RDF/XML relacionado ao conjunto de dados em questão se encontra disponível no Apêndice A.

Os documentos apresentados pelos pesquisadores permitiram descrever o conjunto de dados com todos os elementos recomendados pelo perfil de aplicação, de modo que as métricas de apoio à veracidade calculadas automaticamente obtiveram o maior valor possível (1, em uma escala de 0 a 1 para cada métrica de apoio).

Além da capacidade de serem interpretados por máquina, os formatos *linked data* permitem a geração de grafos interativos, que permitem aos usuários explorar as informações interativamente, selecionando os pontos de maior interesse. Para auxiliar a responder o questionário sobre a confiança do consumidor, os participantes puderam interagir com o grafo dinâmico, ilustrado na Figura 26. Adicionalmente, foi desenvolvida uma ferramenta que permite a geração de grafos dinâmicos a partir de arquivos nos formatos *linked data*, tais como os formatos RDF/XML e JSON-LD.

Figura 26: Grafo interativo representando o conteúdo dos elementos dos metadados que descrevem o conjuntos de dados.



Fonte: o Autor.

4.2 Compreendendo o conjunto de dados: Ciclo diurno de O_3

Descrever conjuntos de dados com os elementos definidos no perfil de aplicação proposto tem por objetivo proporcionar uma maior compreensão do conjunto de dados e dos processos de transformação executados sob o mesmo.

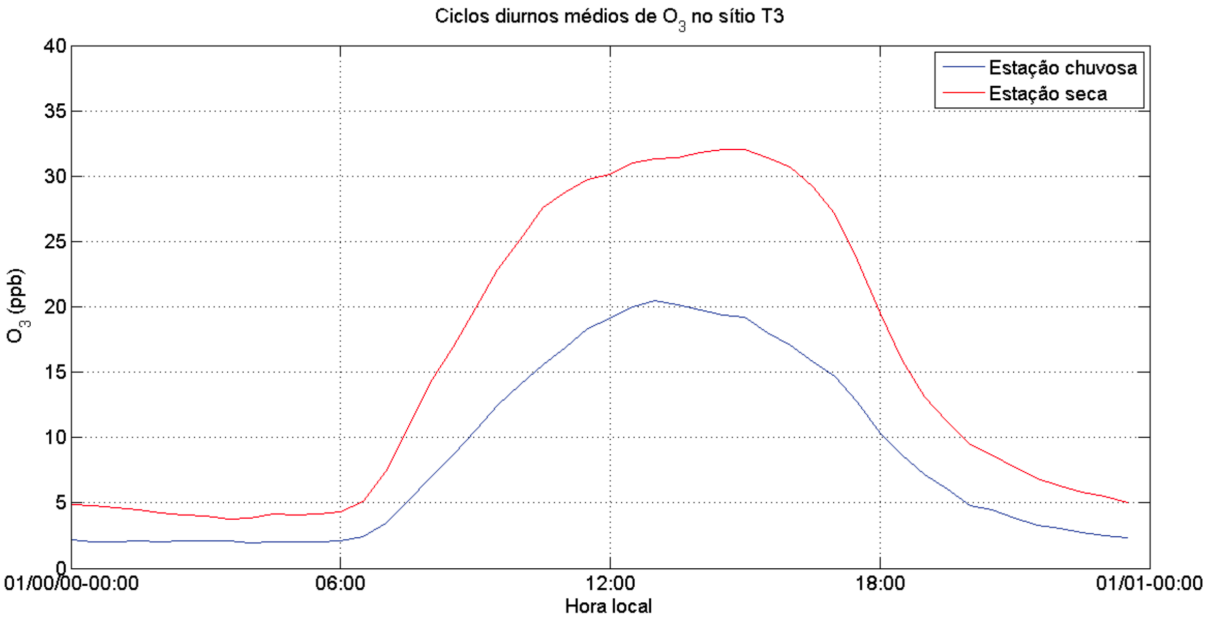
A compreensão de conjuntos de dados foi um dos focos do *III Workshop* em Ciência dos Dados, realizado em outubro de 2017, nas dependências da Escola Politécnica da Universidade de São Paulo. Ao longo do evento, foi possível compreender como pesquisadores das mais diversas áreas buscam compreender um conjunto de dados, seja por uso de ferramentas computacionais de análise exploratória dos dados, até mesmo pela geração de visualizações sobre os dados disponíveis.

O conjunto de dados de concentração de Ozônio em T3 possibilita diversas análises científicas, assim como permite a integração com outros conjuntos de dados. Para explorar a importância do registro de situações adversas como um dos elementos de metadados, buscou-se evidenciar uma característica que já era do conhecimento dos provedores do conjunto de dados, mas que poderia não ser facilmente compreendida e contextualizada para os consumidores de dados, até pelo fato que possuíam níveis de conhecimento diferenciados, além de serem originários de outras áreas do conhecimento.

Ao longo da análise do ciclo diurno da Ozônio na região, espera-se encontrar um comportamento típico, intimamente relacionado com a incidência de luz solar. Dessa forma, observa-se um comportamento característico de substâncias formadas fotoquimicamente, com valores máximos de concentração em horários próximos do meio-dia.

No trabalho de Rodrigues (2015) é apresentado o ciclo diurno médio da concentração de Ozônio no sítio T3, comparando-o entre a estação chuvosa (período de janeiro a junho) e a estação seca (período de julho a dezembro) de Manaus. A Figura 27 apresenta o ciclo de O_3 em T3.

Figura 27: Ciclos diurnos médio da concentração de O_3 nas estações chuvosa (janeiro a junho) e seca (julho a dezembro) no sítio T3.

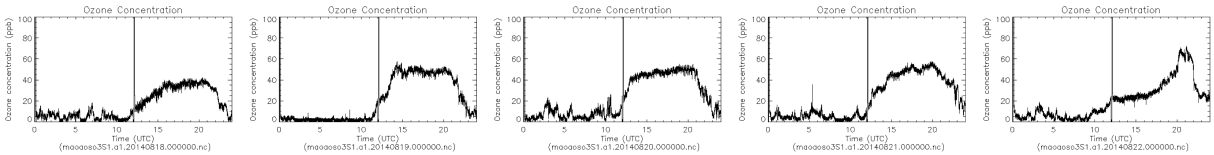


Fonte: (RODRIGUES, 2015).

A Figura 28 apresenta a medição de O_3 em T3 entre as datas 18/08/2014 e 22/08/2014. Com base na análise da série temporal da concentração de O3, é possível notar um aumento da concentração de O_3 no período entre 18 e 22/08/2014, quando comparado a outros dias de coleta, com valores superiores a média para a estação. Por exemplo, em 20 de agosto de 2014, os valores máximos médios chegaram a 50 ppb, cerca de 30 ppb acima da média da estação chuvosa em T3.

O sítio T3 apresenta em sua atmosfera grande disponibilidade de compostos orgânicos voláteis, com destaque para o isopropeno e seus produtos de oxidação, de modo que o reagente limitante para produção de O_3 nesse sítio é o NO_2, cujas principais fontes em T3 são as emissões das queimas de biomassa e as emissões urbanas de Manaus, que são carregadas pelo vento. Logo, há a hipótese de que o aumento da concentração de O_3

Figura 28: Medição da concentração de O_3 em T3 entre 18/08/2014 e 22/08/2014. O horário se encontra na escala de tempo UTC. O horário local em Manaus é dado por UTC - 4.



Fonte: ARM Data Discovery Tool.

neste período se deu devido ao aumento das queimadas, associado com uma mudança na direção dos ventos que trouxe massas de ar provenientes do arco do desflorestamento (parte sul-suldeste da Amazônia legal).

De fato, este conhecimento trazido pelos provedores do conjunto de dados, especialistas do domínio em questão, auxiliou na compreensão do conjunto de dados, uma vez que estes valores incomuns puderam ser esclarecidos aos presentes.

Corrobora a hipótese os dados disponibilizados pelo Banco de Dados de Queimadas[1] do INPE (Instituto Nacional de Pesquisas Espaciais), sobre os focos no Bioma Amazônia para períodos antes, durante e após o período de estudo. De fato, entre 18/08/2014 e 22/08/2014 houve 40.851 focos, enquanto que no período anterior e no posterior houve 25.871 e 13.815 focos, respectivamente, conforme apresentado nas Figuras 29a, 29b e 29c.

O conhecimento desse contexto possibilitou aos organizadores do *Workshop* compreenderem o fenômeno e explorarem os dados já cientes deste fato. Para o perfil de aplicação proposto, o elemento de descrição `<fgdc:complete>` permitiu descrever este fenômeno, possibilitando que os consumidores do conjuntos de dados já saibam, de antemão, um comportamento que irão encontrar nos dados ao longo do processo de análise.

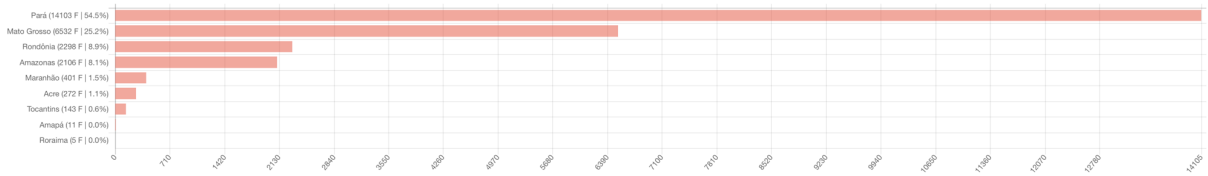
4.3 Validação do Perfil de Aplicação

Ao longo do *III Workshop em Ciência dos Dados* o perfil de aplicação foi apresentado, juntamente com os metadados que descrevem este conjunto de dados.

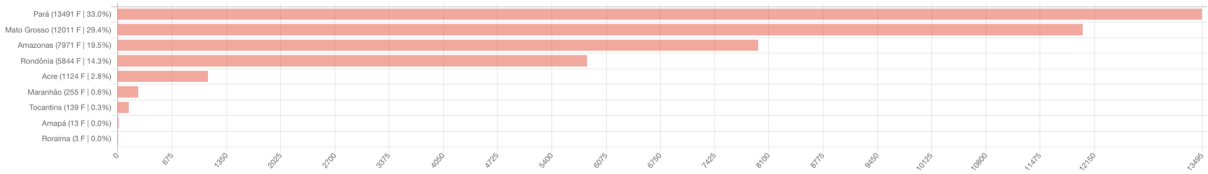
O *workshop* contou com a presença de pesquisadores do *Oak Ridge National Laboratory* (ORNL), responsáveis por ferramentas computacionais de gestão, análise e visualização de dados do programa *Atmospheric Radiation Measurement* (ARM/DoE/US), Giri Palanisamy e Bhargavi Krishna. O projeto GoAmazon tem seus dados armazenados na

[1]Disponível no endereço eletrônico <https://prodwww-queimadas.dgi.inpe.br/bdqueimadas>

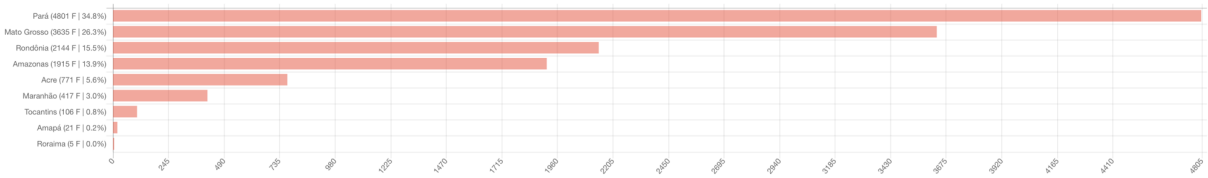
Figura 29: Número de focos em três períodos para o Bioma Amazônia.



(a) Período de 13/08/2014 a 17/08/2014: 25.871 focos.



(b) Período de 18/08/2014 a 22/08/2014: 40.851 focos.



(c) Período de 23/08/2014 a 27/08/2014: 13.815 focos.

Fonte: Banco de Dados de Queimadas do INPE.

infraestrutura computacional do ARM.

O evento contou também com a presença do então cientista chefe em gestão de dados científicos da *United States Geological Survey* (USGS/DoI/US), Mike Frame. O modelo de gestão de dados da USGS foi o escolhido nesta tese como elemento norteador dos requisitos do perfil de aplicação.

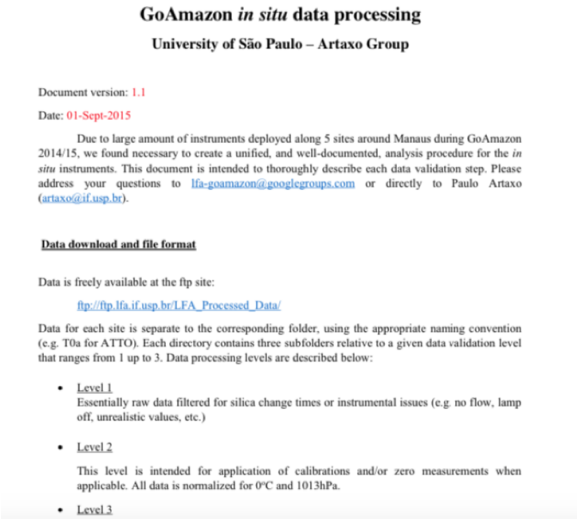
Foi possível também contar com a presença da pesquisadora principal do projeto DataONE (*National Science Foundation - NSF*), Profa. Suzie Allard (*University of Tennessee*).

4.3.1 Avaliando a Confiança no Uso de Conjuntos de Dados

De modo a validar o perfil de aplicação proposto para a descrição de conjuntos de dados científicos, adotou-se a metodologia de *survey* para a captura de opiniões de um grupo de usuários, no caso, os participantes do III WDS. Para isto, os participantes do evento foram divididos em dois grandes grupos: grupo A e grupo B.

Ao longo do III WDS, os participantes receberam um *e-mail* solicitando a participação em um *survey* no qual eles deveriam responder um conjunto de perguntas sobre um

Figura 30: Documentação auxiliar fornecida a cada grupo.



GoAmazon *in situ* data processing

University of São Paulo – Artaxo Group

Document version: 1.1

Date: 01-Sept-2015

Due to large amount of instruments deployed along 5 sites around Manaus during GoAmazon 2014/15, we found necessary to create a unified, and well-documented, analysis procedure for the *in situ* instruments. This document is intended to thoroughly describe each data validation step. Please address your questions to lfa-goamazon@googlegroups.com or directly to Paulo Artaxo (artaxo@if.usp.br).

Data download and file format

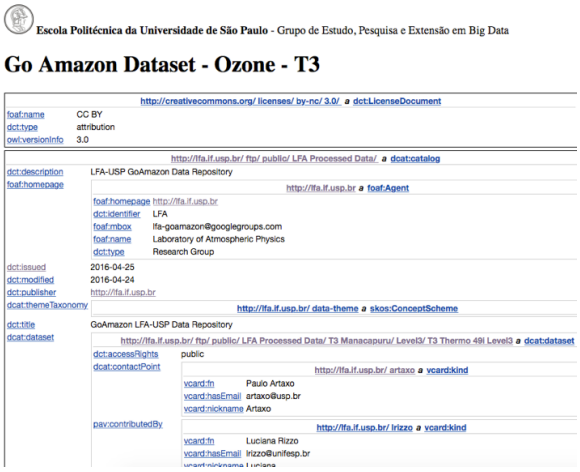
Data is freely available at the ftp site:

ftp://ftp.lfa.if.usp.br/LFA_Processed_Data/

Data for each site is separate to the corresponding folder, using the appropriate naming convention (e.g. T0a for ATTO). Each directory contains three subfolders relative to a given data validation level that ranges from 1 up to 3. Data processing levels are described below:

- Level 1
  Essentially raw data filtered for silica change times or instrumental issues (e.g. no flow, lamp off, unrealistic values, etc.)
- Level 2
  This level is intended for application of calibrations and/or zero measurements when applicable. All data is normalized for 0ºC and 1013hPa.
- Level 3

(a) Grupo A: Relatório Técnico elaborado pelos produtores do conjunto de dados.



Escola Politécnica da Universidade de São Paulo - Grupo de Estudo, Pesquisa e Extensão em Big Data

Go Amazon Dataset - Ozone - T3

http://creativecommons.org/ licenses/ by-nc/ 3.0/ a dct:LicenseDocument
foaf:name CC BY
dct:type attribution
owl:versionInfo 3.0

http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ a dcat:catalog
dct:description LFA-USP GoAmazon Data Repository
foaf:homepage http://lfa.if.usp.br a foaf:Agent
foaf:homepage http://lfa.if.usp.br
dct:identifier LFA
foaf:mbox lfa-goamazon@googlegroups.com
foaf:name Laboratory of Atmospheric Physics
dct:type Research Group
dct:issued 2016-04-25
dct:modified 2016-04-24
dct:publisher http://lfa.if.usp.br
dcat:themeTaxonomy http://lfa.if.usp.br/ data-theme a skos:ConceptScheme
dct:title GoAmazon LFA-USP Data Repository
dcat:dataset http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3 a dcat:dataset
dct:accessRights public
dcat:contactPoint http://lfa.if.usp.br/ artaxo a vcard:kind
vcard:fn Paulo Artaxo
vcard:hasEmail artaxo@usp.br
vcard:nickname Artaxo
pav:contributedBy http://lfa.if.usp.br/ lrizzo a vcard:kind
vcard:fn Luciana Rizzo
vcard:hasEmail lrizzo@unifesp.br

(b) Grupo B: Descrição do conjunto de dados com base no perfil de aplicação proposto.

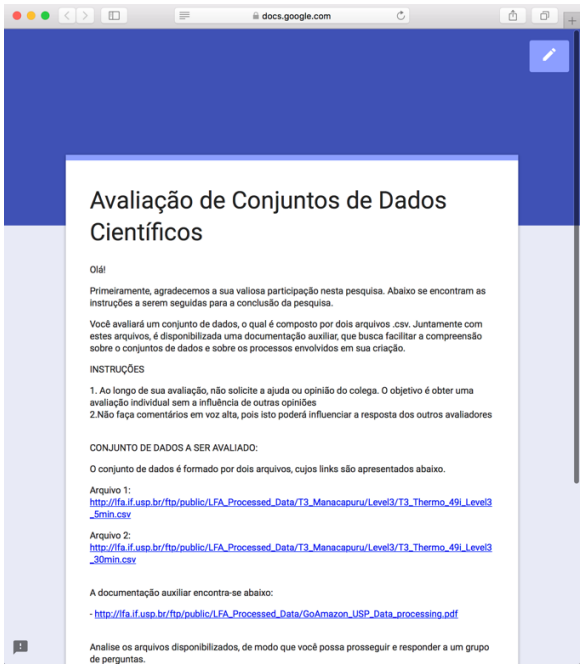
conjunto de dados. Para isto, foi fornecido o conjunto de dados de concentração de Ozônio e uma documentação auxiliar. Ao grupo A, a documentação auxiliar foi o relatório técnico elaborado pelos produtores do conjunto de dados. Já o grupo B recebeu o perfil de aplicação proposto para a Veracidade como elemento de documentação auxiliar. A Figura 30 apresenta a documentação auxiliar fornecida a cada grupo. Ambos documentos são apresentados no Apêndice C.

