UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS AVELLAR DE BARROS

DESENVOLVIMENTO DE UMA PLATAFORMA DE QUALIDADE DE DADOS AGNÓSTICA A DADOS

RIO DE JANEIRO
2022

MATHEUS AVELLAR DE BARROS

DESENVOLVIMENTO DE UMA PLATAFORMA DE QUALIDADE DE DADOS AGNÓSTICA A DADOS

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Geraldo Zimbrão da Silva

RIO DE JANEIRO
2022

CIP - Catalogação na Publicação

B277d

Barros, Matheus Avellar de
Desenvolvimento de uma plataforma de qualidade de dados agnóstica a dados / Matheus Avellar de Barros. -- Rio de Janeiro, 2022.
46 f.

Orientador: Geraldo Zimbrão da Silva.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Computação, Bacharel em Ciência da Computação, 2022.

1. qualidade de dados. 2. análise de dados. 3. protótipo. 4. web. I. Silva, Geraldo Zimbrão da, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

MATHEUS AVELLAR DE BARROS

DESENVOLVIMENTO DE UMA PLATAFORMA DE QUALIDADE DE DADOS AGNÓSTICA A DADOS

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 11 de março de 2022

BANCA EXAMINADORA:

Prof. Geraldo Zimbrão da Silva,
D.Sc. (COPPE/UFRJ)

Prof.ª Silvana Rossetto,
D.Sc. (IC/UFRJ)

Prof. Geraldo Bonorino Xexéo,
D.Sc. (COPPE/UFRJ)

AGRADECIMENTOS

Agradeço primeiramente ao meu orientador, professor Geraldo Zimbrão, pelo apoio e conselhos dados não somente durante o projeto, mas durante minha trajetória acadêmica até hoje. Também agradeço ao professor Mitre Costa pelo apoio e estímulo na fase inicial de pesquisa e investigação de bons temas para a monografia.

"Speed of reporting should not be confused with either machine infallibility or accuracy, for the two are far from equal. Machines are incredibly fast in performing the functions for which they are designed; BUT—should defective data be inserted into the machine, it will produce mistakes just as rapidly."

Capt. Carl M. Guelzo

"Automation in Support of the Battlefield"
National Defense Transportation Journal
v. 15, n. 4, p. 53, jul./ago.1959

RESUMO

Na Era da Informação, nada é mais pervasivo na vida digital do que dados. Dados são atestações factuais elementares da realidade, a partir dos quais é possível construir informações e gerar conhecimento. Acima de tudo, dados são usados para auxiliar a tomada de decisões. Assim, para evitar prejuízos, é imprescindível que os dados utilizados possuam qualidade. Através dos anos, diversas especificações definiram o que faz um dado ter qualidade. Contudo, um campo menos explorado lida com a maneira como deve ser realizada a aferição de qualidade dos dados. Consequentemente, há uma carência de ferramentas de análise de qualidade de dados que sejam totalmente agnósticas para dados; isto é, que não assumam conhecimento prévio sobre os dados a serem analisados. Soluções existentes costumam possuir foco em aplicações específicas, tais como sistemas de informação, e são desenhadas para reconhecer e analisar a qualidade de dados em tais contextos como, por exemplo: nomes, endereços, CEP, entre outros. Prototipamos uma plataforma de análise de qualidade de dados verdadeiramente agnóstica para dados, fazendo uso de tecnologias Web e avaliando 6 dimensões de qualidade de dados, como prova de conceito. Concluímos que uma plataforma generalista como a proposta é viável e possui grande potencial de benefício à área de qualidade de dados, uma vez que suas limitações sejam compreendidas e remediadas com o uso de outras ferramentas.

Palavras-chave: dados; qualidade de dados; análise de dados; protótipo.

ABSTRACT

In the Information Age, nothing is more widespread in the digital life than data. Data are elemental, factual attestations of reality, from which it is possible to construct information and generate knowledge. Above all, data is used to aid the decision making process. Thus, to prevent financial loss, it is paramount that the data used are of quality. Throughout the years, various specifications defined what quality data is. However, a lesser explored field deals with the manner through which the measurement of data quality is executed. Consequently, there is a lack of analysis tools for data quality that are data agnostic; that is, that do not assume previous knowledge on the data to be analyzed. Existing solutions usually focus on specific applications, such as information systems, and are designed to recognize and analyze the quality of data in those contexts, such as, for example: names, addresses, postal codes, among others. We have prototyped a data quality analysis platform that is truty data agnostic, making use of Web technologies and evaluating 6 data quality dimensions, as a proof of concept. We conclude that a generalistic platform such as proposed is viable and has a great potential benefit to the data quality space, given that its limitations are understood and remedied with the use of other tools.

Keywords: data, data quality, data analysis, prototype.

LISTA DE ILUSTRAÇÕES

LISTA DE CÓDIGOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
API	Application Programming Interface
CCPA	California Consumer Privacy Act
CSV	Comma-Separated Values
GDPR	General Data Protection Regulation
HTML	HyperText Markup Language
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LGPD	Lei Geral de Proteção de Dados
RFC	Request For Comments
URL	Uniform Resource Locator

SUMÁRIO

1 INTRODUÇÃO

Na segunda metade do século XX, a sociedade gradualmente transicionou para a chamada "Era da Informação". No decorrer de poucas décadas, houve uma mudança cultural global, de um momento em que informações eram exclusivamente registradas em objetos físicos, que precisavam ser consultados pessoalmente; a um segundo momento em que qualquer dado ou informação pode ser transmitida instantaneamente, para qualquer pessoa, ou para todos que desejarem acesso (CASTELLS, 2010).

Dados estão presentes em tudo que fazemos – de interações em redes sociais, a gastos de gasolina por quilômetro dirigido. Em termos organizacionais, é comum que esses dados sejam registrados em tabelas centralizadas, chamadas bancos de dados. Os bancos de dados podem, então, ser lidos e relidos por aplicações, programas e até pessoas para a tomada de decisões informada. A exemplo, o *rover* Curiosity, enviado a Marte, continha o primeiro sensor de radiação a viajar em uma missão desse porte (NASA, 2013). Possuindo dados de níveis de radiação, pode-se determinar a proteção necessária para que a mesma viagem seja realizada por humanos, futuramente.

Evidencia-se, portanto, a crucialidade em manter agregados de tais dados – chamados conjuntos de dados, ou *datasets* – atualizados, corretos e verídicos. Dados imprecisos ou incorretos têm o potencial de causar – e, historicamente, causaram – bilhões de dólares em prejuízo, devido a, não exaustivamente, gastos supérfluos com correio, impressão e mão de obra adicional (ECKERSON, 2002).

Além disso, percebe-se no mercado atual um descaso generalizado pela aferição de qualidade de dados, por parte não somente de empresas, mas também organizações sem fins lucrativos e entidades governamentais (ECKERSON, 2002; SADIQ; INDULSKA, 2017). A falta de compreensão da escala e do impacto negativo que tal descaso incorre no estado da qualidade de dados atual é devastadora.