De posse do conjunto de dados e da correspondente documentação auxiliar, os participantes responderam questionários eletrônicos, apresentados na Figura 31.

Para ambos os grupos, as seguintes perguntas foram feitas sobre o conjunto de dados.

1. Descreva o que está sendo representado pelo conjunto de dados;
2. Em que período os dados foram coletados?
3. Qual(is) instrumento(s) foi(ram) utilizados para coletar os dados?
4. Quem publicou o conjunto de dados?
5. Você compreendeu como o provedor de dados gerou o conteúdo dos dois arquivos .csv disponibilizados? (sim/não)
6. O que significa cada atributo (coluna) dos arquivos de dados disponibilizados?

Figura 31: Questionários respondidos sobre o conjunto de dados de concentração de Ozônio.



Avaliação de Conjuntos de Dados Científicos

Olá!

Primeiramente, agradecemos a sua valiosa participação nesta pesquisa. Abaixo se encontram as instruções a serem seguidas para a conclusão da pesquisa.

Você avaliará um conjunto de dados, o qual é composto por dois arquivos .csv. Juntamente com estes arquivos, é disponibilizada uma documentação auxiliar, que busca facilitar a compreensão sobre o conjuntos de dados e sobre os processos envolvidos em sua criação.

INSTRUÇÕES

1. Ao longo de sua avaliação, não solicite a ajuda ou opinião do colega. O objetivo é obter uma avaliação individual sem a influência de outras opiniões
2.Não faça comentários em voz alta, pois isto poderá influenciar a resposta dos outros avaliadores

CONJUNTO DE DADOS A SER AVALIADO:

O conjunto de dados é formado por dois arquivos, cujos links são apresentados abaixo.

Arquivo 1:
http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3_5min.csv

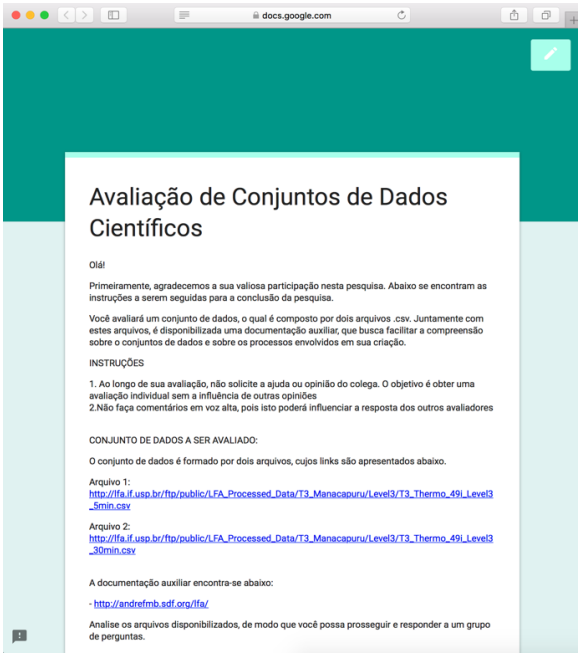
Arquivo 2:
http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3_30min.csv

A documentação auxiliar encontra-se abaixo:

- http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/GoAmazon_USP_Data_processing.pdf

Analise os arquivos disponibilizados, de modo que você possa prosseguir e responder a um grupo de perguntas.

(a) Questionário fornecido ao grupo A.



Avaliação de Conjuntos de Dados Científicos

Olá!

Primeiramente, agradecemos a sua valiosa participação nesta pesquisa. Abaixo se encontram as instruções a serem seguidas para a conclusão da pesquisa.

Você avaliará um conjunto de dados, o qual é composto por dois arquivos .csv. Juntamente com estes arquivos, é disponibilizada uma documentação auxiliar, que busca facilitar a compreensão sobre o conjuntos de dados e sobre os processos envolvidos em sua criação.

INSTRUÇÕES

1. Ao longo de sua avaliação, não solicite a ajuda ou opinião do colega. O objetivo é obter uma avaliação individual sem a influência de outras opiniões
2.Não faça comentários em voz alta, pois isto poderá influenciar a resposta dos outros avaliadores

CONJUNTO DE DADOS A SER AVALIADO:

O conjunto de dados é formado por dois arquivos, cujos links são apresentados abaixo.

Arquivo 1:
http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3_5min.csv

Arquivo 2:
http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3_30min.csv

A documentação auxiliar encontra-se abaixo:

- http://andrefmb.sdf.org/lfa/

Analise os arquivos disponibilizados, de modo que você possa prosseguir e responder a um grupo de perguntas.

(b) Questionário fornecido ao grupo B.

7. Diante de sua análise, se posicione e justifique sobre a existência ou não de comportamentos adversos no conjunto de dados.

Além das perguntas acima, ao término do *survey*, os participantes respondiam a seguinte pergunta: "*Suponha que você necessita utilizar os dados dos arquivos apresentados em um experimento científico. Com base em sua análise, o quão confiante você está em utilizar esses dados em seu experimento?*". As respostas foram coletadas na escala *Likert*, com as seguintes opções:

- Não estou confiante em utilizar o conjunto de dados;
- Sinto-me pouco confiante em utilizar o conjunto de dados;
- Não consigo formar uma opinião sobre a minha confiança em utilizar o conjunto de dados;
- Sinto-me razoavelmente confiante em utilizar o conjunto de dados;
- Sinto-me confiante em utilizar o conjunto de dados.

4.3.1.1 Análise dos Resultados

O estudo de caso foi executado ao longo do III WDS, para o qual os participantes tiveram um período de 24 horas para submeter sua avaliação sobre o conjunto de dados. Ao término desse período, estatísticas foram coletadas e analisadas, as quais são apresentadas nas próximas seções, ponderadas pelas observações efetuadas pelos respondentes em suas respostas dissertativas/qualitativas.

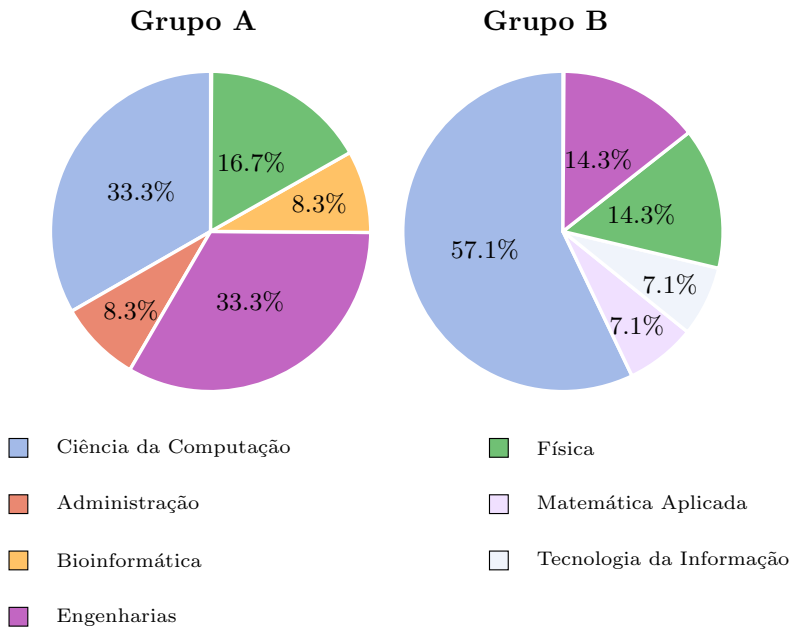
4.3.1.1.1 Número de Respostas Obtidas

A parte prática do III WDS contou com a participação de 49 pessoas, não sendo contabilizados os membros da organização do evento, assim como as pessoas que já conheciam previamente a estratégia do estudo de caso. Por meio de um sorteio, 25 pessoas foram alocadas no grupo A e, consequentemente, 24 pessoas no grupo B.

Ao término do período de submissão das avaliações, 12 avaliações foram enviadas pelo grupo A (48% dos convidados) e 14 pelo grupo B (58% dos convidados).

4.3.1.1.2 Área de Formação dos Respondentes

Figura 32: Área de Formação dos Respondentes.



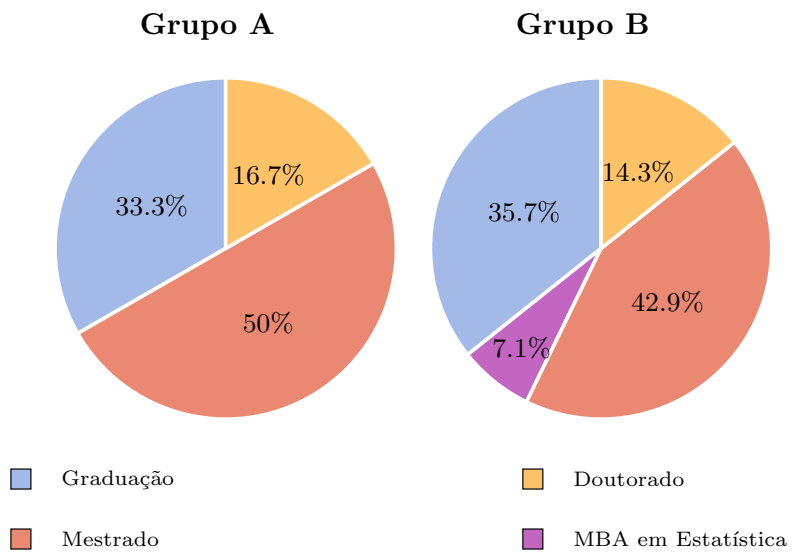
Com relação á área de formação dos respondentes, ambos os grupos apresentaram membros de diferentes áreas de conhecimento. Uma vez que o III WDS foi um evento

hospedado nas dependências da Escola Politécnica da Universidade de São Paulo, esperar-se-ia uma maior participação de pessoas ligadas à Engenharia e à Ciência da Computação. Esta expectativa se confirmou nos resultados obtidos, uma vez que, em ambos os grupos, a maioria dos respondentes é proveniente dessas áreas de conhecimento.

Devido a natureza interdisciplinar dos temas abordados no III WDS, os grupos contaram com pessoas oriundas das áreas de Física, Matemática Aplicada e Bioinformática, conforme apresentado na Figura 32.

4.3.1.1.3 Grau de Instrução dos Respondentes

Figura 33: Formação Acadêmica dos Respondentes.

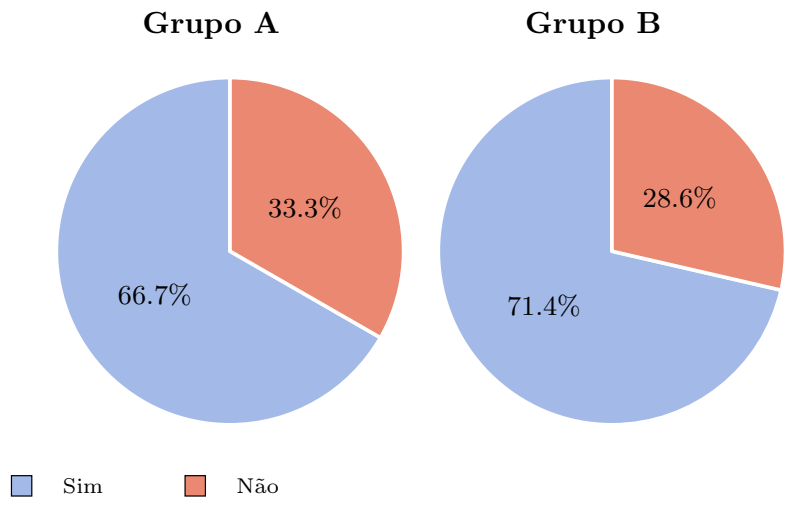


Com relação ao grau de instrução dos respondentes, a Figura 33 demonstra que, em ambos os grupos, aproximadamente 50% dos respondentes possuíam o grau de mestre em suas áreas de especialização. A proporção de respondentes por grau de instrução foi muito similar em ambos os grupos.

4.3.1.1.4 Entendimento sobre o Processo de Geração dos Dados

Com relação à pergunta: "*Você compreendeu como o provedor de dados gerou o conteúdo dos dois arquivos .csv disponibilizados?*", o grupo B, o qual recebeu o conjunto de dados descrito pelos metadados proposto nesta tese, apresentou um maior número de respostas positivas.

Figura 34: Entendimento sobre o processo de Geração de Dados



Com base na análise das respostas dissertativas de ambos os grupos, mais especificamente respostas à pergunta: "*Descreva o que está sendo representado pelo conjunto de dados*", é possível notar uma maior dificuldade de compreender o conjunto de dados pelos membros do grupo A, ressaltada na ausência de detalhes em suas respostas, conforme pode-se observar em alguns exemplos, apresentados a seguir:

- *"Coleta de arquivos, unidade consumida e tempo."*;
- *"Dados de Ozonio."*;
- *"Dados coletados no goAmazon sobre Ozonio."*;
- *"São os dados capturados de Aerosol do projeto GoAmazon."*.

No grupo A, apenas um respondente chegou a detalhar o que estava sendo representado pelo conjunto de dados, respondendo: "*São medidas de Ozônio (O3) na atmosfera realizadas no site Site: T3, Instrumento: Thermo, Modelo:49i . Arquivo 1 e Arquivo 2 diferem apenas nos intervalos de escrita de dados, 5 e 30 minutos, respectivamente. O valor gravado é a média de medidas desde a última gravação (5 ou 30 minutos). Os dados são de nível 3, ou seja, já foram validados contra outros instrumentos.*"

O grupo B, por sua vez, conseguiu detalhar um pouco mais suas respostas, como se observa:

- *"A concentração média de Ozônio (O3) em ppb, em determinado posto de medição (estação T3), já no nível 3 de processamento, em intervalos de 5 e 30 minutos."*;

- *"A concentração de ozônio, em ppb (partes por bilhão), na estação T3 do projeto GoAmazon"*;
- *" Representa o nível de concentração de ozônio (O3) medidos em períodos de 5 min e de 30 min."*;

4.3.1.1.5 Confiança em reutilizar o conjunto de dados

As perguntas sobre o conjunto de dados tinham como objetivo primário estimular o respondente a interagir com os arquivos de dados disponibilizados, juntamente com a documentação auxiliar fornecida. Esperava-se que, ao longo da atividade, o respondente formaria uma opinião sobre o quão confiante se sente em utilizar o conjunto de dados em um experimento próprio, expressada por meio de uma escala *likert*.

A hipótese do estudo de caso era que, com o uso do perfil de aplicação proposto, os respondentes do grupo B estariam mais confiantes em utilizar o conjunto de dados, pelo fato de que poderiam compreender melhor o conjunto de dados e os processos de transformações aos quais este foi submetido.

Como forma de validar tal hipótese, além da pergunta final, a pergunta "*Diante de sua análise, se posicione e justifique sobre a existência ou não de comportamentos adversos no conjunto de dados.*" teve por objetivo verificar se os respondentes encontraram eventuais valores adversos ou premissas dos produtores de dados. Ao longo da preparação do estudo de caso, em colaboração com os produtores dos dados, foi feito um destaque nos metadados sobre o período correspondente a 18/08/2014 e 22/08/2014, o qual apresenta concentrações elevadas de ozônio em T3, conforme apresentado na Seção 4.2.

Com relação a este evento, apenas um respondente do grupo B foi capaz de observar o fato, com base na descrição efetuada nos metadados. A resposta do mesmo foi: "*Não encontrei, mas vi que foi feita uma ressalva sobre altas concentrações em um período específico, em função de ventos e queimadas".*

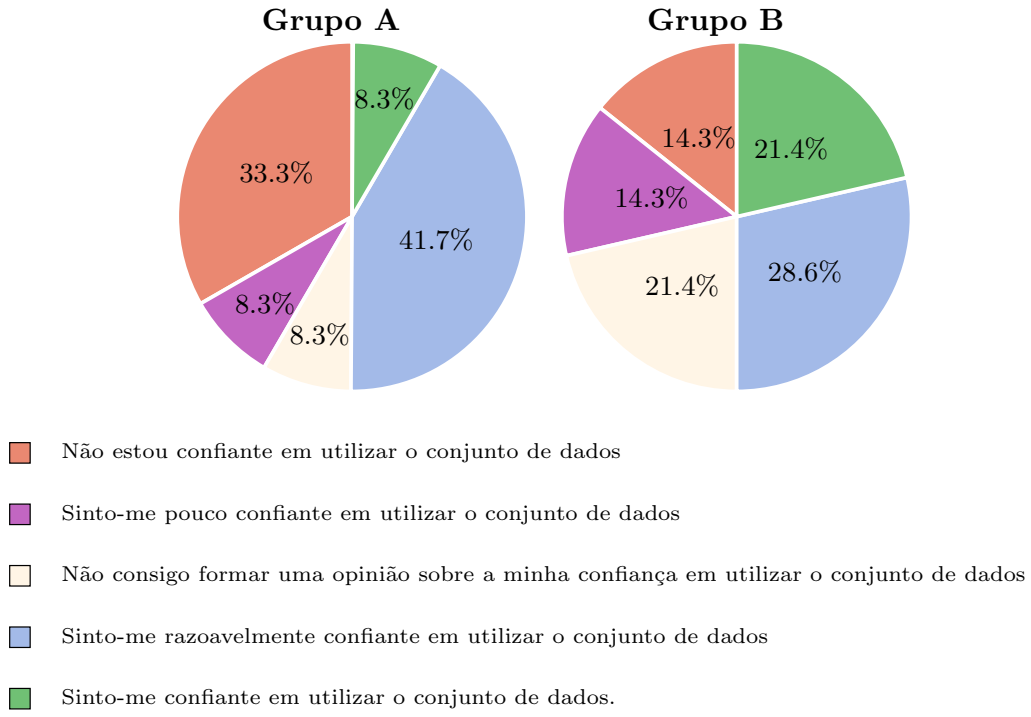
Ambos os grupos fizeram análises estatísticas e considerações sobre valores negativos e ausentes, estranhando o fato de que a concentração de ozônio ser apresentada como um valor negativo, ou até mesmo valores ausentes, conforme respostas:

- *"Existem valores NaN nas medições, o que não é explicado no documento anexo."* (grupo A);
- *"Há também pontos de informação nulos (NaN), mas provavelmente são oriundos dos processamentos de nível 1 e 2."* (grupo B);

- *"There are a lot of data missing, basically in the same period, on both files. In the 5min there are 61.056 (58,08%) and in the 30min 10.176 (58,08%). Probably these NULL values has some meaning and I need more information about this. Lv3 say that data must agree within 20%. Up to the 3rd quartile, mean is aproximatelly 18, but there are some high values like 94.10 (5min) and 86.32(30min), so there are unexpected behaviour."* (grupo A);

- *"Existem vários dados coletados que não seguem o que era esperado para a medição, fugindo bastante (ou dando erro mesmo, NaN) do que foi medido antes e depois dos mesmos. Por exemplo: existem valores que estão negativos e de repente se tornam positivos em apenas um período de medição. Isso denota anomalia, pois não houve uma mudança gradual, para baixo ou para cima, que indicaria de fato que estaria havendo uma mudança ambiental real. Uma mudança de um período apenas, drástica, nesses casos, podem ser consideradas adversas."* (grupo B).

Figura 35: Confiança em reutilizar o conjunto de dados



Com relação à pergunta final, qual seja: "*Suponha que você necessita utilizar os dados dos arquivos apresentados em um experimento científico. Com base em sua análise, o quão confiante você está em utilizar esses dados em seu experimento?*", as respostas obtidas para a mesma demonstram que o uso de um perfil de aplicação para a descrição

de conjuntos de dados aumenta a capacidade de compreensão, levando a um aumento da confiança em reutilizar conjuntos de dados científicos.

Mais especificamente, enquanto que no grupo A apenas 8.3% dos respondentes se sentem confiantes em utilizar o conjunto de dados em um experimento próprio, o grupo B apresentou um percentual de 21.4% de respondentes confiantes no reuso.

Destaca-se também a redução do percentual de respondentes que estão não confiantes em utilizar o conjuntos de dados. De fato, enquanto 33.3% dos respondentes em A estão não confiantes, apenas 14.3% dos respondentes em B apresentam a mesma opinião.

Um ponto importante a ser notado nos resultados é o aumento do percentual de respostas neutras ou indecisas sobre a confiança no reuso para o grupo B (21.4%), em relação ao grupo A (8.3%). Com base nas respostas dissertativas, e nas discussões efetuadas no III WDS após o período de coleta das respostas, observou-se que este aumento foi devido ao formato de apresentação dos metadados de descrição. Optou-se por apresentar os metadados propostos em um formato isento de elementos de interatividade e usabilidade, o que levou a uma maior dificuldade em compreende-los, quando comparado a um relatório de pesquisa, que serviu como referência ao grupo A.

4.4 Considerações sobre o estudo de caso

O estudo de caso proposto permitiu avaliar a percepção de um grupo de usuários sobre os elementos de descrição que acompanham um conjunto de dados, possibilitando avaliar o nível de confiança em reutilizar o conjunto de dados em um novo experimento.

Os respondentes que fizeram parte do grupo B, o qual recebeu as descrições sobre o conjunto de dados com base no perfil de aplicação proposto nesta tese, se sentiram mais confiantes, encontrando valores adversos e suas respostas foram menos superficiais, quando comparadas as respostas fornecidas pelo grupo A.

A ausência de elementos de interatividade e de usabilidade ao usuário dificultaram a compreensão dos metadados de descrição, conforme observado na análise das respostas dissertativas. Os respondentes do grupo B puderam ser mais assertivos e demonstram uma menor rejeição (representada pela maior confiança) ao conjunto de dados.

Ao longo da elaboração da descrição do conjunto de dados por meio do perfil de aplicação proposto, foi feito um trabalho de gestão de dados de modo que a perspectiva intrínseca aos dados obtivesse seu valor máximo (isto é, 1.0). Ou seja, todos os elementos

de metadados do perfil de aplicação foram preenchidos, os recursos do conjunto de dados estavam disponíveis, os responsáveis pelo conjunto de dados estavam identificados, o formato do conjunto de dados era um formato aberto, etc. Com relação à perspectiva social da veracidade, tendo em vista que a avaliação não foi feita por especialistas na área do conjunto de dados, optou-se por apenas validar a hipótese de que o perfil de aplicação proposto para a veracidade melhora a compreensão de um conjunto de dados, refletindo no aumento da confiança dos respondentes.

5 CONSIDERAÇÕES

Como objetivo geral, esta pesquisa buscou caracterizar um modelo de avaliação de conjuntos de dados científicos. Para isto, estabeleceu-se o conceito da dimensão de Veracidade dos dados, sob a qual um perfil de aplicação foi definido, permitindo a elaboração de um índice de veracidade para conjuntos de dados.

O conceito proposto para veracidade dos dados contempla três elementos: legitimidade, transparência e adequabilidade. A legitimidade tem por objetivo evidenciar as decisões ao longo das atividades que produziram o conjunto de dados. A transparência, por sua vez, proporciona acesso ao conjunto de dados, identifica partes e demais processos envolvidos na criação, publicação e reuso do conjunto de dados. O terceiro elemento, adequabilidade, é obtido por meio da revisão e extensão do conjunto de dados pela comunidade científica.

Para o desafio da reprodutibilidade e reuso dos dados científicos, ocasionado principalmente pela ausência de informações que apoiam a compreensão e validação de conjunto de dados e os processos de transformações aos quais esse foi submetido, foi proposta a incorporação da dimensão de Veracidade dos dados à um experimento científico, permitindo assim a elaboração de um modelo de avaliação de conjuntos de dados científicos.

O modelo proposto é baseado no perfil de aplicação desenvolvido ao longo da pesquisa, o qual estabeleceu um conjunto de elementos descritores do conjunto dos dados. A partir do perfil de aplicação, um conjunto de métricas de apoio à veracidade foi proposto. Tais métricas mensuram, dentre outros aspectos, o grau de aderência ao perfil de aplicação, a disponibilidade dos recursos e a manutenibilidade do conjunto de dados.

A interoperabilidade do formato de descrição é uma característica fundamental a ser considerada quando do desenvolvimento de modelos computacionais, de modo que tanto máquinas quanto seres humanos possam fazer uso da estrutura adotada. Dessa forma, fez-se uso de formatos *linked data*, os quais tornam possível o processamento automático das métricas por rotinas computacionais. A comunidade científica, com uso da descrição do

conjunto de dados e dos recursos associados, provê um *feedback* sobre o conjunto de dados, o qual, de uma maneira colaborativa, irá permitir determinar o índice de veracidade de um conjunto de dados para uma determinada área de pesquisa ou de aplicação.

Em um ambiente colaborativo, é preciso modelar estratégias de proteção contra comportamentos ingênuos ou maliciosos, os quais buscam mascarar a avaliação de um conjunto de dados. Para isto, foi incorporado ao modelo um algoritmo de determinação de reputação de avaliadores, evitando situações em que um conjunto de dados obtenha um alto índice de veracidade, quando apenas uma única avaliação foi feita, ou que o índice de veracidade seja afetado por uma avaliação maliciosa, a qual busca apenas penalizar um conjunto de dados avaliado de maneira arbitrária e contrária ao consenso que está sendo formado na rede de avaliações.

Ao modelo também foi incorporado um mecanismo de decaimento temporal das avaliações realizadas para um conjunto de dados, de modo a contemplar situações em o conjunto de dados possui dados sensíveis ao tempo.

De acordo com Wazlawick (2015), os objetivos específicos são subprodutos do objetivo geral, não triviais e verificáveis ao final do trabalho. Desse modo, nos itens a seguir, reproduz-se os objetivos específicos definidos nesta pesquisa e, evidenciam-se, resumidamente, os respectivos resultados obtidos, verificáveis ao longo do trabalho:

- Descrever o estado da arte sobre Ciência dos dados e seus desafios: embora um trabalho de pesquisa tenha como elemento obrigatório uma revisão da literatura, este objetivo específico foi explicitado em virtude da falta de referências que buscam organizar a área de Ciência dos Dados, na busca por contribuir com a comunidade científica com uma organização sobre o tópico e seus desafios. O resultado obtido se encontra apresentado no Capítulo 2;

- Definir o conceito de Veracidade dos Dados: a dimensão de Veracidade dos dados é pouco explorada na literatura científica, embora diversos trabalhos afirmem a sua importância para um processo de análise, reuso e reprodutibilidade de conjuntos de dados científicos. Ao longo do desenvolvimento da pesquisa buscou-se estabelecer o conceito de Veracidade dos dados, o qual foi refinado nos eventos elaborados pelo grupo de estudos, pesquisa e extensão em *Big Data* da Escola Politécnica da Universidade de São Paulo. O conceito proposto se encontra apresentado no Capítulo 3;

- Caracterizar um perfil de aplicação de metadados que suporte a descrição de conjun-

tos de dados científicos, de modo a permitir a criação de métricas que irão compor o índice de veracidade proposto: o perfil de aplicação proposto se encontra apresentado no Capítulo 3. O perfil de dados contemplou o uso de padrões de metadados consolidados, e buscou atender a requisitos estabelecidos por instituições renomadas internacionalmente, com destaque para a *United States Geological Survey* (USGS). Para tanto, contou-se com a valiosa colaboração do então cientista-chefe em gestão de dados científicos da USGS, *Prof. Mike Frame*. O perfil de aplicação desenvolvido foi utilizado em um conjunto de dados proveniente do projeto GoAmazon, apresentado no Capítulo 4;

- Definir um modelo de avaliação de conjuntos de dados, gerando assim um índice de veracidade de conjuntos de dados científicos disponibilizados em portais de dados: o modelo de avaliação se encontra apresentado no Capítulo 3.

Nas seções seguintes são evidenciadas as contribuições obtidas com esta pesquisa, bem como como as considerações que encerram a tese com a visualização de trabalhos futuros percebidos ao longo do desenvolvimento da pesquisa.

5.1 Contribuições da Tese

Para a área de Engenharia da Computação, a contribuição principal foi o desenvolvimento de um modelo computacional que incorpora a dimensão de veracidade dos dados em conjuntos de dados científicos, que são publicados em portais de dados. Tal modelo fornece mecanismos para a determinação da Veracidade em conjuntos de dados científicos, contemplando a determinação da reputação dos consumidores, juntamente com uma estratégia de decaimento temporal das avaliações para conjuntos de dados sensíveis ao tempo.

A partir do modelo proposto, descrições em formatos *linked data* permitirão que soluções computacionais sejam desenvolvidas de modo a calcular as métricas de apoio à Veracidade propostas, que serão elementos norteadores para a avaliação da comunidade científica.

Há um esforço considerável para a publicação de conjuntos de dados científicos. Ao longo do desenvolvimento desta pesquisa, buscou-se a proposição de soluções computacionais que não incorporem em demasia uma complexidade ao processo de publicação e compartilhamento de dados. Para isto, foram utilizados padrões de metadados consolidados e estabelecidos, além de formatos de dados compreensíveis por máquina, permitindo

que os portais de dados possam incorporar o perfil de aplicação proposto, juntamente com o índice de veracidade, quando da publicação de conjuntos de dados.

Esta pesquisa contribuiu com um modelo computacional para validação de requisitos que suportam a veracidade dos dados, consolidados em um perfil de aplicação. A opção por adotar um perfil de aplicação como elemento norteador dos requisitos de descrição de conjuntos de dados foi realizada em virtude de que múltiplos perfis de aplicação poderão ser desenvolvidos, adaptando-se às necessidades específicas de cada área de pesquisa ou de aplicação.

Os elementos de descrição escolhidos foram baseados nas recomendações de órgãos de referência internacional, como a *Data Science Association* e a *United States Geological Survey*, possibilitando que diversas iniciativas que se baseiam em tais recomendações apresentem uma compatibilidade de dados, permitindo assim uma rápida absorção do modelo proposto.

As métricas de apoio à Veracidade foram definidas de uma maneira a se adaptar a diferentes perfis de aplicação, o que possibilitará a evolução e, consequentemente, customização do modelo proposto para diferentes realidades.

O modelo de avaliação proposto buscou fazer uso de algoritmos de determinação da reputação dos avaliadores, pela proposta do algoritmo RBPD* como uma extensão ao algoritmo original proposto por Liu et al. (2017), na busca por aumentar a robustez do índice de veracidade dos dados.

O perfil de aplicação desenvolvido buscou contemplar os requisitos da USGS, de modo que se faz necessário sua extensão quando aplicado em outras áreas do conhecimento, as quais podem possuir requisitos diversos aos da USGS. De todo modo, a metodologia apresentada ao longo do Capítulo 3 possibilita que o perfil de aplicação seja readequado para outras áreas, cabendo também o aperfeiçoamento da perspectiva social da veracidade, refletindo as características sociais da comunidade científica em questão.

5.2 Trabalhos Futuros

Durante a realização deste pesquisa foram identificados alguns desafios e oportunidades que podem ser objetos de trabalhos futuros, quais sejam:

- Ampliar os mecanismos de decaimento temporal do índice de veracidade dos dados com base na dimensão de qualidade dos dados de *timeliness*, atendendo a cená-

rios em que os conjuntos de dados sejam sensíveis em relação à temporalidade da avaliação;

- Modificar o algoritmo RBPD* para que contemple métricas de Análise de Redes Sociais, permitindo que a reputação dos avaliadores seja determinada em função da dinâmica do comportamento de um avaliador juntamente com métricas que analisam a posição do avaliador na rede social de avaliações;

- Desenvolver novos mecanismos e estratégias para a coleta dos elementos de descrição necessários ao conceito de Veracidade, facilitando a sua operacionalização nas diversas etapas do ciclo de vida dos dados;

- Estender o modelo de avaliação de conjuntos de dados para gerar índices que irão incorporar, juntamente com a dimensão da Veracidade, a dimensão de Valor de *Big Data*;

- Incorporar técnicas de aprendizagem de máquina para a análise dos elementos descritores definidos no perfil de aplicação;

- Integrar o modelo de avaliação de conjunto de dados a recursos de captura de proveniência durante a execução de experimentos científicos como, por exemplo, integrando o modelo aqui desenvolvido com a arquitetura computacional de captura de proveniência definida por Silva (2017).

- Ampliar a perspectiva social do índice de veracidade, estendendo o estudo de caso desenvolvido a um grupo de especialistas, de modo que estes validem uma coletânea de conjuntos de dados na área de especialização, possibilitando estender os objetivos de validação social estabelecidos nesta tese.

REFERÊNCIAS

ALEIXO, D. V. B. S.; CATARINO, M. E. Contribuições de um perfil de aplicação à descrição arquivística visando a interoprabilidade nos moldes do modelo resource description framework. In: *XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação*. [S.l.: s.n.], 2015. p. 100. Citado na página 48.

ALEXANDER, K.; CYGANIAK, R.; HAUSENBLAS, M.; ZHAO, J. Describing linked datasets: On the design and usage of void, the "vocabulary of interlinked datasets". In: . [S.l.: s.n.], 2009. v. 538, p. 10. Citado 2 vezes nas páginas 70 e 71.

ALLEN, I. E.; SEAMAN, C. A. Likert scales and data analyses. *Quality progress*, American Society for Quality, v. 40, n. 7, p. 64, 2007. Citado 2 vezes nas páginas 83 e 92.

ARASS, M. E.; TIKITO, I.; SOUISSI, N. Data lifecycles analysis: towards intelligent cycle. In: *The IEEE Second International Conference on Intelligente Systems and Computer Vision - ISCV2017*. [S.l.: s.n.], 2017. Citado na página 44.

BALL, A.; DUKE, M. How to track the impact of research data with metrics. *DCC How-to Guides*, Digital Curation Centre, v. 1, n. 1, p. 2–16, 2015. Citado 2 vezes nas páginas 82 e 95.

BALLARD, C.; COMPERT, C.; JESIONOWSKI, T.; MILMAN, I.; PLANTS, B.; ROSEN, B.; SMITH, H. et al. *Information Governance Principles and Practices for a Big Data Landscape*. 1st edition. ed. [S.l.]: IBM Redbooks, 2014. Citado 2 vezes nas páginas 56 e 57.

BARRY, S.; ELITH, J. Error and uncertainty in habitat models. *Journal of Applied Ecology*, Blackwell Publishing Ltd, v. 43, n. 3, p. 413–423, 2006. ISSN 1365-2664. Disponível em: <http://dx.doi.org/10.1111/j.1365-2664.2006.01136.x>. Citado na página 45.

BASILI, V. R.; ROMBACH, H. D. The tame project: Towards improvement-oriented software environments. *IEEE Transactions on software engineering*, IEEE, v. 14, n. 6, p. 758–773, 1988. Citado na página 92.

BATISTA, A. F. M.; CORRêA, P. L. P.; PALANISAMY, G. Visual Analytics improving Data Understandability in IoT Projects : in IoT Projects: An Overview of the U.S. DOE ARM Program Data Science Tools. *Proceedings of the 2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems Visual*, p. 349–354, 2016. Citado 4 vezes nas páginas 17, 32, 59 e 62.

BERTI-ÉQUILLE, L.; BA, M. L. Veracity of big data: Challenges of cross-modal truth discovery. *J. Data and Information Quality*, v. 7, p. 12:1–12:3, 2016. Citado na página 21.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, p. 205–227, 2009. Citado 2 vezes nas páginas 51 e 80.

BOARD, D. U. *DCMI metadata terms*. [S.l.]: Dublin Core Metadata Initiative, 2008. Citado na página 48.

BOONE, H. N.; BOONE, D. A. Analyzing likert data. *Journal of extension*, Extension Journal, Inc, v. 50, n. 2, p. 1–5, 2012. Citado na página 96.

BORGMAN, C. L. Research data: Who will share what, with whom, when, and why? 2010. Citado na página 60.

BORGMAN, C. L. *Big Data, Little Data, No Data: Scholarship in the Networked World*. [S.l.]: The MIT Press, 2015. ISBN 0262028565, 9780262028561. Citado na página 61.

BORGMAN, C. L. If data sharing is the answer, what is the question? *ERCIM NEWS*, p. 15, 2015. Citado na página 19.

BRADNER, S. *IETF RFC 2119: Key words for use in RFCs to Indicate Requirement Levels*. [S.l.], 1997. Citado na página 71.

BRODIE, M. L. Understanding data Science: An emerging discipline for Data-Intensive discovery. *CEUR Workshop Proceedings*, v. 1536, p. 238–245, 2015. ISSN 16130073. Citado na página 32.

CAO, L. Data science: Challenges and directions. *Communications of the ACM*, ACM, v. 60, n. 8, p. 59–68, 2017. Citado 2 vezes nas páginas 35 e 36.

CHAMBERS, J. M. Greater or lesser statistics: a choice for future research. *Statistics and Computing*, v. 3, n. 4, p. 182–184, 1993. ISSN 1573-1375. Disponível em: <http://dx.doi.org/10.1007/BF00141776>. Citado na página 29.

CHAPMAN, A. Environmental Data Quality: Data Cleaning Tools. n. 6, p. 1–57, 2004. Disponível em: <http://splink.cria.org.br/docs/appendix_i.pdf>. Citado na página 17.