Sendo assim, esse trabalho se propõe a elaborar uma plataforma de análise de qualidade de dados que contribua para sanar as imperfeições de conjuntos de dados existentes.

1.1 ESTRUTURA DO TRABALHO

No capítulo 2, há uma descrição em maiores detalhes sobre a motivação por trás do desenvolvimento do trabalho. O que são exatamente dados? Por que sua qualidade é tão importante? Além disso, menciona-se as contribuições de outros autores, suas respectivas limitações e opiniões sobre o estado da arte.

Por sua vez, no capítulo 3, detalha-se a proposta real do trabalho, explicitando expectativas do resultado e decisões – por tecnologias ou métricas específicas – realizadas

durante o desenvolvimento. Ademais, nela situam-se outras considerações feitas, incluindo no que tange às licenças de *datasets* usados para testes.

No capítulo 4, expõe-se os resultados do desenvolvimento, com trechos dedicados a cada uma das dimensões de qualidade analisadas. Também considera-se as limitações do projeto resultante, e como outras limitações foram contornadas e transformadas em casos de sucesso.

Finalmente, no capítulo 5, há a conclusão do trabalho, que inclui também expectativas para trabalhos futuros relacionados à área. A seção seguinte contém as referências bibliográficas do trabalho. O Apêndice A (p. **??**) detalha cada dimensão definida pelo padrão escolhido e as justificativas por trás da sua inclusão ou não na plataforma desenvolvida.

2 O PROBLEMA DE QUALIDADE DE DADOS

2.1 O QUE SÃO DADOS?

O dicionário Merriam-Webster define dados como "informação factual, como medições ou estatísticas, usada como base para raciocínio, discussão ou cálculo".[1] Dados são, de maneira geral, formas de registrar algum aspecto da realidade para que este possa ser analisado e utilizado para prever aspectos do futuro, ou entender mais amplamente a situação atual.

Crucialmente, dados devem refletir a realidade em algum ponto no tempo, seja esta tangível ou não. A exemplo, o saldo de uma conta em banco não representa algo tangível, mas sim um valor virtual representativo. Ainda assim, este pode ser considerado um dado, pois indica um aspecto factual da realidae. Em contrapartida, valores numéricos inseridos aleatoriamente em uma planilha não são dados, pois não possuem um paralelo no mundo real.

É pertinente explicitar a distinção entre dados e informação. Dados são meramente atestações factuais e elementares da realidade, enquanto que a informação é uma interpretação ou combinação dessas atestações. Como citado por Rafique et al. (2012), quando dados são contextualizados e combinados para formar uma estrutura, surge a informação. Por exemplo, medições históricas de níveis de gás carbônico são, individualmente, apenas dados; a conclusão de que existe uma ameaça na forma de aquecimento global, por outro lado, é uma informação que pode ser interpretada a partir desses dados e algum contexto adicional.

Os dados em si podem vir de incontáveis origens: de aferições da temperatura ambiente, a coordenadas de aeronaves e, claro, dados de interações de usuários em uma plataforma; para citar alguns. Assim, não é surpresa que dados estão cada vez mais na vanguarda de discussões digitais, normalmente no que tange à privacidade pessoal. A GDPR europeia,[2] a CCPA da Califórnia[3] e a LGPD brasileira (LEI GERAL..., 2018) são exemplos de legislação promulgada com o explícito objetivo de aumentar as proteções à privacidade de indivíduos, e impedir o mau uso de seus dados pessoais.

2.2 O QUE SÃO DADOS DE QUALIDADE?

Sendo um fator de tamanha importância na vida digital, e visto que vivemos em uma realidade cada vez mais imersa em – e indistinguível de – um universo digital, é imprescindível que os dados usados, seja por um pequeno desenvolvedor ou por uma

[1] Disponível em: merriam-webster.com/dictionary/data. Acesso em 12 jan.2022.
[2] Disponível em: data.europa.eu/eli/reg/2016/679/oj. Acesso em: 16 fev.2022.
[3] Disponível em: oag.ca.gov/privacy/ccpa. Acesso em: 16 fev.2022

grande empresa, possuam qualidade. Para garantir isso, diversos padrões e especificações já foram desenvolvidos através dos anos, definindo o que são dados de qualidade, ou como a qualidade de certos dados pode ser maximizada (JESIļEVSKA, 2017).

Dados inexoravelmente precisam refletir algum aspecto da realidade (ainda que um aspecto virtual ou intangível), seja em um certo momento no passado, ou na atualidade. Porém, nem todo dado se mantém correto para sempre; por exemplo, uma porcentagem significativa de dados referentes a indivíduos, como nomes, endereços e estados civis, se torna obsoleta no espaço de tempo de um mês, conforme essas pessoas se casam, se divorciam, se mudam ou morrem (ECKERSON, 2002).

Para garantir a qualidade de diversos dados, portanto, é necessário que eles sejam adaptados ou atualizados periodicamente, assim mantendo sua acurácia para com a realidade. Em seu relatório para The Data Warehousing Institute, Eckerson (2002) reporta que, segundo a estimativa do instituto, dados imprecisos ou incorretos causavam a empresas a perda de US$611 bilhões por ano em 2002, devido a gastos supérfluos com correio, impressão e mão de obra – e especula que essa seja uma estimativa muito abaixo do custo real.

Uma multitude de países e organizações, percebendo o potencial impacto negativo de dados sem qualidade, promulgaram regulamentações através dos anos para prevenir o uso ou publicação de dados imprecisos. Citam-se como exemplos o Data Quality Act (ou Information Quality Act),[4] aprovado no começo dos anos 2000 nos Estados Unidos, garantindo que instituições governamentais estabeleceriam padrões mínimos de qualidade para dados e informações publicadas; e o Welsh Health Circular 025 de 2015,[5] uma portaria do Serviço Nacional de Saúde do País de Gales estabelecendo uma equipe para supervisionar a qualidade de dados e informações em serviços da saúde, chamada Information Quality Improvement Initiative. O Brasil, por sua vez, também tomou iniciativa nessa área com a LGDP (LEI GERAL. . . , 2018), cujo artigo 6º possui o seguinte inciso:

> Art. 6º As atividades de tratamento de dados pessoais deverão observar a boa-fé e os seguintes princípios: [...]
> V - qualidade dos dados: garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de

Dados sem qualidade raramente possuem uso real, pois não representam nada de valor. Por exemplo, salvo à pesquisa histórica e à arquivística, uma lista telefônica do século XX não encontra utilidade nos dias de hoje, pois não mais representa aspectos factuais da realidade. Ademais, no caso de dados pessoais, a falta de qualidade pode agir em detrimento da entidade responsável pela manutenção de tais dados pela miríade de motivos citados anteriormente, somando-se ainda onde aplicável os de reparações legais.

[4] Disponível em: govinfo.gov/content/pkg/PLAW-106publ554/html/PLAW-106publ554.htm. Acesso em: 16 fev.2022.

[5] Disponível em: dhcw.nhs.wales/information-services/information-standards/data-quality/data-quality-docs/whc2015027-e-pdf/. Acesso em: 16 fev.2022.