CHEN, M.; TREFETHEN, A.; BAÑARES-ALCÁNTARA, R.; JIROTKA, M.; COECKE, B.; ERTL, T.; SCHMIDT, A. From data analysis and visualization to causality discovery. *Computer*, IEEE, v. 44, n. 10, p. 84–87, 2011. Citado na página 58.

CICCARESE, P.; SOILAND-REYES, S.; BELHAJJAME, K.; GRAY, A. J.; GOBLE, C.; CLARK, T. Pav ontology: provenance, authoring and versioning. *Journal of biomedical semantics*, BioMed Central, v. 4, n. 1, p. 37, 2013. Citado na página 71.

CLEVELAND, W. S. Data science: an action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, Blackwell Publishing Ltd, v. 69, n. 1, p. 21–26, 2001. ISSN 1751-5823. Disponível em: <http://dx.doi.org/10.1111/j.1751-5823.2001.tb00477.x>. Citado 3 vezes nas páginas 30, 31 e 32.

COLLINS. *Collins English Dictionary*. 2017. Citado na página 55.

COMMITTEE, F. G. D. et al. Content standard for digital geospatial metadata. *http://gis. sam. usace. army. mil/General_Information/Standards_And_Reports/Metadata% 20Content% 20Standard. pdf*, 1998. Citado 2 vezes nas páginas 66 e 71.

COYLE, K.; BAKER, T.; INITIATIVE, D. C. M. et al. *Guidelines for Dublin Core application profiles*. [S.l.]: Dublin Core Metadata Initiative, 2009. Citado 2 vezes nas páginas 48 e 67.

CRONE, R. *Big Data Veracity Assessment: Improving risk assessment by adding high veracity data to existing contents insurance models*. Tese (Ph.D Thesis) — Delft University of Technology, dec 2016. Citado 2 vezes nas páginas 55 e 56.

CROSAS, M.; KING, G.; HONAKER, J.; SWEENEY, L. Automating Open Science for Big Data. *ANNALS of the American Academy of Political and Social Science*, n. May, p. 1–20, 2014. ISSN 0002-7162. Disponível em: <http://gking.harvard.edu/files/gking/files/annals-bigdatapaper-v8_0.pdf>. Citado 2 vezes nas páginas 18 e 27.

CUEVAS-VICENTTÍN, V.; KIANMAJD, P.; LUDäSCHER, B.; MISSIER, P.; CHIRIGATI, F.; WEI, Y.; KOOP, D.; DEY, S. Provenance Storage, Querying, and Visualization in PBase. *Lecture Notes in Computer Science*, p. 239–241, 2015. Disponível em: <http://dx.doi.org/10.1007/978-3-319-16462-5_24>. Citado na página 17.

CYGANIAK, R.; WOOD, D.; LANTHALER, M. *RDF 1.1 concepts and abstract syntax. Recommendation, W3C (2014)*. 2014. Citado na página 71.

DASGUPTA, A.; POCO, J.; WEI, Y.; COOK, R.; BERTINI, E.; SILVA, C. T. Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison. *IEEE Transactions on Visualization and Computer Graphics*, v. 21, n. 9, p. 996–1014, 2015. ISSN 1077-2626. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7061479>. Citado na página 17.

DAVENPORT, T. H.; PATIL, D. J. Data Scientist: The Sexiest Job Of the 21st Century. *Harvard Business Review*, v. 90, n. 10, p. 70–76, 2012. Disponível em: <http://ezproxy.auckland.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=79996214&site=ehost-live&scope=site>. Citado na página 31.

DCAT-AP. *DCAT application profile for data portals in Europe*. 2015. Disponível em: <https://joinup.ec.europa.eu/asset/dcat_application_profile/description>. Citado 3 vezes nas páginas 69, 71 e 72.

DEKHTYAR, A. Lecture Notes on Data Science - DATA 301. California Polytechnic State University. 2016. Disponível em: <http://users.csc.calpoly.edu/~dekhtyar/DATA301-Spring2016/lectures/lec03.301.pdf>. Citado 4 vezes nas páginas 32, 35, 36 e 37.

DHAR, V. Data science and prediction. *Commun. ACM*, ACM, New York, NY, USA, v. 56, n. 12, p. 64–73, dez. 2013. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2500499>. Citado na página 34.

DHILLON, S. K.; SIDHU, A. S. *Data Intensive Computing for Biodiversity*. Springer, 2013. v. 485. 1-118 p. (Studies in Computational Intelligence, v. 485). ISBN 978-3-642-38046-4. Disponível em: <http://dx.doi.org/10.1007/978-3-642-38047-1>. Citado 2 vezes nas páginas 25 e 26.

DIEPENBROEK, M.; GLÖCKNER, F. O.; GROBE, P.; GÜNTSCH, A.; HUBER, R.; KÖNIG-RIES, B.; KOSTADINOV, I.; NIESCHULZE, J.; SEEGER, B.; TOLKSDORF, R.; TRIEBEL, D. Towards an integrated biodiversity and ecological research data management and archiving platform: The german federation for the curation of biological data (gfbio). In: PLÖDEREDER, E.; GRUNSKE, L.; SCHNEIDER, E.; ULL, D. (Ed.). *44. Jahrestagung der Gesellschaft für Informatik, Informatik 2014, Big Data - Komplexität meistern, 22.-26. September 2014 in Stuttgart, Deutschland*. GI, 2014. (LNI, v. 232), p. 1711–1721. ISBN 978-3-88579-626-8. Disponível em: <http://subs.emis.de/LNI/Proceedings/Proceedings232.html>. Citado na página 26.

DING, Y.; LI, X. Time weight collaborative filtering. In: ACM. *Proceedings of the 14th ACM international conference on Information and knowledge management*. [S.l.], 2005. p. 485–492. Citado na página 91.

DONOHO, D. 50 years of Data Science. 2015. Disponível em: <http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf>. Citado 4 vezes nas páginas 28, 29, 31 e 32.

DUMONTIER, M.; GRAY, A. J.; MARSHALL, M. S.; ALEXIEV, V.; ANSELL, P.; BADER, G.; BARAN, J.; BOLLEMAN, J. T.; CALLAHAN, A.; CRUZ-TOLEDO, J. et al. The health care and life sciences community profile for dataset descriptions. *PeerJ*, PeerJ Inc., v. 4, p. e2331, 2016. Citado 3 vezes nas páginas 68, 72 e 75.

DUNCAN, A. *Focus on the "Three Vs" of Big Data Analytics: Variability, Veracity and Value. Gartner*. 2016. Disponível em: <https://www.gartner.com/doc/2921417/focus-vs-big-data-analytics>. Citado na página 56.

DYER, J.; HAVER, M.; NOWLAN, M. *ASPIRE - A Study on the Prospects of the Internet for Research and Education 2014-2020*. [S.l.]: TERENA, 2012. Citado 2 vezes nas páginas 22 e 54.

EPSRC. *Clarifications of EPSCR expectation on research data management*. EPSRC, 2014. Disponível em: <https://www.epsrc.ac.uk/about/standards/researchdata/expectations/>. Citado na página 82.

ESTEVES, D.; RULA, A.; REDDY, A. J.; LEHMANN, J. Toward veracity assessment in rdf knowledge bases: An exploratory analysis. *J. Data and Information Quality*, ACM, New York, NY, USA, v. 9, n. 3, p. 16:1–16:26, fev. 2018. ISSN 1936-1955. Disponível em: <http://doi.acm.org/10.1145/3177873>. Citado na página 57.

FAUNDEEN, J. L.; BURLEY, T. E.; CARLINO, J. A.; GOVONI, D. L.; HENKEL, H. S.; HOLL, S. L.; HUTCHISON, V. B.; MARTÍN, E.; MONTGOMERY, E. T.; LADINO, C. et al. *The United States geological survey science data lifecycle model*. [S.l.], 2014. Citado 4 vezes nas páginas 41, 42, 44 e 66.

FERRELL, W.; KOCH, D.; KUPERBERG, M. *GOAmazon2014 Workshop Report*. 2014. <http://science.energy.gov/~/media/ber/pdf/GOAmazon2014_200dpi.pdf?id=81>. Acessado em: 24 out. 2015. Citado 2 vezes nas páginas 15 e 27.

FGDC. *Geospatial Line of Business. OMB Circular A-16 Supplemental Guidance*. [S.l.], 2008. Disponível em: https://www.fgdc.gov/policyandplanning/A-16-supplemental-guidance-endorsed-dec08.pdf. Citado na página 67.

FOUNDATION, N. S. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*. [S.l.], 2005. Citado na página 31.

FOWLER, D.; AMANN, M.; ANDERSON, F.; ASHMORE, M.; COX, P.; DEPLEDGE, M.; DERWENT, D.; GRENNFELT, P.; HEWITT, N.; HOV, O. et al. Ground-level ozone in the 21st century: future trends, impacts and policy implications. *Royal Society Science Policy Report*, v. 15, n. 08, 2008. Citado na página 114.

FOWLER, M. *UML Essencial: um breve guia para linguagem padrão*. [S.l.]: Bookman Editora, 2014. Citado na página 49.

FRANK, M.; WALKER, J. User centred methods for measuring the value of open data. *The Journal of Community Informatics*, v. 12, n. 2, 2016. Citado na página 97.

GAO, J.; DONG, Y.-W.; SHANG, M.-S.; CAI, S.-M.; ZHOU, T. Group-based ranking method for online rating systems with spamming attacks. *EPL (europhysics letters)*, IOP Publishing, v. 110, n. 2, p. 28003, 2015. Citado na página 83.

HARDISTY, A.; ROBERTS, D.; ADDINK, W.; AELTERMAN, B.; AGOSTI, D.; AMARAL-ZETTLER, L.; nO, A. H. A.; ARVANITIDIS, C.; BACKELJAU, T.; BAILLY, N.; BELBIN, L.; BERENDSOHN, W.; BERTRAND, N.; CAITHNESS, N.; CAMPBELL, D.; COCHRANE, G.; CONRUYT, N.; CULHAM, A.; DAMGAARD, C.; DAVIES, N.; FADY, B.; FAULWETTER, S.; FEEST, A.; FIELD, D.; GARNIER, E.; GESER, G.; GILBERT, J.; GROSCHE; GROSSER, D.; HERBINET, B.; HOBERN, D.; JONES, A.; JONG, Y. de; KING, D.; KNAPP, S.; KOIVULA, H.; LOS, W.; MEYER, C.; MORRIS, R. a.; MORRISON, N.; MORSE, D.; OBST, M.; PAFILIS, E.; PAGE, L. M.; PAGE, R.; PAPE, T.; PARR, C.; PATON, A.; PATTERSON, D.; PAYMAL, E.; PENEV, L.; POLLET, M.; PYLE, R.; RAAB-STRAUBE, E. von; ROBERT, V.; ROBERTSON, T.; ROVELLOTTI, O.; SAARENMAA, H.; SCHALK, P.; SCHAMINEE, J.; SCHOFIELD, P.; SIER, A.; SIERRA, S.; SMITH, V.; SPRONSEN, E. van; THORNTON-WOOD, S.; TIENDEREN, P. van; TOL, J. van; TUAMA, E. O.; UETZ, P.; VAAS, L.; Vignes Lebbe, R.; VISION, T.; VU, D.; De Wever, A.; WHITE, R.; WILLIS, K.; YOUNG, F. A decadal view of biodiversity informatics: challenges and priorities. *BMC ecology*, v. 13, n. 1, p. 16, 2013. ISSN 1472-6785. Disponível em: <http://www.biomedcentral.com/1472-6785/13/16>. Citado 4 vezes nas páginas 17, 25, 26 e 27.

HEERY, R.; PATEL, M. Application profiles: mixing and matching metadata schemas. *Ariadne*, v. 1, n. 25, 2000. Citado na página 48.

HERITAGE, A. *American Heritage® Dictionary of the English Language - Fifth Edition*. 2016. Citado na página 55.

HEY, T.; TANSLEY, S.; TOLLE, K. M. et al. *The fourth paradigm: data-intensive scientific discovery*. [S.l.]: Microsoft research Redmond, WA, 2009. v. 1. Citado na página 60.

HOLANDA, B. A. *Absorção da radiação por aerossóis na Amazônia*. Tese (Doutorado) — Universidade de São Paulo, 2015. Citado na página 112.

HUBBARD, D. W. *How to measure anything : finding the value of "intangibles"in business*. 2nd ed.. ed. Hoboken, N.J. :: Wiley,, 2010. Citado na página 105.

JAIN, M. K. Uncertain Unstructured Data Management : Put All Apple in Nine Basket. *INternational Journal on Big Data Security Intelligence*, v. 2, n. 1, p. 1–6, 2015. Disponível em: <http://dx.doi.org/10.14257/ijbdsi.2015.2.1.01>. Citado na página 44.

KOWALD, D.; SEITLINGER, P.; TRATTNER, C.; LEY, T. Long time no see: The probability of reusing tags as a function of frequency and recency. In: ACM. *Proceedings of the 23rd International Conference on World Wide Web*. [S.l.], 2014. p. 463–468. Citado na página 91.

KUČERA, J.; CHLAPEK, D.; NEČASKỲ, M. Open government data catalogs: Current approaches and quality perspective. In: SPRINGER. *International Conference on Electronic Government and the Information Systems Perspective*. [S.l.], 2013. p. 152–166. Citado na página 75.

LARRAIN, S.; TRATTNER, C.; PARRA, D.; GRAELLS-GARRIDO, E.; NØRVÅG, K. Good times bad times: A study on recency effects in collaborative filtering for social tagging. In: ACM. *Proceedings of the 9th ACM Conference on Recommender Systems*. [S.l.], 2015. p. 269–272. Citado 4 vezes nas páginas 91, 92, 100 e 107.

LAWSON, D. *Big Data Veracity and the Value of Doubt*. 2014. Disponível em: <https://goo.gl/rorx2F>. Citado 2 vezes nas páginas 54 e 55.

LEBO, T.; SAHOO, S.; MCGUINNESS, D.; BELHAJJAME, K.; CHENEY, J.; CORSAR, D.; GARIJO, D.; SOILAND-REYES, S.; ZEDNIK, S.; ZHAO, J. Prov-o: The prov ontology. *W3C recommendation*, v. 30, 2013. Citado na página 69.

LI, D.; CAO, P.; GUO, Y.; LEI, M. Time weight update model based on the memory principle in collaborative filtering. *JCP*, v. 8, n. 11, p. 2763–2767, 2013. Citado na página 91.

LIAO, H.; ZENG, A.; XIAO, R.; REN, Z.-M.; CHEN, D.-B.; ZHANG, Y.-C. Ranking reputation and quality in online rating systems. *PloS one*, Public Library of Science, v. 9, n. 5, p. e97146, 2014. Citado na página 83.

LIU, J.; LI, J.; LI, W.; WU, J. Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 115, p. 134–142, 2016. Citado na página 21.

LIU, X. L.; LIU, J. G.; YANG, K.; GUO, Q.; HAN, J. T. Identifying online user reputation of user-object bipartite networks. *Physica A: Statistical Mechanics and its Applications*, Elsevier B.V., v. 467, p. 508–516, 2017. ISSN 03784371. Disponível em: <http://dx.doi.org/10.1016/j.physa.2016.10.031>. Citado 7 vezes nas páginas , 83, 84, 85, 87, 88 e 132.

LOUKIDES, M. What is data science? *O'Reilly radar*, 6 2010. Disponível em: <http://radar.oreilly.com/2010/06/what-is-data-science.html>. Citado na página 33.

MACEACHREN, a. M. Visual Analytics and Uncertainty : It ' s Not About the Data. 2015. Citado na página 44.

MARR, B. *Big data in practice*. [S.l.: s.n.], 2016. Citado na página 56.

MARTIN, S.; ARTAXO, P.; MACHADO, L.; MANZI, A.; SOUZA, R.; SCHUMACHER, C.; WANG, J.; BISCARO, T.; BRITO, J.; CALHEIROS, A. et al. The green ocean amazon experiment (goamazon2014/5) observes pollution affecting gases, aerosols, clouds, and rainfall over the rain forest. *Bulletin of the American Meteorological Society*, v. 98, n. 5, p. 981–997, 2017. Citado na página 112.

MATTMANN, C. a. Computing: A vision for data science. *Nature*, v. 493, n. 7433, p. 473–475, 2013. ISSN 1476-4687. Disponível em: <http://www.nature.com/nature/journal/v493/n7433/full/493473a.html>. Citado na página 18.

MCCLURE, S. *Data Science and Big Data: Two very Different Beasts*. 2015. <http://www.kdnuggets.com/2015/07/data-science-big-data-different-beasts.html>. Acessado em: 24 out. 2015. Citado na página 27.

MELOROSE, J.; PERROY, R.; CAREAS, S. *Doing Data Science*. [S.l.: s.n.], 2015. v. 1. ISSN 1098-6596. ISBN 9788578110796. Citado na página 32.

MERRIAM-WEBSTER. *Merriam-Webster's Dictionary*. 2017. Disponível em: <http://www.dictionary.com/browse/veracity>. Citado na página 55.

MICHENER, W. K. Ecological data sharing. *Ecological Informatics*, Elsevier B.V., v. 29, p. 33–44, 2015. ISSN 15749541. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1574954115001004>. Citado 3 vezes nas páginas 16, 26 e 45.

MOHAMMADI, N. G.; PAULUS, S.; BISHR, M.; METZGER, A.; KÖNNECKE, H.; HARTENSTEIN, S.; WEYER, T.; POHL, K. Trustworthiness attributes and metrics for engineering trusted internet-based software systems. In: HELFERT, M.; DESPREZ, F.; FERGUSON, D.; LEYMANN, F. (Ed.). *Cloud Computing and Services Science*. Cham: Springer International Publishing, 2014. p. 19–35. ISBN 978-3-319-11561-0. Citado 3 vezes nas páginas 57, 58 e 75.

MORISETTE, J. T.; JARNEVICH, C. S.; HOLCOMBE, T. R.; TALBERT, C. B.; IGNIZIO, D.; TALBERT, M. K.; SILVA, C.; KOOP, D.; SWANSON, A.; YOUNG, N. E. Vistrails sahm: visualization and workflow management for species habitat modeling. *Ecography*, Blackwell Publishing Ltd, v. 36, n. 2, p. 129–135, 2013. ISSN 1600-0587. Disponível em: <http://dx.doi.org/10.1111/j.1600-0587.2012.07815.x>. Citado na página 17.

MUNAFÒ, M. R.; NOSEK, B. A.; BISHOP, D. V.; BUTTON, K. S.; CHAMBERS, C. D.; SERT, N. P. du; SIMONSOHN, U.; WAGENMAKERS, E.-J.; WARE, J. J.; IOANNIDIS, J. P. A manifesto for reproducible science. *Nature Human Behaviour*, Nature Publishing Group, v. 1, p. 0021, 2017. Citado 3 vezes nas páginas 22, 58 e 75.

NAUR, P. *Concise survey of computer methods*. [S.l.]: Petrocelli Books, 1974. Citado na página 28.

NILSSON, M.; BAKER, T.; JOHNSTON, P. *Singapore Framework for Dublin Core Application Profiles. Retrieved, April 30, 2011*. 2008. Citado 4 vezes nas páginas 48, 49, 50 e 65.

NISO. *Understanding metadata*. NISO, 2004. Disponível em: <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>. Citado na página 46.

OLIVEIRA, A. C. S. de. *Desvendando a autoralidade colaborativa na e-science sob a ótica dos direitos de propriedade intelectual*. Tese (Tese de Doutorado) — Universidade Federal da Paraíba, nov 2016. Citado 11 vezes nas páginas 39, 41, 42, 43, 60, 61, 67, 68, 70, 79 e 83.

OWAIS, S. S.; HUSSEIN, N. S. Extract five categories cpivw from the 9v's characteristics of the big data. *International Journal of Advanced Computer Science and Applications*, SCIENCE & INFORMATION SAI ORGANIZATION LTD 19 BOLLING RD, BRADFORD, WEST YORKSHIRE, 00000, ENGLAND, v. 7, n. 3, p. 254–258, 2016. Citado na página 56.

PETERSON, A. T.; KNAPP, S.; GURALNICK, R.; SOBERóN, J.; HOLDER, M. T. The big questions for biodiversity informatics The big questions for biodiversity informatics. v. 8, n. 2, p. 159–168, 2010. Citado 3 vezes nas páginas 16, 26 e 27.

PETERSON, a. T.; SOBERóN, J. Species distribution modeling and ecological niche modeling: Getting the Concepts Right. *Natureza a Conservacao*, v. 10, n. 2, p. 102–107, 2012. ISSN 16790073. Citado na página 44.

PETERSON, a. T.; SOBERóN, J.; KRISHTALKA, L. A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC ecology*, BioMed Central, v. 15, n. 1, p. 15, 2015. ISSN 1472-6785. Disponível em: <http://www.biomedcentral.com/1472-6785/15/15>. Citado na página 16.

PÔÇAS, I.; GONÇALVES, J.; MARCOS, B.; ALONSO, J.; CASTRO, P.; HONRADO, J. P. Evaluating the fitness for use of spatial data sets to promote quality in ecological assessment and monitoring. *International Journal of Geographical Information Science*, v. 1, n. June, p. 1–16, 2014. ISSN 1365-8816. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/13658816.2014.924627>. Citado na página 56.

PORTER, M. E.; MILLAR, V. E. *How information gives you competitive advantage*. [S.l.]: Harvard Business Review, Reprint Service Watertown, Massachusetts, USA, 1985. Citado na página 62.

PRESS, G. *A Very Short History Of Data Science*. 2013. <http://goo.gl/isvN0h>. Acessado em: 03 set. 2016. Citado 3 vezes nas páginas 28, 29 e 31.

PRESSER, K.; WEBER, D.; NORRIE, M. C. A study of data quality requirements for empirical data in the food sciences. In: *Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014*. [S.l.]: Twenty Second European Conference on Information Systems, 2014. ISBN 978-0-9915567-0-0. ECIS 2014; Conference Location: Tel Aviv, Israel; Conference Date: June 9-11, 2014; . Citado na página 21.

PROVOST, F.; FAWCETT, T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, Mary Ann Liebert, Inc., publishers, v. 1, n. 1, p. 51–59, fev. 2013. ISSN 2167-6461. Disponível em: <http://dx.doi.org/10.1089/big.2013.1508>. Citado na página 33.

PRUD, E.; SEABORNE, A. et al. Sparql query language for rdf. 2006. Citado na página 52.

QIN, H.; LI, Z. Research on the Method of Big Data Analysis. *Information Technology Journal*, v. 12, n. 10, p. 1974–1980, out. 2013. ISSN 18125638. Disponível em: <http://www.scialert.net/abstract/?doi=itj.2013.1974.1980>. Citado na página 45.

REIMER, A. P.; MADIGAN, E. A. Veracity in big data: How good is good enough. *Health Informatics Journal*, v. 1, n. 1, p. 1460458217744369, 2018. PMID: 29388495. Disponível em: <https://doi.org/10.1177/1460458217744369>. Citado na página 59.

RESNICK, P. *RFC 2822: Internet Message Format*. [S.l.], 2001. Disponível em: <http://www.rfc-archive.org/getrfc.php?rfc=2822>. Citado na página 78.

RODRIGUES, N. P. *A influência da pluma urbana de Manaus sobre a razão de mistura de O3 troposférico na Amazônia*. 2015. Trabalho de Conclusão de Curso (Bacharel em Química), UNIFESP (Universidade Federal de São Paulo), São Paulo, Brasil. Citado 3 vezes nas páginas 112, 114 e 117.

ROGET. *the American heritage Roget's Thesaurus*. 2017. Citado na página 55.

ROSS, M.; DOUGLAS, M. Visualization for data science: Adding credibility, legitimacy, and saliency. *Big Data*, v. 4, n. 2, p. 73–74, 2016. Citado na página 58.

SACHA, D.; SENARATNE, H.; KWON, B. C.; ELLIS, G.; KEIM, D. A. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, IEEE, v. 22, n. 1, p. 240–249, 2016. Citado na página 54.

SAYÃO, L. F.; SALES, L. F. Curadoria digital: um novo patamar para preservac?o de dados digitais de pesquisa. *Informação & Sociedade*, Universidade Federal da Paraíba-Programa de Pós-Graduação em Ciência da Informação, v. 22, n. 3, 2012. Citado na página 62.

SCANNAPIECO, M.; BERTI, L. Quality of web data and quality of big data: Open problems. In: *Data and Information Quality*. [S.l.]: Springer, 2016. p. 421–449. Citado na página 58.

SCHUNN, C. D.; TRAFTON, J. G. The psychology of uncertainty in scientific data analysis. In: *Handbook of the Psychology of Science*. [S.l.]: Springer Publishing Company, 2012. Citado 2 vezes nas páginas 44 e 45.

SELTZER, M. L.; ZHANG, L. The Data Deluge: Challenges and Opportunities of unlimited data in statistical signal processing. p. 3701–3704, 2009. Citado 2 vezes nas páginas 18 e 60.

SILVA, D. L. d. *Estratégia computacional para apoiar a reprodutibilidade e reuso de dados científicos baseado em metadados de proveniência*. Tese (Doutorado) — Universidade de São Paulo, 2017. Citado 2 vezes nas páginas 50 e 133.

SILVA, G. G. C. d. et al. Caracterização fisico-química de aerossóis no experimento goamazon2014/15: a interação entre emissões urbanas de manaus com emissões naturais da floresta. Instituto Nacional de Pesquisas da Amazônia, 2015. Citado na página 113.

SIVARAJAH, U.; KAMAL, M. M.; IRANI, Z.; WEERAKKODY, V. Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, Elsevier, v. 70, p. 263–286, 2017. Citado 2 vezes nas páginas 18 e 19.

SMITH, J. R.; SCHIRLING, P. Metadata standards roundup. *IEEE MultiMedia*, IEEE, v. 13, n. 2, p. 84–88, 2006. Citado na página 47.

e-infrastructures for data publishing in biodiversity science. Sofia, p. 1–417, 11/2011 2011. Disponível em: <http://www.pensoft.net/journals/zookeys/issue/150/>. Citado na página 17.

STANTIC, B. *Lecture notes - Database Systems - Griffith University.* 2015. Disponível em: <http://hobbit.ict.griffith.edu.au/~s106441/DS/Topic7/BIGData.ppt>. Citado 2 vezes nas páginas 56 e 57.

TAYI, G. K.; BALLOU, D. P. Examining data quality. *Communications of the ACM*, ACM, v. 41, n. 2, p. 54–57, 1998. Citado na página 84.

TENOPIR, C.; KING, D. W. *Communication patterns of engineers.* [S.l.]: John Wiley & Sons, 2004. Citado na página 62.

THANOS, C.; RAUBER, A. Scientific data sharing and re-use. *ERCIM NEWS*, p. 15, 2015. Citado na página 20.

THEORET, C. *The oncoming Social Data Analysis Revolution: Chapter 2 – The 4th V: Data Veracity.* 2009. Disponível em: <http://www.slideshare.net/cgtheoret/infopresse-veracity>. Citado 2 vezes nas páginas 56 e 57.

TUKEY, J. W. The future of data analysis. *Ann. Math. Statist.*, The Institute of Mathematical Statistics, v. 33, n. 1, p. 1–67, 03 1962. Disponível em: <http://dx.doi.org/10.1214/aoms/1177704711>. Citado na página 28.

TUKEY, J. W. *Exploratory Data Analysis.* [S.l.]: Addison-Wesley, 1977. Citado na página 29.

UMBRICH, J.; NEUMAIER, S.; POLLERES, A. Quality assessment and evolution of open data portals. In: IEEE. *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on.* [S.l.], 2015. p. 404–411. Citado na página 75.

USGS. *USGS Data Review Checklist.* 2014. Disponível em: <https://www2.usgs.gov/datamanagement/documents/DataReviewChecklist_2014.pdf>. Citado na página 93.

USGS. *U.S. Geological Survey - Data Management Guidance Materials / White Papers.* 2015. Disponível em: <https://www2.usgs.gov/datamanagement/documents/MetadataQuestionnaire.pdf>. Citado na página 66.

VELLUCCI, S. L. Metadata. *Annual review of information science and technology*, Information Today, v. 33, p. 187–222, 1998. Citado na página 46.

VINES, T. H.; ALBERT, A. Y.; ANDREW, R. L.; DÉBARRE, F.; BOCK, D. G.; FRANKLIN, M. T.; GILBERT, K. J.; MOORE, J.-S.; RENAUT, S.; RENNISON, D. J. The availability of research data declines rapidly with article age. *Current biology*, Elsevier, v. 24, n. 1, p. 94–97, 2014. Citado 3 vezes nas páginas 75, 77 e 78.

WALKER, M. A. The professionalisation of data science. *International Journal of Data Science*, Inderscience Publishers (IEL), v. 1, n. 1, p. 7–16, 2015. Citado na página 93.

WANG, J.-L.; CHIOU, J.-M.; MUELLER, H.-G. Review of Functional Data Analysis. p. 1–47, 2015. Disponível em: <http://arxiv.org/abs/1507.05135>. Citado na página 46.

WAZLAWICK, R. *Metodologia de pesquisa para ciência da computação*. [S.l.]: Elsevier Brasil, 2015. v. 2. Citado na página 130.

WEIBEL, S.; KUNZE, J.; LAGOZE, C.; WOLF, M. *Dublin core metadata for resource discovery*. [S.l.], 1998. Citado 4 vezes nas páginas 47, 66, 69 e 71.

WHO. World health organization. particulate matter, ozone, nitrogen dioxide and sulfur dioxide. *Air Quality Guidelines: Global Update*, p. 76–100, 2005. Citado na página 114.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J.-W.; SANTOS, L. B. da S.; BOURNE, P. E. et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, Nature Publishing Group, v. 3, p. 160018, 2016. Citado na página 70.

WILSON, A.; DOWNS, R. R.; LENHARDT, W. C.; MEYER, C.; MICHENER, W.; RAMAPRIYAN, H.; ROBINSON, E. Realizing the value of a national asset: Scientific data. v. 95, p. 477–478, 2014. ISSN 2324-9250. Disponível em: <http://dx.doi.org/10.1002/2014EO500006>. Citado 2 vezes nas páginas 15 e 39.

WONG, P. C.; SHEN, H. W.; JOHNSON, C. R.; CHEN, C.; ROSS, R. B. The top 10 challenges in extreme-scale visual analytics. *IEEE Computer Graphics and Applications*, v. 32, n. 4, p. 63–67, July 2012. ISSN 0272-1716. Citado 2 vezes nas páginas 53 e 54.

WU, D.; YUAN, Z.; YU, K.; PAN, H. Temporal social tagging based collaborative filtering recommender for digital library. *The outreach of digital libraries: A globalized resource network*, Springer, p. 199–208, 2012. Citado na página 91.

ZAVERI, A.; RULA, A.; MAURINO, A.; PIETROBON, R.; LEHMANN, J.; AUER, S. Quality assessment for linked data: A survey. *Semantic Web*, IOS Press, v. 7, n. 1, p. 63–93, 2016. Citado na página 75.

ZHU, Y.; XIONG, Y. Defining Data Science. *CoRR*, abs/1501.0, 2015. Disponível em: <http://arxiv.org/abs/1501.05039>. Citado 4 vezes nas páginas 32, 34, 35 e 38.