Em artigo ao Tribunal de Contas do Estado de São Paulo, Xavier (2021) alerta sobre o amplo benefício de dados de qualidade, e cita que uma das características comum entre países em desenvolvimento é a falta de conhecimento atualizado e refinado sobre sua própria realidade; em contraste com países desenvolvidos, que têm o costume de valorizar e privilegiar essa forma de autoconhecimento preciso.

2.3 TRABALHOS RELACIONADOS

Independentemente da forma em que se quantifica a qualidade de dados (ou a falta dela), a aferição em si ainda não possui uma padronização. Isto é, entidades diferentes usam ferramentas diferentes para determinar se seus dados possuem ou não qualidade. Em muitas ocasiões, tais ferramentas de avaliação são desenvolvidas internamente e, portanto, são construídas com um ou poucos bancos de dados em mente, podendo variar bruscamente em seus usos e limitações (ECKERSON, 2002). Além disso, por serem aplicações internas e não públicas, a existência de similares não exclui a utilidade do desenvolvimento de uma de uso geral.

As ferramentas disponíveis no mercado, por sua vez, normalmente possuem foco em dados empresariais e análises de clientes, ao invés de abordagens mais generalizadas. A plataforma da Precisely[6] possui uma suíte de ferramentas de qualidade e análise de dados; porém, essas ferramentas lidam exclusivamente com *big data* e aplicações empresariais, e não se baseiam em um padrão pré-determinado de qualidade de dados. Similarmente, a plataforma Ataccama ONE[7] realiza diversas análises para gerar *insights* sobre a qualidade e a característica dos dados, mas também parece ter sua própria definição interna de qualidade de dados. Ademais, ambas possuem viés para dados de indivíduos e não são agnósticas em suas análises, avaliando somente dados como endereços, datas de nascimento e nomes próprios.

Por outro lado, como reporta Eckerson (2002), em situações (não incomuns) de indiferença com qualidade de dados, pode não existir nenhuma forma de aferição de dados em uma organização. As decisões – ou, mais realisticamente, inações – que levam à carência do uso de uma ferramenta de análise se dá muitas vezes pelo alto custo de ferramentas de terceiros, ou pela falta de uma boa aplicação genérica que não seja tão limitada em seus usos. A exemplo, para a plataforma de dados abertos do governo federal, não há utilidade em uma aplicação que somente verifica dados de indivíduos, pois os dados lá presentes em sua maioria não estão relacionados a pessoas. Esse descaso para com a qualidade de dados pode, então, provocar a produção de informação incorreta, ou a tomada de decisão desinformada, ambos resultados com potenciais repercussões negativas posteriormente.

Sadiq e Indulska (2017) discutem exatamente a qualidade de dados abertos, que são muitas vezes disponibilizados por entidades governamentais. Os autores indicam múlti-

[6] Disponível em: precisely.com/products/verify. Acesso em: 26 fev.2022.
[7] Disponível em: one.ataccama.com. Acesso em: 26 fev.2022.

plos problemas encontrados em seu estudo, como a falta de entendimento da escala e do impacto causados pela má qualidade de dados abertos, e até o próprio desconhecimento de que os dados não possuem qualidade por parte de quem os publica. Ademais, como dados abertos muitas vezes são usados para propósitos diferentes daqueles pelos quais eles foram coletados em primeiro lugar, existem ocasiões em que há qualidade suficiente em um contexto, mas não em outro.

Em termos de métricas de qualidade medidas, nos anos 80, Ballou e Pazer (1985) criaram um modelo complexo de análise de qualidade de dados, fazendo uso de derivadas parciais e multiplicações de matrizes para quantificar a magnitude de erro em um conjunto de dados. Era uma análise totalmente objetiva e rígida, e usava exclusivamente dados numéricos. De maneira similar, Pipino et al. (2005) apontam a falta de objetividade na medição de qualidade de dados em trabalhos anteriores, e desenvolvem fórmulas para atribuir valores numéricos a 5 diferentes dimensões – apesar de concordarem que, em certos contextos, suas soluções não são as ideais, e que outras podem ser usadas.

Também em meados dos anos 80, Laudon (1986) avaliou a qualidade de registros criminais dos Estados Unidos, e usou 3 métricas para tal: completude, acurácia e ambiguidade. Foi um ótimo ponto de partida para a padronização de tais análises, e seu trabalho indicou que entre 50% e 75% dos registros possuíam pelo menos um problema em alguma das métricas verificadas – o que ressalta o valor que análises semelhantes em outros sistemas críticos podem trazer à sociedade.

Pipino, Lee e Wang (2002), por sua vez, percebem a falta de uma padronização para a medição de qualidade de dados, e sugerem 16 dimensões a serem medidas, tanto de valor objetivo, quanto de valor subjetivo. Os autores concluem que não é possível, ou desejável, ter uma solução única, fixa e rígida para toda e qualquer análise de qualidade de dados. Ao invés disso, soluções devem ser adaptadas conforme cada caso de uso. Essa visão é compartilhada por Eckerson (2002, p. 26), que diz: "a tecnologia por si só não pode resolver o problema de qualidade de dados de uma companhia, mas ela tem um papel importante".

Redman (1998) argumenta de um outro ponto de vista: dado o maior esforço e tempo necessário para encontrar e consertar erros em datasets depois que eles já estão prontos, seria potencialmente mais benéfico dedicar os esforços na geração dos dados em si, para garantir sua qualidade. Contudo, nem sempre tem-se controle sobre a fonte de dados usados. Assim, em um futuro utópico, talvez não seja mais necessário analisar a qualidade de dados existentes. Mas, no momento, essa etapa ainda é crucial.

Fica evidente, portanto, o quão benéfico seria para essa área a existência de uma ferramenta generalizada para a análise de qualidade de dados, que não seja dependente a um banco de dados específico, nem que esteja restrita unicamente a dados numéricos ou de indivíduos.

3 DESENVOLVIMENTO DA PLATAFORMA

3.1 EXPECTATIVAS

Temos como objetivo com esse trabalho desenvolver um protótipo de plataforma que realize automaticamente análises sobre a qualidade de certos dados, com a condição de que ela não dependa dos dados usados em si. Em outras palavras, queremos produzir uma prova de conceito para uma plataforma de qualidade de dados *agnóstica* para dados. Realizada essa etapa, poderemos confirmar se uma ferramenta como esta seria útil e benéfica para a área de qualidade de dados, ou não.

Usando essa plataforma, esperamos que, ao inserir um conjunto de dados arbitrário (ainda que no formato esperado pelo programa), a plataforma consiga realizar uma análise o mais precisa possível de problemas do *dataset*, no que tange à qualidade de seu conteúdo. Por exemplo, dado uma lista de medições de temperatura, valores incorretos, como medições abaixo do zero absoluto, devem ser sinalizados para que possam ser consertados.

3.2 DECISÕES

3.2.1 Métricas de qualidade de dados

Como Batini et al. (2009) apontam, o conjunto de métricas utilizadas para a aferição da qualidade de dados não é sempre o mesmo entre análises similares realizadas por outros autores:

> *[...] No general agreement exists either on which set of dimensions defines the quality of data, or on the exact meaning of each dimension.*

Autores através dos anos utilizaram diferentes conjuntos, constituídos de apenas 3 dimensões, como Laudon (1986), até conjuntos com mais de 15, como Pipino, Lee e Wang (2002). Em contrapartida, uma pesquisa por Wang e Strong (1996) enumera um total de 179 dimensões diferentes pensadas por entrevistados para avaliar a qualidade de um *dataset*. Jesiļevska (2017) argumenta que a qualidade de dados é contextual, pois o usuário define o que são dados de qualidade dependendo do contexto de uso, então não é possível existir somente uma resposta fixa para quais métricas devem ser medidas. Dessa maneira, uma das primeiras decisões do trabalho é exatamente essa: em qual conjunto de métricas a plataforma irá se basear para realizar suas análises?

Nesse aspecto, Wang, Storey e Firth (1995) facilitam a decisão, ao realizarem uma meta-análise sobre trabalhos que tratam de qualidade de dados. Os autores defendem o uso de padrões da ISO da seguinte forma:

> *The main strength of the ISO approach is that it is a set of well established standards and guidelines that has been widely adopted by the international community.*

Em dezembro de 2008, a ISO, em parceria com a IEC, publicou o ISO/IEC 25012. O padrão estabelece 15 dimensões mensuráveis (Tabela 1) – algumas inerentes aos dados, e outras dependentes do sistema em si – a serem analisadas para determinar a qualidade dos dados de um dado sistema. A ISO é uma organização de reconhecimento internacional, cujos padrões são utilizados em uma pletora de aplicações de diferentes áreas do conhecimento; assim, nos pareceu adequado escolher em favor das dimensões definidas por esse padrão.

Tendo em mente o uso da ISO/IEC 25012 de 2008, foi inicialmente necessário enumerar e compreender as métricas definidas pela especificação para determinar de que maneira poderiam ser exploradas individualmente na plataforma. Durante essa etapa, percebeu-se que não seria possível analisar automaticamente todas as 15 dimensões abordadas pela especificação, por dois grandes motivos: primeiramente, algumas dimensões dependem inteiramente de subjetividade, tornando sua automatização não trivial; e, além disso, algumas dimensões dependem de análises internas dos sistemas originais que possuem o *dataset*, tornando sua avaliação por uma ferramenta externa como a desenvolvida nesse trabalho impossível. Portanto, dentre as possibilidades, foram escolhidas 6 dimensões que seriam adequadas para uso pela ferramenta.

A lista completa das 15 dimensões especificadas e suas respectivas traduções para o português está presente na Tabela 1, onde destaca-se, em cinza, as 6 dimensões que foram escolhidas para serem analisadas pela plataforma. Uma descrição detalhada de cada dimensão, com justificativas de suas aplicabilidades e restrições se encontra no Apêndice A, na página 37.

3.2.2 Arquitetura lógica

A arquitetura da plataforma divide-se em 3 partes: dados, *back-end* e *front-end*. Os dados são importados de alguma fonte externa, e são estáticos; o *back-end*, então, disponibiliza diversas funcionalidades para o *front-end* através de acessos de API, e faz a consulta dos dados conforme requisitado. O *front-end*, por sua vez, permite ao usuário a interação com a ferramenta, e exibe os resultados da consulta retornados pelo *back-end*.

Abaixo estão descritas em maiores detalhes cada uma das partes:

Dados Os dados devem ser importados e padronizados em somente um formato para que possam ser usados pela plataforma. Por exemplo, se a ferramenta realizar a leitura de arquivos CSV, então todos os dados a serem utilizados devem ser convertidos para CSV previamente.

Tabela 1 – Dimensões de qualidade de dados definidas pelo padrão ISO/IEC

Número	Dimensão original	Dimensão traduzida
1	Accuracy	Acurácia
2	Completeness	Completude
3	Consistency	Consistência
4	Credibility	Credibilidade
5	Currentness	Atualidade
6	Accessibility	Acessibilidade
7	Compliance	Conformidade
8	Confidentiality	Confidencialidade
9	Efficiency	Eficiência
10	Precision	Precisão
11	Traceability	Rastreabilidade
12	Understandability	Compreensibilidade
13	Availability	Disponibilidade
14	Portability	Portabilidade
15	Recoverability	Recuperabilidade

Fonte: (ISO/IEC 25012, 2008).

Back-end Existe a necessidade de alguma interface intermediária para a interação apropriada entre os dados (e metadados) e o usuário em si. Esse papel é cumprido pelo *back-end*, que recebe inquisições do *front-end*, realiza a consulta dos dados, e retorna a resposta para a aplicação.

Front-end Refere-se à parte visual da aplicação, interativa e acessível ao usuário. É nela que o usuário definirá em qual conjunto de dados está interessado e quais análises devem ser realizadas nele. Enfatiza-se que o usuário, através do *front-end*, não possui acesso direto aos dados, mas somente ao *back-end*, que atua como intermediador entre essas duas partes.

É relevante ressaltar que a etapa de preparo dos dados envolve também a criação do arquivo de metadados, descrito em maiores detalhes no Apêndice B (p. 44). Ele contém informações adicionais não disponíveis diretamente no banco, associando campos do banco com dados que podem contribuir para a avaliação de qualidade. Os campos usados foram definidos e adicionados durante o processo de desenvolvimento conforme fazia-se a necessidade para a análise adequada de uma das métricas avaliadas.

3.2.3 Tecnologia

Na aplicação implementada, os dados são guardado em um arquivo SQLite,[1] pela portabilidade e praticidade que oferece. Os metadados, por sua vez, são guardados em

[1] Disponível em: sqlite.org. Acesso em: 12 mar.2022.

um arquivo JSON.[2] Uma descrição detalhada do formato e campos presentes no arquivo implementado está disponível no Apêndice B, na página 44.

O *back-end*, por sua vez, tem como único objetivo simular uma API, isto é, uma interface para a interação do *front-end* com o banco de dados e seus metadados. Portanto, ele foi desenvolvido em Node.js,[3] para tomar proveito da facilidade de integração com o banco SQLite, e da maior familiaridade dos autores com a linguagem.

O *front-end*, finalmente, é o verdadeiro produto a ser testado nesse trabalho; é a plataforma com a qual o usuário irá interagir com os dados, e que lhe dará informações sobre eles. Dessa maneira, ele foi desenvolvido para Web, para que seja acessível a virtualmente qualquer computador ou dispositivo móvel moderno, e utiliza o *framework* Svelte.[4] A decisão em favor do *framework* foi feita, além da familiaridade, pela possibilidade de modularização de componentes que representam cada dimensão medida. Assim, modificações à implementação de uma métrica não afetam as outras. Essa compartimentalização da aplicação é interessante pois permite modificações futuras – como a adição, remoção ou alteração da análise de uma dimensão – sem o risco de perda acidental de funcionalidade, ou necessidade de manutenção adicional.