APÊNDICE A – CONJUNTO DE DADOS GOAMAZON EM RDF/XML

A seguir apresenta-se o conjunto de dados utilizado no Estudo de Caso descrito pelos metadados estabelecidos pelo perfil de aplicação proposto, no formato RDF/XML.

```
1  <?xml version="1.0"?>
2  <!DOCTYPE rdf:RDF [
3    <!ENTITY schema "http://schema.org" >
4    <!ENTITY dct "http://purl.org/dc/terms/" >
5    <!ENTITY adms "http://www.w3.org/ns/adms#" >
6    <!ENTITY dcat "http://www.w3.org/ns/dcat#" >
7    <!ENTITY foaf "http://xmlns.com/foaf/0.1/" >
8    <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
9    <!ENTITY vcard "http://www.w3.org/2006/vcard/ns#" >
10   <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
11   <!ENTITY xml "http://www.w3.org/XML/1998/namespace" >
12   <!ENTITY skos "http://www.w3.org/2004/02/skos/core#" >
13   <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
14   <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
15   <!ENTITY pav "http://purl.org/pav/" >
16   <!ENTITY vera "http://wds.poli.usp.br/andrefmb/ap#" >
17   <!ENTITY fgdc "http://www.opengis.net/cat/csw/csdgm#" >
18
19 ]>
20
21 <rdf:RDF
22   xmlns="http://www.semanticweb.org/owl/owlapi/turtle#"
23   xml:base="http://www.semanticweb.org/owl/owlapi/turtle"
24   xmlns:schema="http://schema.org/"
25   xmlns:adms="http://www.w3.org/ns/adms#"
26   xmlns:owl="http://www.w3.org/2002/07/owl#"
27   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
28   xmlns:skos="http://www.w3.org/2004/02/skos/core#"
29   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
30   xmlns:vcard="http://www.w3.org/2006/vcard/ns#"
31   xmlns:dct="http://purl.org/dc/terms/"
32   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
33   xmlns:xml="http://www.w3.org/XML/1998/namespace"
34   xmlns:dcat="http://www.w3.org/ns/dcat#"
35   xmlns:foaf="http://xmlns.com/foaf/0.1/"
36   xmlns:pav="http://purl.org/pav/"
37   xmlns:vera="http://wds.poli.usp.br/andrefmb/ap#"
38   xmlns:void="http://rdfs.org/ns/void#"
39   xmlns:void-ext="http://ldf.fi/void-ext#"
40   xmlns:fgdc="http://www.opengis.net/cat/csw/csdgm#">
41
42   <!--http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T0z_ZF2_TT34_K34/Level3/ -->
43   <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/">
44
45     <rdf:type rdf:resource="&dcat;catalog"/>
46     <dct:title xml:lang="en">GoAmazon LFA-USP Data Repository</dct:title>
```

```
47    <dct:description xml:lang="en">LFA-USP GoAmazon Data Repository</dct:description>
48    <dct:issued rdf:datatype="&xsd;date">2016-04-25</dct:issued>
49
50    <dct:modified rdf:datatype="&xsd;date">2016-04-24</dct:modified>
51
52    <dct:publisher rdf:resource="http://lfa.if.usp.br"/>
53
54    <dcat:dataset rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3"/>
55    <!--
56        <dcat:dataset rdf:resource="another dataset"/>
57        -->
58
59    <dcat:themeTaxonomy rdf:resource="http://lfa.if.usp.br/data-theme"/>
60
61    <foaf:homepage rdf:resource="http://lfa.if.usp.br"/>
62
63  </rdf:Description>
64
65  <!-- THEMES -->
66  <!-- http://lfa.if.usp.br/data-theme -->
67
68  <rdf:Description rdf:about="http://lfa.if.usp.br/data-theme">
69    <rdf:type rdf:resource="&skos;ConceptScheme"/>
70    <dct:title xml:lang="it">Vocabulario LFAIFUSP</dct:title>
71  </rdf:Description>
72
73  <rdf:Description rdf:about="http://lfa.if.usp.br/data-theme/GoAmazon">
74    <rdf:type rdf:resource="&skos;Concept"/>
75    <skos:prefLabel xml:lang="en">Aerosols</skos:prefLabel>
76    <skos:prefLabel xml:lang="en">GoAmazon</skos:prefLabel>
77    <skos:prefLabel xml:lang="en">Ozone</skos:prefLabel>
78  </rdf:Description>
79
80  <!-- LICENCA  -->
81
82  <!-- http://creativecommons.org/licenses/by/4.0/ -->
83  <rdf:Description rdf:about="http://creativecommons.org/licenses/by-nc/3.0/">
84    <rdf:type rdf:resource="&dct;LicenseDocument"/>
85    <dct:type>Attribution</dct:type>
86    <owl:versionInfo>3.0</owl:versionInfo>
87    <foaf:name>CC BY</foaf:name>
88  </rdf:Description>
89
90  <!-- AGENTES -->
91
92  <rdf:Description rdf:about="http://lfa.if.usp.br">
93
94    <rdf:type rdf:resource="&foaf;Agent"/>
95    <dct:identifier>LFA</dct:identifier>
96    <foaf:name xml:lang="en">Laboratory of Atmospheric Physics</foaf:name>
97    <foaf:mbox>lfa-goamazon@googlegroups.com</foaf:mbox>
98    <foaf:homepage>http://lfa.if.usp.br</foaf:homepage>
99    <dct:type>Research Group</dct:type>
100 </rdf:Description>
101
102 <!-- DATASET -->
103
104 <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3">
105
106   <rdf:type rdf:resource="&dcat;dataset"/>
107   <dct:identifier>http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/</dct:identifier>
108   <dct:title xml:lang="en">T3_Thermo_49i_Level3_5min</dct:title>
109   <dct:description xml:lang="en">Time series of ozone concentration, in ppb, at the T3 station of the GoAmazon project.</
        dct:description>
110   <foaf:page rdf:resource="http://lfa.if.usp.br/"/>
111   <dct:language rdf:resource="http://publications.europa.eu/mdr/authority/language/ENG"/>
112   <dct:source rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/GoAmazon_USP_Data_processing.pdf"/>
113   <pav:hasCurrentVersion rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
        T3_Thermo_49i_Level3#1"/>
114   <pav:createdOn rdf:datatype="&xsd;date">2015-02-09</pav:createdOn>
115   <dct:issued rdf:datatype="&xsd;date">2015-02-09</dct:issued>
116   <pav:createdBy rdf:resource="http://www.researcherid.com/rid/E-8874-2010"/>
117   <pav:providedBy rdf:resource="http://lfa.if.usp.br"/>
118   <dcat:contactPoint rdf:resource="http://lfa.if.usp.br/artaxo"/>
```

```
119   <pav:contributedBy rdf:resource="http://lfa.if.usp.br/lrizzo"/>
120
121   <dct:spatial rdf:resource="http://www.geonames.org/3663529"/>
122   <dcat:keyword xml:lang="en">Ozone</dcat:keyword>
123   <dcat:keyword xml:lang="en">Aerosol</dcat:keyword>
124   <dcat:theme rdf:resource="http://lfa.if.usp.br/data-theme/GoAmazon"/>
125   <void:vocabulary rdf:resource="http://purl.org/dc/terms/"/>
126   <dcat:provenance>The T3 station is in a forest region, where biogenic emissions predominate in the rainy season (Jan-Jun) and
        burnings in the dry season (Jul-Dec). All seasons are influenced, to a greater or lesser extent, by dry season burning
127     emissions. The concentrations and properties of pollutants change drastically as a function of the predominant sources of pollution and
        the aging processes of pollutants as they are transported into the atmosphere. This dataset was processed in three
128     steps, namely: Level 1 - Flow in both cells must be >= 0.55 LPM; Intensity in both cells must be >= 75000 Hz. The subsequent data
        record must also be removed due to the need of instrument stabilization. O3 mixing ratios should be between -50 and 1000 ppb.
        Eventual clock delay adjustments.
129     Level 2 - apply calibration factors, baseline adjustments by linear functions, remove periods of maintenance.
130     Level 3 - intercomparison with other O3 monitor if available. Data must agree within 20%.
131
132   </dcat:provenance>
133   <dct:accessRights>public</dct:accessRights>
134   <dct:license rdf:resource="http://creativecommons.org/licenses/by/3.0/"/>
135   <dcat:datasetVersion rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
        T3_Thermo_49i_Level3#1"/>
136
137
138 </rdf:Description>
139
140 <!-- DATASET VERSION -->
141
142 <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#1"
      >
143
144   <rdf:type rdf:resource="&vera;datasetVersion"/>
145   <dct:identifier>http://lfa.if.usp.br/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#1</dct:identifier>
146   <dct:title xml:lang="en">T3_Thermo_49i_Level3 - Version 01</dct:title>
147   <dct:description xml:lang="en">Version 01: Time series of Ozone Concentration at the T3 station of the GoAmazon project.</
        dct:description>
148   <pav:createdOn rdf:datatype="&xsd;date">2015-02-09</pav:createdOn>
149   <dct:issued rdf:datatype="&xsd;date">2015-02-09</dct:issued>
150   <pav:createdBy rdf:resource="http://www.researcherid.com/rid/E-8874-2010"/>
151   <pav:version>1</pav:version>
152     <dct:temporal rdf:resource="http://lfa.if.usp.br/PeriodOfTime"/>
153   <dct:isVersionOf rdf:resource="ttp://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3"/
      >
154   <pav:previousVersion rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
        T3_Thermo_49i_Level3#1"/>
155   <fgdc:complete>Data transformations steps are described in http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/
        GoAmazon_USP_Data_processing.pdf.
156
157   Between 20140818 and 20140822 there was an increase in the number of fire outbreaks in the period, associated with a change in the
        direction of the winds which brought air masses from the deforestation arc (south-south-east of Amazonia part). This has led to a
        high concentration of ozone in this period. More information on fires in the period can be found in INPE Burning Database.
158
159   </fgdc:complete>
160
161   <dcat:distribution rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
        T3_Thermo_49i_Level3_5min"/>
162   <dcat:distribution rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
        T3_Thermo_49i_Level3_30min"/>
163
164 </rdf:Description>
165
166
167 <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
      T3_Thermo_49i_Level3_5min">
168
169   <rdf:type rdf:resource="&dcat;distribution"/>
170   <dct:identifier>http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3_5min</
        dct:identifier>
171   <dct:title>T3_Thermo_49i_Level3_5min</dct:title>
172   <dct:description>
173     5min time series csv file
174   </dct:description>
175   <pav:createdOn rdf:datatype="&xsd;date">2015-02-09</pav:createdOn>
176   <dct:issued rdf:datatype="&xsd;date">2015-02-09</dct:issued>
```

```
177   <pav:createdBy rdf:resource="http://www.researcherid.com/rid/E-8874-2010"/>
178   <dct:accrualPeriodicity rdf:resource="http://purl.org/cld/freq/irregular"/>
179   <dct:format>text/csv</dct:format>
180   <dct:conformsTo rdf:resource="ttp://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3
        #1"/>
181   <dcat:downloadURL rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
        T3_Thermo_49i_Level3_5min.csv"/>
182   <dcat:byteSize>3700000</dcat:byteSize>
183   <fgdc:eainfo rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
        EAInfo"/>
184
185 </rdf:Description>
186
187
188 <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
        T3_Thermo_49i_Level3_30min">
189
190   <rdf:type rdf:resource="&dcat;distribution"/>
191   <dct:identifier>http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3_30min</
        dct:identifier>
192   <dct:title>T3_Thermo_49i_Level3_30min</dct:title>
193   <dct:description>
194    30min extracted mean from 5min file.
195   </dct:description>
196   <pav:createdOn rdf:datatype="&xsd;date">2015-02-09</pav:createdOn>
197   <dct:issued rdf:datatype="&xsd;date">2015-02-09</dct:issued>
198   <pav:createdBy rdf:resource="http://www.researcherid.com/rid/E-8874-2010"/>
199   <dct:accrualPeriodicity rdf:resource="http://purl.org/cld/freq/irregular"/>
200   <dct:format>text/csv</dct:format>
201   <dct:conformsTo rdf:resource="ttp://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3
        #1"/>
202   <dcat:downloadURL rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/
        T3_Thermo_49i_Level3_30min.csv"/>
203   <dcat:byteSize>639000</dcat:byteSize>
204   <fgdc:eainfo rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
        EAInfo"/>
205
206 </rdf:Description>
207
208 <!-- CONTACT POINTS -->
209
210 <rdf:Description rdf:about="http://lfa.if.usp.br/artaxo">
211
212   <rdf:type rdf:resource="&vcard;kind"/>
213   <vcard:fn rdf:datatype="&vcard;fn">Paulo Artaxo</vcard:fn>
214   <vcard:nickname rdf:datatype="&vcard;Nickname">Artaxo</vcard:nickname>
215   <vcard:hasEmail rdf:datatype="&vcard;Email">artaxo@usp.br</vcard:hasEmail>
216 </rdf:Description>
217
218 <rdf:Description rdf:about="http://lfa.if.usp.br/lrizzo">
219
220   <rdf:type rdf:resource="&vcard;kind"/>
221   <vcard:fn rdf:datatype="&vcard;fn">Luciana Rizzo</vcard:fn>
222   <vcard:nickname rdf:datatype="&vcard;Nickname">Luciana</vcard:nickname>
223   <vcard:hasEmail rdf:datatype="&vcard;Email">lrizzo@unifesp.br</vcard:hasEmail>
224 </rdf:Description>
225
226 <!-- PERIODS OF TIME  -->
227
228 <rdf:Description rdf:about="http://lfa.if.usp.br/periodOfTime">
229   <rdf:type rdf:resource="&dct;periodOfTime"/>
230   <schema:startDate rdf:datatype="&xsd;date">2014-01-01</schema:startDate>
231   <schema:endDate rdf:datatype="&xsd;date">2015-01-01</schema:endDate>
232 </rdf:Description>
233
234 <!--- EAINFO -->
235
236 <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
        EAInfo">
237   <rdf:type rdf:resource="&fgdc;eainfo"/>
238   <fgdc:overview rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
        Overview"/>
239   <fgdc:attr rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
        AttrTime"/>
```

```
240     <fgdc:attr rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
            AttrUnixTime"/>
241     <fgdc:attr rdf:resource="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
            AttrO3"/>
242   </rdf:Description>
243
244   <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
            Overview">
245     <rdf:type rdf:resource="&fgdc;overview"/>
246     <fgdc:eaover>O3 concentration data were collected using Thermo Fisher Scientific's 49i UV Photometric Analyzer. In T3, the device was
             installed at 5 meters in height. Its lower limit of detection is 1 ppb (parts per billion). Observations were made at 5 minute intervals
            </fgdc:eaover>
247   </rdf:Description>
248
249   <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
            AttrTime">
250     <rdf:type rdf:resource="&fgdc;attr"/>
251     <fgdc:attrlabl>Time (UTC)</fgdc:attrlabl>
252     <fgdc:attrdef>date and hour in UTC format</fgdc:attrdef>
253   </rdf:Description>
254
255   <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
            AttrUnixTime">
256     <rdf:type rdf:resource="&fgdc;attr"/>
257     <fgdc:attrlabl>Unix Time (UTC)</fgdc:attrlabl>
258     <fgdc:attrdef>date and hour in Unix Time format, since 1970-01-01</fgdc:attrdef>
259   </rdf:Description>
260
261   <rdf:Description rdf:about="http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/T3_Manacapuru/Level3/T3_Thermo_49i_Level3#
            AttrO3">
262     <rdf:type rdf:resource="&fgdc;attr"/>
263     <fgdc:attrlabl>O3</fgdc:attrlabl>
264     <fgdc:attrdef>O3  concentration in ppb</fgdc:attrdef>
265   </rdf:Description>
266
267
268
269 </rdf:RDF>
```

APÊNDICE B – QUESTIONÁRIO - CONFIANÇA DO CONSUMIDOR

Este questionário tem por objetivo verificar a confiança do consumidor de um conjunto de dados, descrito por meio de metadados baseados no perfil de aplicação proposto para suporte à dimensão de Veracidade dos Dados.

O conjunto de dados em questão é proveniente do Workshop em Ciência dos Dados, realizado pela Escola Politécnica da Universidade de São Paulo. Com base no conjunto de dados disponibilizado pelo Laboratório de Física aplicada da USP, juntamente com documentos que buscam evidenciar como o conjunto foi gerado, elaborou-se o RDF/XML em questão. Para isto, foi feito uso da ferramenta automatizada, desenvolvida ao longo desta tese de doutorado[1].

Para compreender melhor o perfil de aplicação, acesse a tabela resumo dos elementos de metadados[2].

[1]Disponível em: <http://andrefmb.sdf.org/veracity>
[2]Disponível em <http://andrefmb.sdf.org/lfa/pa.png>

Sobre você (Opcional)

Nome: _____

Instituição: _____

E-mail: _____

Confiança do Consumidor

A determinação da confiança do consumidor sobre um conjunto de dados visa obter o feedback com relação ao conjunto de dados, e consequentemente, dos metadados que descrevem tal conjunto, possibilitando a determinação da adequabilidade do conjunto de dados para uso e reuso pela comunidade científica.

A determinação é baseada em um questionário, o qual o consumidor irá julgar um conjunto de afirmações, em uma escala de 1 a 5. As afirmações abaixo foram elaboradas com base nas discussões efetuadas pelos membros do grupo de estudos, pesquisa e extensão da Escola Politécnica da Universidade de São Paulo.

Julgue as afirmações a seguir.

1. **É possível compreender os critérios e os procedimentos realizados pelo provedor de dados para aquisição do conjunto de dados.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

2. **É possível compreender como se deram as transformações nos dados ao longo do experimento.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

3. **Os elementos de metadados e suas descrições permitem compreender o contexto em que o conjunto de dados foi utilizado.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

4. **O provedor de dados e demais agentes envolvidos na criação/liberação encontram-se bem identificados.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

5. **Os dados gerados são provenientes de uma pesquisa científica apoiada por uma instituição de pesquisa reconhecida.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

6. **As informações de contato e demais fontes adicionais fornecidas pelo provedor permitem obter maiores informações sobre o conjunto de dados.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

7. **As unidades de medida, símbolos e definições usadas no conjunto de dados se encontram sob um padrão conhecido e estabelecido.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

8. **O provedor de dados apresenta comentários para explicar valor adversos, se houver.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

9. **O conjunto de dados apresenta valores consistentes com fontes externas ou experiências prévias.**
   Discordo totalmente ○—○—○—○—○ Concordo totalmente

10. **O conjunto de dados apresenta valores consistentes com fontes externas ou experiências prévias.**
    Discordo totalmente ○—○—○—○—○ Concordo totalmente

11. **É possível recriar o conjunto de dados a partir das descrições fornecidas.**
    Discordo totalmente ○—○—○—○—○ Concordo totalmente

12. **A documentação sobre a metodologia é facilmente encontrada e apresenta elementos suficientes que permitem a sua compreensão.**
    Discordo totalmente ○—○—○—○—○ Concordo totalmente

Comentários Adicionais

13. **Caso tenha algum comentário a fazer, utilize o espaço abaixo.**

    _____

    _____

    _____

APÊNDICE C – DOCUMENTOS RELACIONADOS AO ESTUDO DE CASO - GOAMAZON

Este apêndice apresenta os documentos relacionados ao estudo de caso desenvolvido e apresentado no Capítulo 4.

C.1 Relatório Técnico

A seguir tem-se o relatório técnico apresentado ao grupo A.

GoAmazon *in situ* data processing

University of São Paulo – Artaxo Group

Document version: 1.1

Date: 01-Sept-2015

Due to large amount of instruments deployed along 5 sites around Manaus during GoAmazon 2014/15, we found necessary to create a unified, and well-documented, analysis procedure for the *in situ* instruments. This document is intended to thoroughly describe each data validation step. Please address your questions to lfa-goamazon@googlegroups.com or directly to Paulo Artaxo (artaxo@if.usp.br).

Data download and file format

Data is freely available at the ftp site:

ftp://ftp.lfa.if.usp.br/LFA_Processed_Data/

Data for each site is separate to the corresponding folder, using the appropriate naming convention (e.g. T0a for ATTO). Each directory contains three subfolders relative to a given data validation level that ranges from 1 up to 3. Data processing levels are described below:

- Level 1
  Essentially raw data filtered for silica change times or instrumental issues (e.g. no flow, lamp off, unrealistic values, etc.)

- Level 2

  This level is intended for application of calibrations and/or zero measurements when applicable. All data is normalized for 0ºC and 1013hPa.

- Level 3

  Data is validated against other instruments, e.g., BC cannot be higher than PM2.5 and so forth.

Data has been processed into unified time steps starting 01-Jan-2014. Two files are created for each instrument/site/level, one 5-min data (01-Jan-2014 00:00, 00:05, 00:10 and so on) and one 30-min data (01-Jan-2014 00:00, 00:30, 01:00 and so on).

File format is an ASCII file, using semi-colon as separator. The header of each data file contains information on site, instrument, model, serial number and amount of data that was removed applying the corresponding filter.

Time is in UTC (Local time + 4). Two date formats were used, one text in the format dd-mmm-yyyy HH:MM (e.g. 14-Jul-2014 18:45), and epoch (or Unix) time, i.e., the number of seconds since 01-Jan-1970 00:00. The time stamp stands for the beginning of that given period.

The following list describes data processing for each instrument.

AETHELOMETER (Magee) – Applies to models AE30 and AE33

Level 1: Air flow has to be between 1lpm and 8lpm; status must be 0; 1-min BC data cannot be smaller than -1 μg m^{-3} or larger than 100 μg m^{-3}.

Level 2: Apply filter loading and multiple scattering corrections according to (Rizzo et al., 2011; Weingartner et al., 2003). Apply STP corrections.

Level 3: Aethalometer BC at 880nm has to be within 30% of MAAP.

MAAP (Thermo) – Model 5012

Level 1: Iteration number has to be smaller than 50 and ≠ 0; 1-min BC data cannot be smaller than -1 μg m^{-3} or larger than 100 μg m^{-3}.

Level 2: Apply 5% correction according to Müller et al., (2011). Apply STP corrections.

Level 3: MAAP BC must be smaller than 50% of $PM_{2.5}$ measured from TEOM (Thermo). If no TEOM data is available, use SMPS integrated mass and average density of 1.4 g cm^{-3}.

Nephelometer (Ecotech) – Model Aurora 300

Level 1: 5-sec aerosol light scattering at 450nm must be within -10Mm-1 and 900 Mm-1 . Relative humidity must be smaller than 60%. Silica dryer maintenance periods were removed based on field logbook, and based on the rule: if the absolute value of scat450 variation is greater than 2Mm-1/sec, data within +/-4min is excluded.

Level 2: Correct for truncation error according to Müller et al., (2011b). Apply STP corrections.

Level 3: SSA compared with MAAP BC (MAC of 6.6 m^2 g^{-1}) must be higher than 0.1

Obs: a PM1 inlet was installed at ATTO Nephelometer on 2014 May 4th.

Nephelometer (TSI) – Model 3563

Level 1: 1-min aerosol light scattering at 550nm cannot be smaller than -10 mM^{-1} or larger than 1000 Mm^{-1}. Backscatter must be larger than 2% of forward scattering and smaller than 80% of forward scattering. Relative humidity must be smaller than 70%.

Level 2: Correct for truncation error according to Anderson and Ogren, (1998). Apply STP corrections. Apply STP corrections and zero correction. SAE must be larger than 0 and smaller than 3.

Level 3: SSA compared with MAAP BC (MAC of 6.6 m^2 g^{-1}) must be higher than 0.3 and smaller than 1. If no MAAP data is available, use Aethalometer data.

CPC (TSI) – Models 3772 (butanol) and 3787 (water)

Level 1: 1-min aerosol concentration cannot be smaller than 50 cm^{-3} or larger than 10 000 cm^{-3}. No error reported.

Level 2: Apply STP corrections.

Level 3: Intercomparison with other CPC if available. Particle counts comparison must be within 20%.

CPC (GRIMM) – Model 5412 (butanol)

Level 1: 1-min aerosol concentration cannot be smaller than 50 cm^{-3} or larger than 10 000 cm^{-3}. If the following error codes were present 16, 128, 512 and 1024, data were removed.

Level 2: Apply STP corrections.

Level 3: Intercomparison with other CPC if available. Particle counts comparison must be within 20%.

SMPS (TSI) – Models 3081 and 3082

Level 1: 1-min aerosol concentration cannot be smaller than 50 cm^{-3} or larger than 100 000 cm^{-3}. Sheath flow must be higher than 3lpm and smaller than 12lpm. Sample flow must be higher than 0.3lpm and smaller than 2lpm. Status flag must be "Normal Scan". Multiple charge correction must have been applied.

Level 2: Apply STP corrections.

Level 3: Integrated aerosol number must be within 30% CPC measurement. Considering an aerosol density of 1.4 g cm^{-3}, integrated mass must be higher than 50% of TEOM $PM_{2.5}$ and smaller than 120%.