3.2.4 Considerações adicionais

Durante a pesquisa para a elaboração do trabalho, um dos livros analisados foi Visualize This (YAU, 2011). Apesar de grandes conselhos relacionados a *design* de visualizações, um ponto destoante de possível descuido por parte de Yau foi a não inclusão de ressalvas sobre daltonismo, e como ele pode afetar a compreensão de visualizações que baseiam-se primariamente em cores. Assim, uma das considerações extras feitas para a reportagem da plataforma sobre a qualidade dos dados foi a paleta de cores utilizada, e a adição de funcionalidade que permite trocá-la caso necessário.

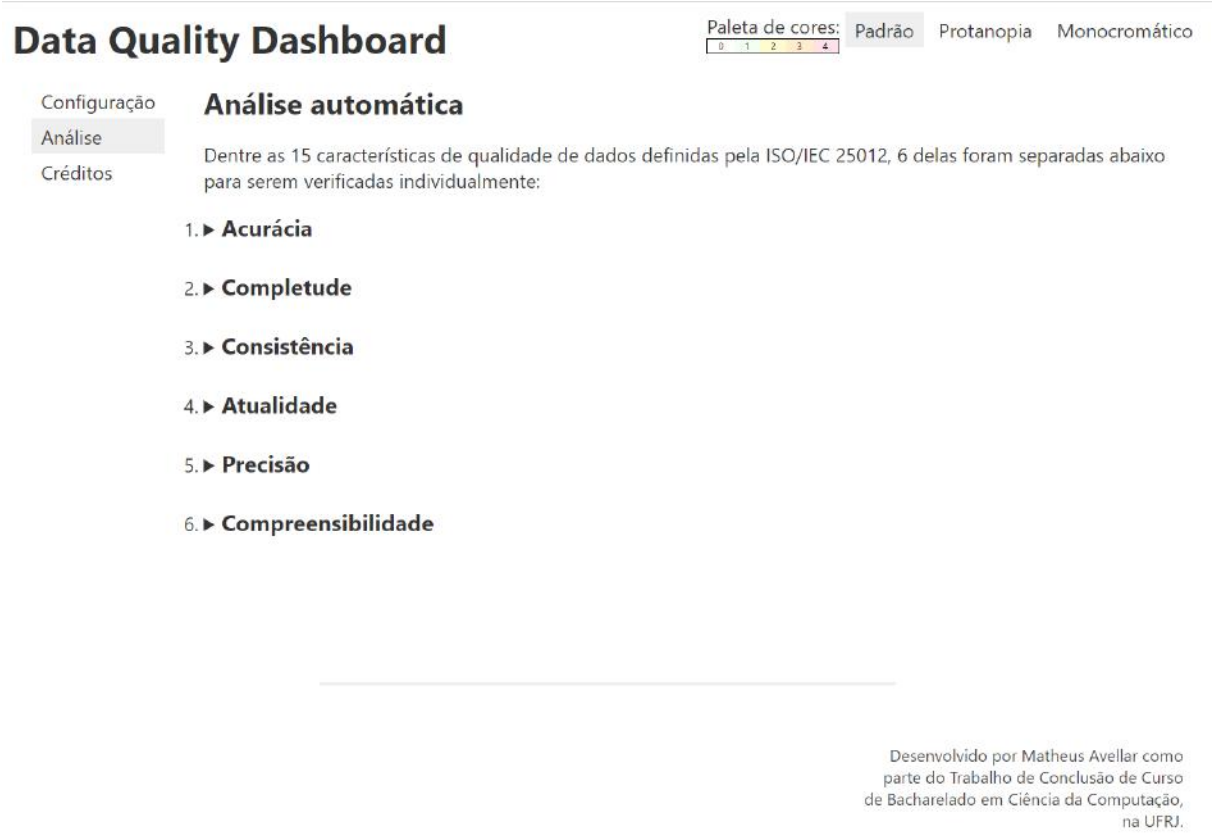
[2] Disponível em: json.org. Acesso em: 12 mar.2022.
[3] Disponível em: nodejs.org. Acesso em: 12 mar.2022.
[4] Disponível em: svelte.dev. Acesso em: 12 mar.2022.

4 RESULTADOS E DISCUSSÃO

4.1 PERSPECTIVA GERAL

A Figura 1 abaixo demonstra parte da tela de análises da plataforma, após a seleção pelo usuário do *dataset* a ser utilizado. No topo esquerdo da imagem, é possível ver a área de navegação da plataforma, com opções para Configuração (onde ocorre a seleção do conjunto de dados), Análise e Créditos. No topo direito, há a seleção de paleta de cores, para permitir o uso por daltônicos, pessoas com baixa visão, ou em monitores com pouca fidelidade. A Figura 2 possui três situações em que cada uma das opções de paletas de cores está selecionada, para melhor demonstrar a diferença entre elas.

Figura 1 – Tela de análises da plataforma

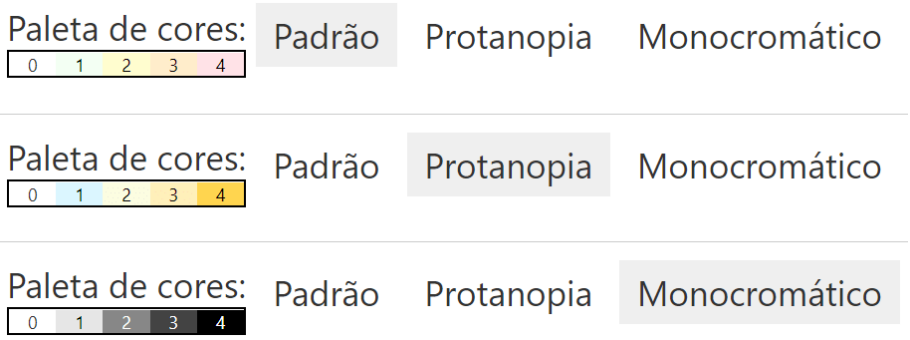


4.2 SUCESSOS

4.2.1 Dados de teste

Para realizar a população do banco de dados de teste da plataforma, utilizou-se dados abertos disponibilizados pela Prefeitura do Rio de Janeiro, no portal de dados abertos

Figura 2 – Seleção de paleta de cores da plataforma



Data.Rio, e pelo Governo Federal do Brasil, no Portal Brasileiro de Dados Abertos (Dados.gov.br). Ambos disponibilizaram os dados sob a licença CC-BY,[1] que permite seu uso para quaisquer objetivos, desde que com atribuição – algo que também foi incluído na plataforma, sob a aba de créditos.

Os conjuntos de dados usados separam-se em dois:

Monitoramento horário da qualidade do ar Parte do projeto municipal MonitorAr, consiste em dados coletados uma vez a cada hora por diversas estações de qualidade do ar localizadas na cidade do Rio de Janeiro. O conjunto de dados final é a junção de dois *datasets* similares, referentes a estações fixas[2] e a estações móveis.[3] Ambos são disponibilizados no formato CSV (RFC 4180).[4]

Monitoramento mensal da produção de pré-sal Disponibilizados pela ANP, consiste em um agregado mensal de dados sobre a produção e qualidade do pré-sal. O conjunto de dados usado refere-se aos dados publicados entre março e dezembro de 2021.[5] Também estão disponibilizados no formato CSV.

4.2.2 Acurácia

Um caso de sucesso é a análise de acurácia do *dataset* de monitoramento de qualidade do ar, disponibilizado através do Data.Rio. Ele possui medições de diversos aspectos da qualidade do ar no Rio de Janeiro (como concentração de gases estufa e de material particulado), e inclui até fatores de cunho mais analítico, como temperatura e umidade relativa do ar.

[1] Disponível em: creativecommons.org/licenses/by/4.0. Acesso em: 9 fev.2022.
[2] Disponível em: data.rio/datasets/PCRJ::dados-horários-do-monitoramento-da-qualidade-do-ar-monitorar. Acesso em: 12 mar.2022.
[3] Disponível em: data.rio/datasets/PCRJ::dados-horários-das-campanhas-móveis-de-monitoramento-da-qualidade-do-ar-monitorar. Acesso em: 12 mar.2022.
[4] Disponível em: ietf.org/rfc/rfc4180. Acesso em: 12 mar.2022.
[5] Disponível em: dados.gov.br/dataset/producao-de-petroleo-e-gas-natural-por-poco. Acesso em: 12 mar.2022.

Usando esse conjunto de dados como exemplo, a análise da plataforma encontra 106 medições de temperatura ambiente no Rio de Janeiro (de um total de 716,563 medições) que registraram 0 °C – valores incorretos e fora do intervalo esperado. O intervalo esperado, de 8 a 50 °C, é definido no arquivo de metadados, externamente ao banco de dados em si.

Algo similar ocorre com a coluna de pressão do ar, que possui 90 medições de 800 mbar, valor fora do intervalo esperado de 870 a 1080 mbar. Ambos os casos podem ser especulativamente explicados por falhas físicas, ou configuração errônea do sistema de coleta de dados.

A Figura 3 possui imagens da plataforma reportando ambos os resultados citados, e a indicação de que nenhuma anomalia adicional foi encontrada em outras medições.

4.2.3 Completude

A métrica de completude, na plataforma, é analisada em termos de valores não nulos – isto é, para cada entrada da tabela analisada, se existe um valor para determinada coluna, então a entrada é dita "completa"; se não existe valor, ou seja, se o valor presente é nulo, então a entrada não é considerada "completa".

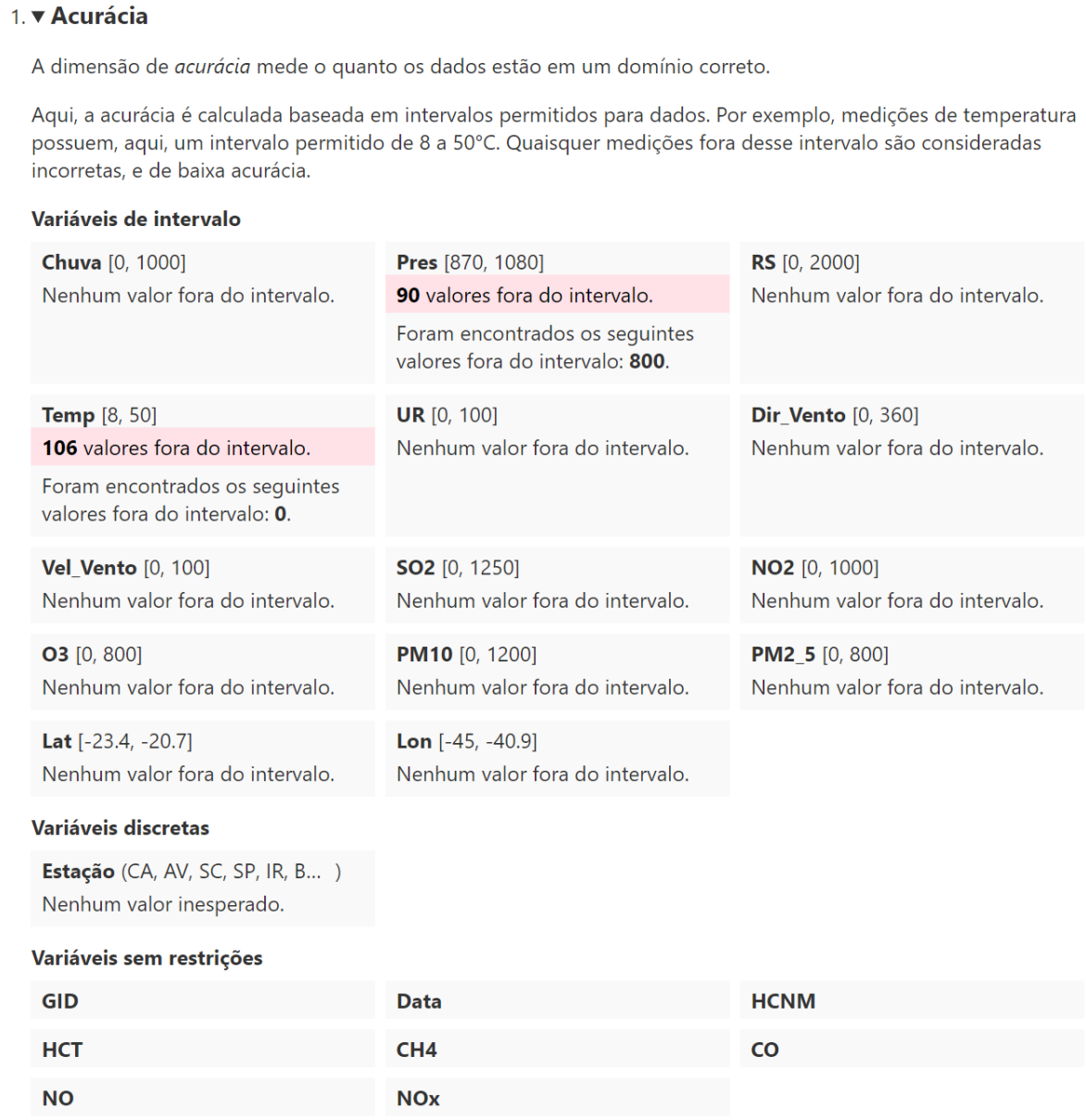
Nem sempre a presença de valores nulos representa problemas no *dataset*, mas em muitas situações pode indicar alguma falha na medição ou escrita de dados para o banco. Dessa maneira, uma indicação da presença ou não de tais valores é significativa para administradores de sistemas de bancos de dados, pois pode contribuir para a descoberta de problemas até então ignorados.

A Figura 4 mostra a análise de completude dos dados de monitoramento da qualidade do ar. A atenção do usuário é mais enfaticamente chamada, com cores gradualmente mais acentuadas, às métricas que possuem maior percentual de valores nulos. O pior caso presente nesse conjunto de dados é o de material particulado $< 2.5\mu m$, do qual mais de 80% das entradas possuem valores nulos; um exemplo presente no outro lado do espectro é o campo de latitude, que não possui um único valor nulo em todas as suas mais de 700 mil entradas.