TEOM (Thermo) – Model 1405A

Level 1: 5-min aerosol loading cannot be smaller than -20 µg m^{-3} or larger than 100 µg m^{-3}. Condition parameter must be equal 0, Mode parameter must be below 4.

Level 2: Apply STP corrections.

Level 3: Intercomparison with other TEOM.

O3 (Thermo) – Model 49i

Level 1: Flow in both cells must be >= 0.55 LPM; Intensity in both cells must be >= 75000 Hz. The subsequent data record must also be removed due to the need of instrument stabilization. O3 mixing ratios should be between -50 and 1000 ppb. Eventual clock delay adjustments.

Level 2: Apply calibration factors, baseline adjustments by linear functions, remove periods of maintenance.

Level 3: Intercomparison with other O3 monitor if available. Data must agree within 20%.

SO2 (Thermo) – Model 43iTLE

L1:- Flow must be >=0.38 LPM; Reaction chamber temperature (rctt) must be >=43oC, Lamp intensity (lmpi) must be >=88%. The subsequent data record must also be removed due to the need of instrument stabilization. SO2 mixing ratios should be between -10 and 100 ppb. Eventual clock delay adjustments.

L2: Apply calibration factors, baseline adjustments by linear functions, remove periods of maintenance.

Level 3: Intercomparison with other SO2 monitor if available. Data must agree within 20%.

NO2-CAPS (Aerodyne)

Level 1: Temp>315K; Pressure>700Torr; -10<NO2<200ppb; remove pump alarm and baseline flagged data points

Level 2: remove calibration and maintenance periods and eventually corrected the baseline.

Level 3: Intercomparison with other NO2 monitor if available. Data must agree within 20%.

CO+N2O (Los Gatos)

Level 1: Reported CO and N2O mixing ratios are relative to ambiente air (not dry air), and must be in the ranges: -1<CO<3 ppm and -1<N2O<50 ppm. The following conditions should apply for instrument internal variables: 0.67<=AIN5>=0.68; LTC0>0.2 V; 26<=GasT<=40 oC. Eventual clock delay adjustments based on the correlation to BC concentration timeseries.

Level 2: Apply calibration factors, baseline adjustments.

Level 3: Intercomparison with other CO/N2O monitor if available. Data must agree within 20%.

CO+CO2 (PICARRO)

Level 1: Reported CO and CO2 mixing ratios are relative to ambient air (not dry air), with: -0.1<CO<1 ppm and "alarm"=0. Periods of silica change must be removed.

Level 2: remove calibration and maintenance periods and eventually correct the baseline.

Level 3: Intercomparison with other CO/CO2 monitor if available. Data must agree within 20%.

ACSM (Aerodyne)

Due to its complexity, only level 3 data will be made available of this instrument.

PTR-Q-MS (Ionicon)

Due to its complexity, only level 3 data will be made available of this instrument.

References

Anderson, T.L., Ogren, J. a., 1998. Determining Aerosol Radiative Properties Using the TSI 3563 Integrating Nephelometer. Aerosol Sci. Technol. 29, 57–69. doi:10.1080/02786829808965551

Müller, T., Henzing, J.S., de Leeuw, G., Wiedensohler, A., Alastuey, A., Angelov, H., Bizjak, M., Collaud Coen, M., Engström, J.E., Gruening, C., Hillamo, R., Hoffer, A., Imre, K., Ivanow, P., Jennings, G., Sun, J.Y., Kalivitis, N., Karlsson, H., Komppula, M., Laj, P., Li, S.-M., Lunder, C., Marinoni, A., dos Santos, S., Moerman, M., Nowak, A., Ogren, J.A., Petzold, A., Pichon, J.M., Rodriquez, S., Sharma, S., Sheridan, P.J., Teinilä, K., Tuch, T., Viana, M., Virkkula, A., Weingartner, E., Wilhelm, R., Wang, Y.Q., 2011. Characterization and intercomparison of aerosol absorption photometers: result of two intercomparison workshops. Atmos. Meas. Tech. 4, 245–268. doi:10.5194/amt-4-245-2011

Müller, T., Laborde, M., Kassell, G., Wiedensohler, A., 2011. Design and performance of a three-wavelength LED-based total scatter and backscatter integrating nephelometer. Atmos. Meas. Tech. 4, 1291–1303. doi:10.5194/amt-4-1291-2011

Rizzo, L. V., Correia, a. L., Artaxo, P., Procópio, a. S., Andreae, M.O., 2011. Spectral dependence of aerosol light absorption over the Amazon Basin. Atmos. Chem. Phys. 11, 8899–8912. doi:10.5194/acp-11-8899-2011

Weingartner, E., Saathoff, H., Schnaiter, M., Streit, N., Bitnar, B., Baltensperger, U., 2003. Absorption of light by soot particles: determination of the absorption coefficient by means of aethalometers. J. Aerosol Sci. 34, 1445–1463. doi:10.1016/S0021-8502(03)00359-8

C.2 Descrição do Conjunto de Dados com base no perfil de aplicação

A seguir tem-se a descrição do conjunto de dados utilizado no estudo de caso, com base no perfil de aplicação proposto para a veracidade. Este documento foi apresentado ao grupo B.

Escola Politécnica da Universidade de São Paulo - Grupo de Estudo, Pesquisa e Extensão em Big Data

Go Amazon Dataset - Ozone - T3

http://creativecommons.org/licenses/by-nc/3.0/ *a* dct:LicenseDocument	
foaf:name	CC BY
dct:type	attribution
owl:versionInfo	3.0

http://lfa.if.usp.br/ftp/public/LFA Processed Data/ *a* dcat:catalog	
dct:description	LFA-USP GoAmazon Data Repository
foaf:homepage	http://lfa.if.usp.br *a* foaf:Agent foaf:homepage http://lfa.if.usp.br dct:identifier LFA foaf:mbox lfa-goamazon@googlegroups.com foaf:name Laboratory of Atmospheric Physics dct:type Research Group
dct:issued	2016-04-25
dct:modified	2016-04-24
dct:publisher	http://lfa.if.usp.br
dcat:themeTaxonomy	http://lfa.if.usp.br/data-theme *a* skos:ConceptScheme
dct:title	GoAmazon LFA-USP Data Repository
dcat:dataset	http://lfa.if.usp.br/ftp/public/LFA Processed Data/T3 Manacapuru/Level3/T3 Thermo 49i Level3 *a* dcat:dataset

http://lfa.if.usp.br/ftp/public/LFA Processed Data/T3 Manacapuru/Level3/T3 Thermo 49i Level3 *a* dcat:dataset	
dct:accessRights	public
dcat:contactPoint	http://lfa.if.usp.br/artaxo *a* vcard:kind vcard:fn Paulo Artaxo vcard:hasEmail artaxo@usp.br vcard:nickname Artaxo
pav:contributedBy	http://lfa.if.usp.br/lrizzo *a* vcard:kind vcard:fn Luciana Rizzo vcard:hasEmail lrizzo@unifesp.br vcard:nickname Luciana
pav:createdBy	http://www.researcherid.com/rid/E-8874-2010
pav:createdOn	2015-02-09
dct:description	Time series of ozone concentration, in ppb, at the T3 station of the GoAmazon project.
pav:hasCurrentVersion	http://lfa.if.usp.br/ftp/public/LFA Processed Data/T3 Manacapuru/Level3/T3 Thermo 49i Level3#1
dct:identifier	http://lfa.if.usp.br/ftp/public/LFA Processed Data/T3 Manacapuru/Level3/
dct:issued	2015-02-09
dcat:keyword	Ozone *and* Aerosol
dct:language	http://publications.europa.eu/mdr/authority/language/ENG
dct:license	http://creativecommons.org/licenses/by/3.0/
foaf:page	http://lfa.if.usp.br/
dcat:provenance	The T3 station is in a forest region, where biogenic emissions predominate in the rainy season (Jan-Jun) and burnings in the dry season (Jul-Dec). All seasons are influenced, to a greater or lesser extent, by dry season burning emissions. The concentrations and properties of pollutants change drastically as a function of the predominant sources of pollution and the aging processes of pollutants as they are transported into the atmosphere. This dataset was processed in three steps, namely: Level 1 - Flow in both cells must be >= 0.55 LPM; Intensity in both cells must be >= 75000 Hz. The subsequent data record must also be removed due to the need of instrument stabilization. O3 mixing ratios should be between -50 and 1000 ppb. Eventual clock delay adjustments. Level 2 - apply calibration factors, baseline adjustments by linear functions, remove periods of maintenance. Level 3 - intercomparison with other O3 monitor if available. Data must agree within 20%.
pav:providedBy	http://lfa.if.usp.br
dct:source	http://lfa.if.usp.br/ftp/public/LFA Processed Data/GoAmazon USP Data processing.pdf
dct:spatial	http://www.geonames.org/3663529
dcat:theme	http://lfa.if.usp.br/data-theme/GoAmazon *a* skos:Concept skos:prefLabel Ozone , GoAmazon *and* Aerosols
dct:title	T3_Thermo_49i_Level3_5min
void:vocabulary	http://purl.org/dc/terms/
dcat:datasetVersion	http://lfa.if.usp.br/ftp/public/LFA Processed Data/T3 Manacapuru/Level3/T3 Thermo 49i Level3#1 *a* vera:datasetVersion fgdc:complete Data transformations steps are described in http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/GoAmazon_USP_Data_processing.pdf. Between 20140818 and 20140822 there was an increase in the number of fire outbreaks

	in the period, associated with a change in the direction of the winds which brought air masses from the deforestation arc (south-south-east of Amazonia part). This has led to a high concentration of ozone in this period. More information on fires in the period can be found in INPE Burning Database.
pav:createdBy	http://www.researcherid.com/ rid/ E-8874-2010
pav:createdOn	2015-02-09
dct:description	Version 01: Time series of Ozone Concentration at the T3 station of the GoAmazon project.
dcat:distribution	**http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3 5min** ***a*** **dcat:distribution**

dct:accrualPeriodicity	http://purl.org/ cld/ freq/ irregular
dcat:byteSize	3700000
dct:conformsTo	ttp://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#1
pav:createdBy	http://www.researcherid.com/ rid/ E-8874-2010
pav:createdOn	2015-02-09
dct:description	5min time series csv file
dcat:downloadURL	http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3 5min.csv
fgdc:eainfo	**http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#EAInfo** ***a*** **fgdc:eainfo**

fgdc:attr	**http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#AttrO3** ***a*** **fgdc:attr**
fgdc:attrdef	O3 concentration in ppb
fgdc:attrlabl	O3
	,
	http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#AttrUnixTime ***a*** **fgdc:attr**
fgdc:attrdef	date and hour in Unix Time format, since 1970-01-01
fgdc:attrlabl	Unix Time (UTC)
	and
	http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#AttrTime ***a*** **fgdc:attr**
fgdc:attrdef	date and hour in UTC format
fgdc:attrlabl	Time (UTC)
fgdc:overview	**http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#Overview** ***a*** **fgdc:overview**
fgdc:eaover	O3 concentration data were collected using Thermo Fisher Scientific's 49i UV Photometric Analyzer. In T3, the device was installed at 5 meters in height. Its lower limit of detection is 1 ppb (parts per billion). Observations were made at 5 minute intervals

dct:format	text/csv
dct:identifier	http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3 5min
dct:issued	2015-02-09
dct:title	T3_Thermo_49i_Level3_5min

and

http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3 30min ***a*** **dcat:distribution**

dct:accrualPeriodicity	http://purl.org/ cld/ freq/ irregular
dcat:byteSize	639000
dct:conformsTo	ttp://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#1
pav:createdBy	http://www.researcherid.com/ rid/ E-8874-2010
pav:createdOn	2015-02-09
dct:description	30min extracted mean from 5min file. time series csv file
dcat:downloadURL	http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3 30min.csv
fgdc:eainfo	http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#EAInfo

dct:format	text/csv
dct:identifier	http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3 30min
dct:issued	2015-02-09
dct:title	T3_Thermo_49i_Level3_30min

dct:identifier	http://lfa.if.usp.br/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#1
dct:isVersionOf	ttp://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3
dct:issued	2015-02-09
pav:previousVersion	http://lfa.if.usp.br/ ftp/ public/ LFA Processed Data/ T3 Manacapuru/ Level3/ T3 Thermo 49i Level3#1
dct:temporal	http://lfa.if.usp.br/ PeriodOfTime
dct:title	T3_Thermo_49i_Level3 - Version 01
pav:version	1

http://lfa.if.usp.br/ periodOfTime ***a*** **dct:periodOfTime**

schema:endDate	2015-01-01
schema:startDate	2014-01-01

C.3 Respostas obtidas

A seguir são apresentadas as respostas obtidas no estudo de caso apresentado no Capítulo 4.

RESPOSTAS - GRUPO A

Área de formação	Grau de instrução	1. Descreva o quê está sendo representado pelo conjunto de dados.	2. Em que período os dados foram coletados?	3. Qual(is) instrumento(s) foi(ram) utilizados para coletar os dados?	4. Quem publicou o conjunto de dados?	5. Você compreendeu como o provedor de dados gerou o conteúdo dos dois arquivos .csv disponibilizados?	6. O que significa cada atributo (coluna) dos arquivos de dados disponibilizados?	7. Diante de sua análise, se posicione e justifique sobre a existência ou não de comportamentos adversos no conjunto de dados.	Suponha que você necessita utilizar os dados dos arquivos apresentados em um experimento científico. Com base em sua análise, o quão seguro você está em utilizar esses dados em seu experimento?	Justifique a sua resposta anterior.	Caso tenha algum comentários sobre a sua avaliação, por favor utilize o espaço abaixo.
Ciência da Computação	Graduação	Medições de valores médios de O3 por intervalo de tempo.	01/01/2014 a 01/01/2015	Thermo 49i	Pelo pdf informado o Artaxo Group. Não há informação no dataset.	Sim	Data da medição, UTC da medição e valor médio de O3 medido	Existem valores NaN nas medições, o que não é explicado no documento anexo.	Sinto-me razoavelmente seguro em utilizar o conjunto de dados	Falta da informação de quem publicou e valores NaN não explicados	
Engenheiro Eletricista	Graduação	O conjunto de dados representa o valor médio de ozônio (O3) medido para diferentes intervalos de tempo (5 min e 30 min).	2014/15	Thermo – Model 49i	Artaxo Group da USP	Sim	O dois primeiros atributos são o tempo no horário UTC (formatado de forma legível na primeira coluna e em segundo no segundo caso). O terceiro atributo representa a média de ozônio medida naquele intervalo de tempo em ppb.	A primeira vista não, pois todos os dados se encontram dentro do limite especificado e o número de medições do time series de 30 min é 6 vezes menor do que o número de medições do time series de 5, o que é o esperado.	Sinto-me razoavelmente seguro em utilizar o conjunto de dados	Me sinto seguro em utilizar os dados sim, entretando, seria interessante aplicar mais ferramentas para entender melhor os dados e fazer uma análise exploratória mais complexa antes de afirmar com 100% de certeza.	
Administração	Mestrado	Coleta de arquivos, unidade consumida e tempo.	01-01-2014 a 31-12-2014	não sei informar.	GoAmazon	Não	Tempo de resposta e consumo	Houve variação, mas não sei informar o que representa.	Não consigo formar uma opinião sobre a minha segurança em utilizar o conjunto de dados	Por não conhecer com profundidade os dados, não é possível dar uma opinião mais concreta.	
Bioinformatica	Mestrado	Tempo versus concentração de ozonio	de 1/1/2014 ate 31/12/2014	Thermo 49i	Paulo Artaxo	Sim	Momento da coleta (human readable), Momento da coleta (machine readable), e concentração de ozonio.	A média dos tempos em meia hora do arquivo 5min não equivale exatamente a um pounto no arquivo 30min.	Sinto-me razoavelmente seguro em utilizar o conjunto de dados	Apesar de alguns dados serem negarticos e de o valor minimo por dia aumentar linearmente, os dados podem ser usados após ajustes pequenos	
Eng. Elétrica / Computação	Doutorado	a concentração de ozônio a cada instante (a cada 30 min / 5 min) num certo local.	nos primeiros dois arquivos ao longo de 2014 (diariamente, a cada 30 min / 5 min)	nao sei	USGS	Sim	de fato das varias colunas apenas 3 foram preenchidas: 1) dia/hora, 2) tempo na representacao UTC, 3) concentracao de O3	não cheguei a fazer a analise, mas deveria comparar as duas tabelas para ver se há inconsistencias (por exemplo no mesmo instante valores de O3 bem discrepantes entre as duas tabelas, ou mesmo a presenca de um valor numa e não na outra)	Sinto-me seguro em utilizar o conjunto de dados	Se por um lado eu nao conheca de fato a fonte dos dados, eu considero que seja confiavel. Mas a questao aqui me parece ser quanto a possibilidade de avaliar a coerencia dos dados a partir de comparacoes entre diferentes arquivos (fontes, formas de medida, ...) E neste sentido me parece que eventuais discrepancias (que podem ocorrer) seriam facilmente identificaveis. E quando nao verificadas, reforçam a qualidade dos dados.	
Ciência da Computação	Doutorado	Dados de Ozonio	2014-2015	instrumentos Thermo	LFA	Não	Não consegui compreender com base na documentação fornecedida	Não encontrei valores adversos.	Não estou seguro em utilizar o conjunto de dados	A documentação não permite compreender completamente como os dados foram gerados. Fiquei com bastante dúvidas sobre como os dados foram gerados e manipulados. Não compreendi muito bem a diferença entre os dois arquivos de dados.	
Computação	Mestrado	Dados coletados no goAmazon sobre Ozonio	2014 e 2015	Não pude identificar o equipamento.	Artaxo Group	Não	Fiz uma análise manual, é o tempo, unix time e concentração de O3	Não notei, apenas valores ausentes.	Não estou seguro em utilizar o conjunto de dados	Como não sou da área, não me sinto seguro. a documentação é muito especifica e não permite que eu possa compreende-la. Sinto falta de entender qual o objetivo do conjunto de dados. No evento foi muito enfatizado a importância dos metadados e de um conjunto de documentações, acho que isto poderia ter sido utilizado.	Acho que poderíamos ter maiores informações sobre o conjunto de dados para poder avalia-lo.

RESPOSTAS - GRUPO A

Área de formação	Grau de instrução	1. Descreva o quê está sendo representado pelo conjunto de dados.	2. Em que período os dados foram coletados?	3. Qual(is) instrumento(s) foi(ram) utilizados para coletar os dados?	4. Quem publicou o conjunto de dados?	5. Você compreendeu como o provedor de dados gerou o conteúdo dos dois arquivos .csv disponibilizados?	6. O que significa cada atributo (coluna) dos arquivos de dados disponibilizados?	7. Diante de sua análise, se posicione e justifique sobre a existência ou não de comportamentos adversos no conjunto de dados.	Suponha que você necessita utilizar os dados dos arquivos apresentados em um experimento científico. Com base em sua análise, o quão seguro você está em utilizar esses dados em seu experimento?	Justifique a sua resposta anterior.	Caso tenha algum comentários sobre a sua avaliação, por favor utilize o espaço abaixo.
Físico	Mestrado	O3 - concentração	2014	Thermo	USGS	Não	Entendi apenas 2: tempo (primeira coluna) e O3 (segunda coluna)		Não estou seguro em utilizar o conjunto de dados	Não me sinto totalmente seguro, gostaria de saber mais sobre quem criou o conjunto, quais seus trabalhos, se houve alguma premissa antes da coleta.	
Engenharia da Computação	Mestrado	At Lv3, data is validated against other instruments. These documents represent the mean intercomparison with other ozone monitors. It should not range more than 20%.	Frequency: 5min and 30min. Period: From 01/01/2014 to 01/01/2015	An UV Photometric O3 Analyzer sent the data, and a computer received it for further analysis.	Lv1 data is published by Photometric O3 Analyzer, and Lv3 data is already processed and compared with other O3 Analyzers.	Sim	Fisrt column is UTC date formated as dd-mmm-yyyy HH:MM; Second Column is epoch (or Unix) time, i.e., the number of seconds since 01-Jan-1970 00:00. Third Column is the O3 mean in ppb compared to other O3 analyzers.	There are a lot of data missing, basically in the same period, on both files. In the 5min there are 61.056 (58,08%) and in the 30min 10.176 (58,08%). Probabilily these NULL values has some meaning and I need more information about this. Lv3 say that data must agree within 20%. Up to the 3rd quartile, mean is aproximatelily 18, but there are some high values like 94.10 (5min) and 86.32(30min), so there are unexpected behaviour.	Sinto-me pouco seguro em utilizar o conjunto de dados	If I use just the documentation available at http://lfa.if.usp.br/ftp/public/LFA_Processed_Data/GoAmazon_USP_Data_processing.pdf there are not enought information to use data. I do not know what was its methodology, description are very general, there are some values missing without explanation and so on.	
Engenharia Elétrica	Mestrado	São medidas de Ozônio (O3) na atmosfera realizadas no site Site: T3, Instrumento: Thermo, Modelo:49i . Arquivo 1 e Arquivo 2 diferem apenas nos intervalos de escrita de dados, 5 e 30 minutos, respectivamente. O valor gravado é a média de medidas desde a última gravação (5 ou 30 minutos). Os dados são	Durante todo o ano de 2014	Instrumento: Thermo, Modelo:49i .	GoAmazon	Sim	Data-hora no formato DD-MM-YYYY hh:mm, data-hora no formato timestamp unix e o valor medido de 03 em partes por bilhão	Há valores negativos de Ozônio ppb, o que conceitualmente não é aceitável, visto que essa unidade representa proporção do ozônio em relação aos demais gases atmosféricos. Exemplo: 20-Apr-2014 02:00; 1397959200; -0.07. Como os valores negativos são sempre muito próximos de zero, considero que seja erro intrínseco à construção do instrumento de medição, dentro da tolerância do próprio instrumento e da validação nível 3.	Sinto-me razoavelmente seguro em utilizar o conjunto de dados	Em pesquisa rápida na Internet, verifiquei que a quantidade de ozônio na atmosfera é de 70 ppb, coerente com os valores medidos. Faria apenas análise prévia para encontrar e remover outliers.	
Física Computacional	Graduação	São os dados capturados de Aerosol do projeto GoAmazon	De 01/01/2014 até 01/01/2015	5012 Multi Angle Absorption Photometer (MAAP) black carbon monitor	GoAmazon in situ data processing	Sim	Timestamp da coleta, Timestamp em formato numérico e o Dado Coletado		Sinto-me razoavelmente seguro em utilizar o conjunto de dados	O Tratamento de dados deve fazer parte da análise dos mesmos. Outro ponto importante é que muitas vezes a o descasamento de informação em si pode trazer ganhos para a análise.	
Ciência Computação	Graduação	Dados de sensores de 5 estações na Amazônia, mais precisamente do sensor O3 (Thermo).	2014/2015	AETHELOMETER (Magee), MAAP (Thermo), Nephelometer (Ecotech), Nephelometer (TSI) , CPC (TSI) , CPC (GRIMM), SMPS (TSI), TEOM (Thermo), O3 (Thermo) ,SO2 (Thermo) , NO2-CAPS (Aerodyne), CO+N2O (Los Gatos) ,CO+CO2 (PICARRO) , ACSM (Aerodyne)e PTR-Q-MS (Ionicon)	Artaxo Group	Sim	Time (UTC) data/hora da medicao, O3 Mean (ppb) dados do sensor, mas não consigo detalhar seu objetivo.	Na coluna O3 Mean (ppb) os dados são todos NAN (Not a Number), o que impossibilita qualquer analise.	Não estou seguro em utilizar o conjunto de dados	A informacao NAN não permite analise	

RESPOSTAS - GRUPO B

Área de formação	Grau de instrução	1. Descreva o quê está sendo representado pelo conjunto de dados.	2. Em que período os dados foram coletados?	3. Qual(is) instrumento(s) foi(ram) utilizados para coletar os dados?	4. Quem publicou o conjunto de dados?	5. Você compreendeu como o provedor de dados gerou o conteúdo dos dois arquivos .csv disponibilizados?	6. O que significa cada atributo (coluna) dos arquivos de dados disponibilizados?	7. Diante de sua análise, se posicione e justifique sobre a existência ou não de comportamentos adversos no conjunto de dados.	Suponha que você necessita utilizar os dados dos arquivos apresentados em um experimento científico. Com base em sua análise, o quão seguro você está em utilizar esses dados em seu experimento?	Justifique a sua resposta anterior.	Caso tenha algum comentários sobre a sua avaliação, por favor utilize o espaço abaixo.
Ciência da Computação	Graduação	Série temporal de concentração de ozônio na estação T3 do projeto GoAmazon	01/01/2014 - 01/01/2015	Thermo	Paulo Artaxo	Sim	A primeira coluna apresenta a data e hora em que a coleta foi feita, a segunda a o tempo relativo desde o início e a terceira a média de ozônio.	Nesta etapa não pude encontrar nehum comportamento adverso no conjunto de dados.	Sinto-me razoavelmente seguro em utilizar o conjunto de dados	Consegui compreender o que eles estão medindo, porém não tenho certeza das análises que posso realizar a partir deles nem quais métricas são mais ou menos significativas.	A documentação dos dados e seus metadados permitem uma rápida avaliação do conteúdo disponível e dá mais segurança no uso dos mesmos em um momento inicial.
Eng. Civil	Mestrado	A concentração média de Ozônio (O3) em ppb, em determinado posto de medição (estação T3), já no nível 3 de processamento, em intervalos de 5 e 30 minutos.	01/01/2014 - 31/12/2014	Células de medição instaladas na estação.	LFA USP	Sim	Time (UTC) - Data da medição; Unit Time (UTC) - Data da medição (UNIX); O3 Mean (ppb) - Concentração de Ozônio em ppb.	A princípio, não deveria haver valores negativos (dado que se trata da concentração de determinado poluente). Seria necessário identificar possíveis outliers dos dados para o controle de qualidade. Há também pontos de informação nulos (NaN) mas provavelmente são oriundos dos processamentos de nível 1 e 2.	Não consigo formar uma opinião sobre a minha segurança em utilizar o conjunto de dados	Não tenho experiência prévia nesta área específica de conhecimento científico, então não consigo elaborar um julgamento a respeito da qualidade dos dados. Se fosse utilizá-los, buscaria opiniões de especialistas, bem como de outras fontes de informação que convergissem para a análise a ser realizada. Quanto aos dados em si, necessitaria verificar todas as etapas de pré-processamento para garantir a sua adequabilidade à questão em análise. Potencialmente, preferiria optar por dados brutos, em vez de dados já processados.	
Engenharia da Computação	Graduação	Uma série temporal da Concentração de ozônio, em partículas por milhão, na estação T3 do projeto GoAmazon	Jan-Dez 2014	T3 Thermo 49i nível 3	Paulo Artaxo e Luciana Rizzo	Sim	Time e Unit Time UTC (Momento da Coleta de Dados), O3 Mean ppb (Concentração de O3 em p.p.b)	Existem vários dados do dataset sem o valor da concentração de ozonio	Não consigo formar uma opinião sobre a minha segurança em utilizar o conjunto de dados	Faltam informações sobre a calibração do sensor, entre outras	
Ciência da computação	Graduação	Medições em partes por bilhão da concentração de ozônio (O3) na estação T3 do projeto GOAmazon.	Intervalos de 5 minutos e de 30 minutos.	As medições foram obtidas pelo equipamento Thermo Fisher Scientific's 49i UV Photometric Analyzer.	Paulo Artaxo e Luciana Rizzo	Sim	As duas primeiras colunas são dois padrões de mostrar a data e hora e a terceira coluna mostra a concentração média de ozônio em partes por bilhão no intervalo de tempo.	Existem valores médios negativos de concentração de O3.	Sinto-me pouco seguro em utilizar o conjunto de dados	Por não entender os valores negativos de concentração média de O3	
Computação e Matemática Aplicada	Doutorado	Evolução temporal da concentração de Ozonio no site T3 de Manacapuru	01/01/2014 a 01/01/2015	Nao encontrei o instrumento, apenas que foram instrumentos in situs	Instituto de Fisica da USP	Não	data (formato dd-mm-aa); hora (formato hh:mm) data em formato UTC,media da concentração de ozonio (em ppb)	Dados NaN dificultam a análise. Os dados existentes permitem uma boa análise.	Sinto-me seguro em utilizar o conjunto de dados	Devido a origem confiável da publicação e a documentação disponível.	
Ciência da Computação	Mestrado	Uma série temporal de concentração de ozônio	01-01-2014 até 01-01-2015	Thermo Fisher Scientific's 49i UV Photometric Analyzer	Laboratory of Atmospheric Physics	Sim	Timestamp formatado; timestamp em segundos; concentração média de ozônio em ppb	Existem valores faltantes (NaN) na série temporal, provavelmente devido a períodos de calibração e/ou manutenção no instrumento	Sinto-me razoavelmente seguro em utilizar o conjunto de dados	Os metadados são claros quanto a origem dos dados, instrumento, unidades de medida e etc. Os documentos citados (pdfs) auxiliam na descrição dos dados.	Não houve tempo suficiente para dedicação ao survey, devido às demais atividades do workshop.
Ciência da Computação	Doutorado	Aparentemente dados relativos a condições temáticas	01/01/14 a 31/12/14	Não consegui achar essa informação	http://lfa.if.usp.br	Não	não tenho certeza, mas parece-me que eram informações relativas a 03 na camada de ozônio	Os números apresentados são bastante próximos e parecidos. Não me parece que há algum com comportamento adverso. Uma das colunas apresenta sempre o mesmo valor: NaN (não sei o que isso representa)	Não estou seguro em utilizar o conjunto de dados	Não consegui ter certeza do que os dados representam.	Com base nas informações apresentadas (arquivos e informações de consulta) não consegui entender o significado dos dados apresentados, o que representam e qual a importância dos mesmos. Não teria como utilizá-los. Deve haver uma melhor definição dos mesmos na questão de Apresentação da Informação, um dos focos da Usabilidade de Ambientes Computacionais.

RESPOSTAS - GRUPO B

Área de formação	Grau de instrução	1. Descreva o quê está sendo representado pelo conjunto de dados.	2. Em que período os dados foram coletados?	3. Qual(is) instrumento(s) foi(ram) utilizados para coletar os dados?	4. Quem publicou o conjunto de dados?	5. Você compreendeu como o provedor de dados gerou o conteúdo dos dois arquivos .csv disponibilizados?	6. O que significa cada atributo (coluna) dos arquivos de dados disponibilizados?	7. Diante de sua análise, se posicione e justifique sobre a existência ou não de comportamentos adversos no conjunto de dados.	Suponha que você necessita utilizar os dados dos arquivos apresentados em um experimento científico. Com base em sua análise, o quão seguro você está em utilizar esses dados em seu experimento?	Justifique a sua resposta anterior.	Caso tenha algum comentários sobre a sua avaliação, por favor utilize o espaço abaixo.
Física	Mestrado	A concentração de ozônio, em ppb (partes por bilhão), na estação T3 do projeto GoAmazon	2014-01-01 a 2015-01-01	Thermo Fisher Scientific's 49i UV Photometric Analyzer	Laboratory of Atmospheric Physics (LFA) - IF-USP	Não	A primeira coluna corresponde ao momento no qual o dado foi coletado (data/tempo em UTC). A segunda coluna corresponde à unidade de tempo em segundos desde 01-Jan-1970 00:00 (uma outra forma de representar a primeira coluna). A terceira coluna corresponde ao valor medido de concentração de ozônio (em ppb).	Parece não haver consistência entre os conjuntos de dados. A diferença entre eles é o intervalo de tempo no qual os dados são apresentados. Neste sentido, o valor observado para o momento 31-Dec-2014 23:30 (ou 1420068600), por exemplo, deveria ser o mesmo para os dois conjuntos. No entanto, observa-se que, para um mesmo momento, os valores são distintos. No caso do momento 1420068600 encontramos o valor 6.76 na base 30' e 7.13 na base 5'.	Não estou seguro em utilizar o conjunto de dados	Não me sentiria confortável pois os conjuntos de dados me pareceram discrepantes.	
Física	Mestrado	Série temporal de concentração de ozonio.	2014-01-01 - 2015-01-01	Thermo Fisher Scientific's 49i UV Photometric Analyzer	Laboratory of Atmospheric Physics	Sim	Time - Time (UTC), UnixTime - Unix Time (UTC), O3 concentration in ppb	Não encontrei, mas vi que foi feita uma ressalva sobre altas concentrações em um período especifico, em função de ventos e queimadas	Sinto-me seguro em utilizar o conjunto de dados	Acredito que a descrição dos dados foi muito benéfica. Embora alguns pontos possam ser melhor esclarecidos, eu pude facilmente encontrar as respostas para as perguntas efetuadas. A página poderia ser melhorada para que ficasse mais fácil de utilizar, mas em geral eu pude compreender o proposito do conjunto de dados	
Tecnologia da Informação	Pós-Graduado em Estatística Aplicada	Concentração de Ozônio (O3) coletado na estação T3 do Projeto GoAmazon	2014	Thermo Fisher Scientific's 49i UV Photometric Analyzer	LFA-USP ou Laboratory of Atmospheric Physics-USP	Sim	Horário no formato UTC, Horário no formato Unix e Concentração de Ozônio (em ppb)	A concentração de ozônio é geralmente maior no período da tarde. No mesmo horário, a medição de 30 em 30 minutos gera uma informação menor do que na medição de 5 em 5 minutos.	Sinto-me razoavelmente seguro em utilizar o conjunto de dados	O material fornece um conjunto objetivo de informações sobre o experimento e com a aplicação de alguns filtros e rotinas de tratamento de dados, além de rotinas de machine learning para completar dados faltantes, é possível tirar algumas conclusões sobre a Concentração de Ozônio ao longo do ano. Com outro dataset do mesmo período, é possível correlacionar os eventos, para identificar possíveis causas e efeitos colaterais.	
Ciência Computação	Mestrado	A concentração de Ozonio durante o ano de 2014 na Estaçao T3	Durante o ano de 2014	Termostato Fisher Scientific's 49i UV Photometric Analyzer	O FLA	Sim	Da primeira coluna para ultima: Data da coleta, a etapa do processo de experimento, a versão de processamento; os dados usados da primeira etapa para estabilização do termostato; a estação ou campo de coleta; o modelo do termostato; o serial number do termostato; e um link para ver maiores informações sobre os dados.	O unico problema que achei foi a coluna O3 mean, ser um NaN, dificultando a leitura e identificação da informação ali presente.	Sinto-me pouco seguro em utilizar o conjunto de dados	Devido não compreender os dados presentes em um dos campos.	
Ciências da Computação	Graduação	Representa o nível de concentação de ozônio (O3) medidos em períodos de 5 min e de 30 min.	Entre 01/01/2014 e 01/01/2015	Thermo Fisher Scientific's 49i UV Photometric Analyzer, ou Analisador Fotométrico UV 49i da Thermo Fisher Scientific.	O Laboratório de Física Atmosférica (LFA) da Universidade de São Paulo (USP).	Sim	A primeira coluna é da data/hora (UTC) da coleta daquele dado. A segunda coluna é a unidade de tempo (UTC). A terceira coluna é a média medida da concentração de ozônio (O3) naquele dado período.	Existem vários dados coletados que não seguem o que era esperado para a medição, fugindo bastante (ou dando erro mesmo, NaN) do que foi medido antes e depois dos mesmos. Por exemplo: existem valores que estão negativos e de repente se tornam positivos em apenas um período de medição. Isso denota anomalia, pois não houve uma mudança gradual, para baixo ou para cima, que indicaria de fato que estaria havendo uma mudança ambiental real. Uma mudança de um período apenas, drástica, nesses casos, podem ser consideradas adversas.	Sinto-me seguro em utilizar o conjunto de dados	O conjunto de dados poderá ser filtrado com base no que ser pode apreender e aprender dos mesmos, do comportamento das medidas. Assim, pode-se estudar melhor se os fenômenos observados pelos dados seriam mapeáveis ou não em algum tipo de realidade adversa, desconhecida ou não. O uso desse conhecimento do ambiente origem dos dados e da base estatística, pode nos dar a segurança. No caso em questão, podemos ficar tranquilos de que os resultados corretos das medidas ocorrem em maior número do que os adversos.	Acredito que sou bastante iniciante em Ciência dos Dados e que minha avaliação pode ser equivalente. No entanto, espero ter colaborado e, de fato, gostei muito de ter feito esse trabalho! Um grande abraço e continuem no bom caminho do ensino teórico e prático sempre em paralelo!
Computação	Mestrado	Medição da Concentração de Ozônio em Partes por Bilhão (ppb)	01/01/2014 - 01/01/2015	Thermo Fisher Scientific's 49i UV Photometric Analyzer	Laboratory of Atmospheric Physics	Sim	Coluna O3 Mean: valor de Ozônio coletado. Unit Time: Horário?!?!	Não sei responder a questão	Não consigo formar uma opinião sobre a minha segurança em utilizar o conjunto de dados	Falta de conhecimento específico na área dos dados e falta de um exemplo de utilização para analisar a aplicação dos dados	
Computacao	Graduação	Medicoes	Ano de 2014	Thermo	IF USP no GOAMAZON	Não	dia, horario, media de ozonio	Muitas medicoes ausentes	Sinto-me razoavelmente seguro em utilizar o conjunto de dados	O fato de ter poucas variaveis facilita a utilizacao	