4.2.4 Consistência

Outra análise bem sucedida é a medição de consistência dos *datasets* testados, que também funciona com ajuda dos metadados. Restrições de variáveis – como, no caso das medições de qualidade de ar, o campo "NOx" alegadamente sendo igual ao somatório dos campos de "NO" e "NO2" – são especificadas no arquivo de metadados, e podem em seguida ser analisadas automaticamente. Como pode ser visto na Figura 5, a restrição não está sendo respeitada em quase 200 mil medições (novamente, de um total de pouco mais de 700 mil).

Figura 3 – Parte da análise da dimensão de acurácia sobre dados de monitoramento de ar



1. ▾ **Acurácia**

A dimensão de *acurácia* mede o quanto os dados estão em um domínio correto.

Aqui, a acurácia é calculada baseada em intervalos permitidos para dados. Por exemplo, medições de temperatura possuem, aqui, um intervalo permitido de 8 a 50°C. Quaisquer medições fora desse intervalo são consideradas incorretas, e de baixa acurácia.

Variáveis de intervalo

Chuva [0, 1000] Nenhum valor fora do intervalo.	**Pres** [870, 1080] **90** valores fora do intervalo. Foram encontrados os seguintes valores fora do intervalo: **800**.	**RS** [0, 2000] Nenhum valor fora do intervalo.
Temp [8, 50] **106** valores fora do intervalo. Foram encontrados os seguintes valores fora do intervalo: **0**.	**UR** [0, 100] Nenhum valor fora do intervalo.	**Dir_Vento** [0, 360] Nenhum valor fora do intervalo.
Vel_Vento [0, 100] Nenhum valor fora do intervalo.	**SO2** [0, 1250] Nenhum valor fora do intervalo.	**NO2** [0, 1000] Nenhum valor fora do intervalo.
O3 [0, 800] Nenhum valor fora do intervalo.	**PM10** [0, 1200] Nenhum valor fora do intervalo.	**PM2_5** [0, 800] Nenhum valor fora do intervalo.
Lat [-23.4, -20.7] Nenhum valor fora do intervalo.	**Lon** [-45, -40.9] Nenhum valor fora do intervalo.	

Variáveis discretas

Estação (CA, AV, SC, SP, IR, B...) Nenhum valor inesperado.

Variáveis sem restrições

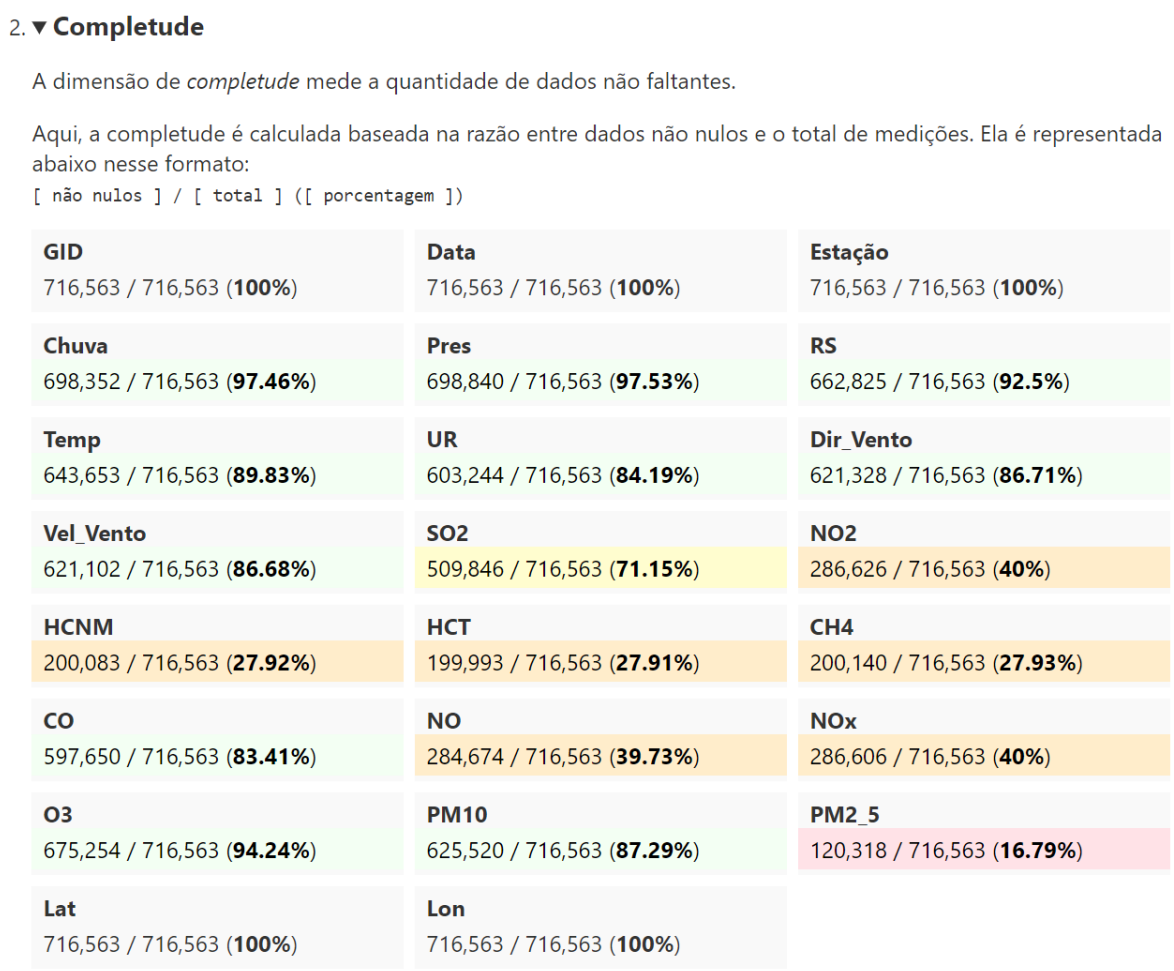
GID	**Data**	**HCNM**
HCT	**CH4**	**CO**
NO	**NOx**	

Adicionalmente, ao realizar a análise nos dados de medições do pré-sal, pode-se concluir que este é um *dataset* mais consistente do que o de monitoramento da qualidade do ar. Os resultados da análise estão presentes na Figura 6, onde nenhuma medição desrespeita as restrições estabelecidas.

4.2.5 Atualidade

A análise da dimensão de atualidade também se demonstrou bem sucedida nos testes realizados. A avaliação da plataforma verifica os intervalos entre as datas de cada entrada sucessiva do banco de dados. Dessa maneira, é possível perceber a existência de atrasos

Figura 4 – Análise da dimensão de completude sobre dados de monitoramento de ar



2. ▾ **Completude**

A dimensão de *completude* mede a quantidade de dados não faltantes.

Aqui, a completude é calculada baseada na razão entre dados não nulos e o total de medições. Ela é representada abaixo nesse formato:

`[ não nulos ] / [ total ] ([ porcentagem ])`

GID 716,563 / 716,563 (**100%**)	**Data** 716,563 / 716,563 (**100%**)	**Estação** 716,563 / 716,563 (**100%**)
Chuva 698,352 / 716,563 (**97.46%**)	**Pres** 698,840 / 716,563 (**97.53%**)	**RS** 662,825 / 716,563 (**92.5%**)
Temp 643,653 / 716,563 (**89.83%**)	**UR** 603,244 / 716,563 (**84.19%**)	**Dir_Vento** 621,328 / 716,563 (**86.71%**)
Vel_Vento 621,102 / 716,563 (**86.68%**)	**SO2** 509,846 / 716,563 (**71.15%**)	**NO2** 286,626 / 716,563 (**40%**)
HCNM 200,083 / 716,563 (**27.92%**)	**HCT** 199,993 / 716,563 (**27.91%**)	**CH4** 200,140 / 716,563 (**27.93%**)
CO 597,650 / 716,563 (**83.41%**)	**NO** 284,674 / 716,563 (**39.73%**)	**NOx** 286,606 / 716,563 (**40%**)
O3 675,254 / 716,563 (**94.24%**)	**PM10** 625,520 / 716,563 (**87.29%**)	**PM2_5** 120,318 / 716,563 (**16.79%**)
Lat 716,563 / 716,563 (**100%**)	**Lon** 716,563 / 716,563 (**100%**)	

ou irregularidades dentre as medições realizadas.

Um caso bastante interessante é o dos dados mensais do pré-sal (de aproximadamente 3200 medições), o qual a análise reporta medições no intervalo de 28 a 35 dias, com média de aproximadamente 30 dias e 10 horas (Figura 7). Para um *dataset* que se dispõe a ser mensal, essa média está com precisão bastante próxima do esperado.

O conjunto de dados horários de monitoramento da qualidade do ar possui variação inexistente, em que todas as medições possuem diferença de tempo de 1 hora entre si. Assim, a avaliação da plataforma é menos intrigante, reportando média de exata 1 hora (Figura 8). De qualquer forma, é um caso de sucesso.

4.2.6 Precisão

Como não seria trivial prever programaticamente a necessidade, ou falta dela, de uma maior precisão para dados numéricos, a análise da plataforma restringe-se a somente informar a quantidade de casas decimais de dados guardados como ponto flutuante. A dimensão de precisão pode englobar dados para além dos números fracionários – como, por

Figura 5 – Análise da dimensão de consistência sobre dados de monitoramento de ar

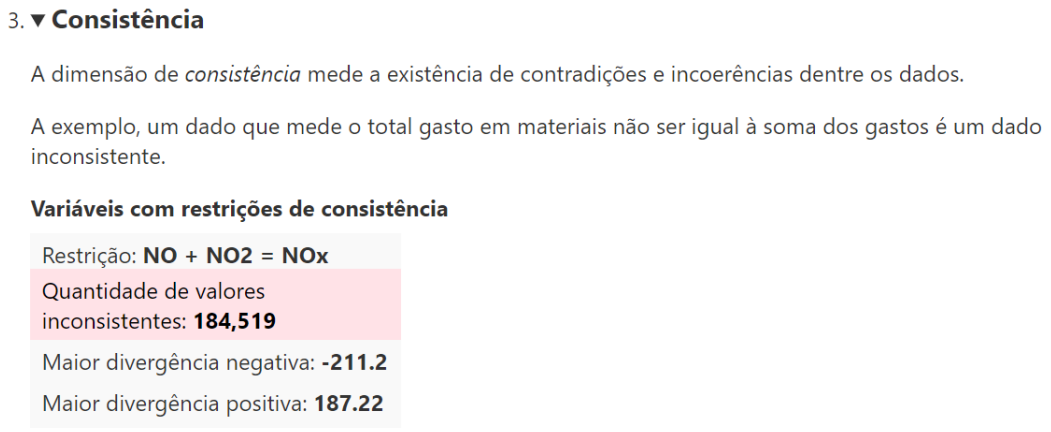


Figura 6 – Análise da dimensão de consistência sobre dados de pré-sal

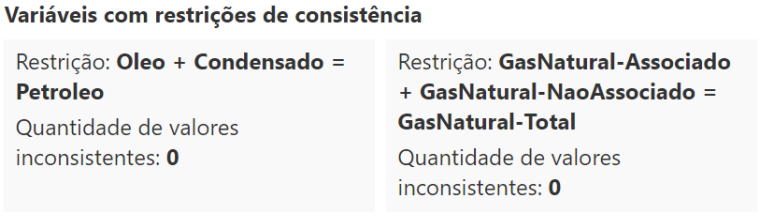
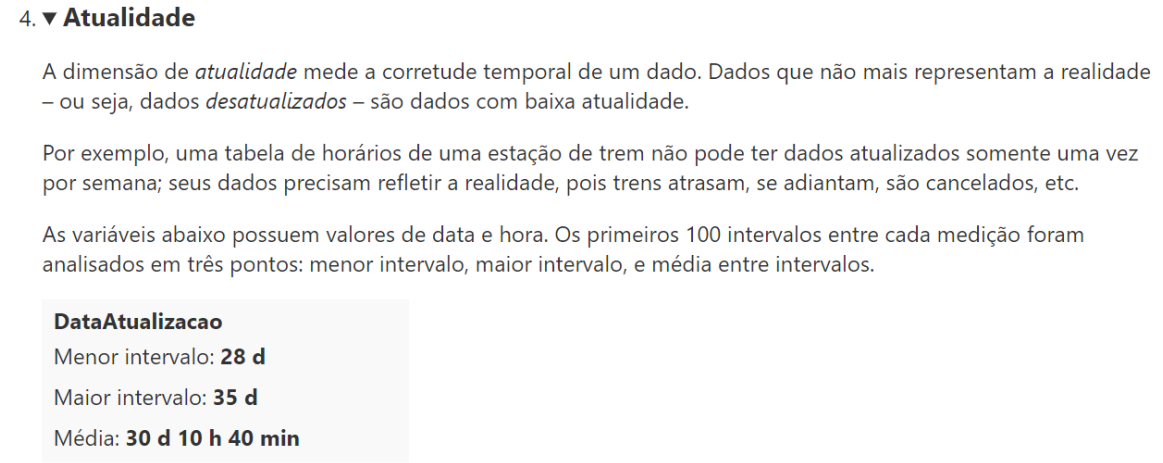


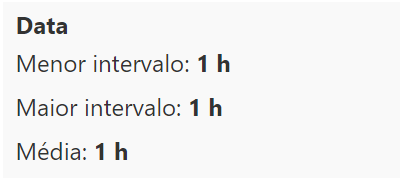
Figura 7 – Análise da dimensão de atualidade sobre dados de pré-sal



exemplo, a precisão de endereços ou nomes próprios. Contudo, a análise de tal maneira não seria trivial, e essa limitação será discutida em uma seção posterior. Porém, dado o escopo selecionado para a análise, sua execução pode ser considerada um caso de sucesso, pois cumpre o que propõe fazer.

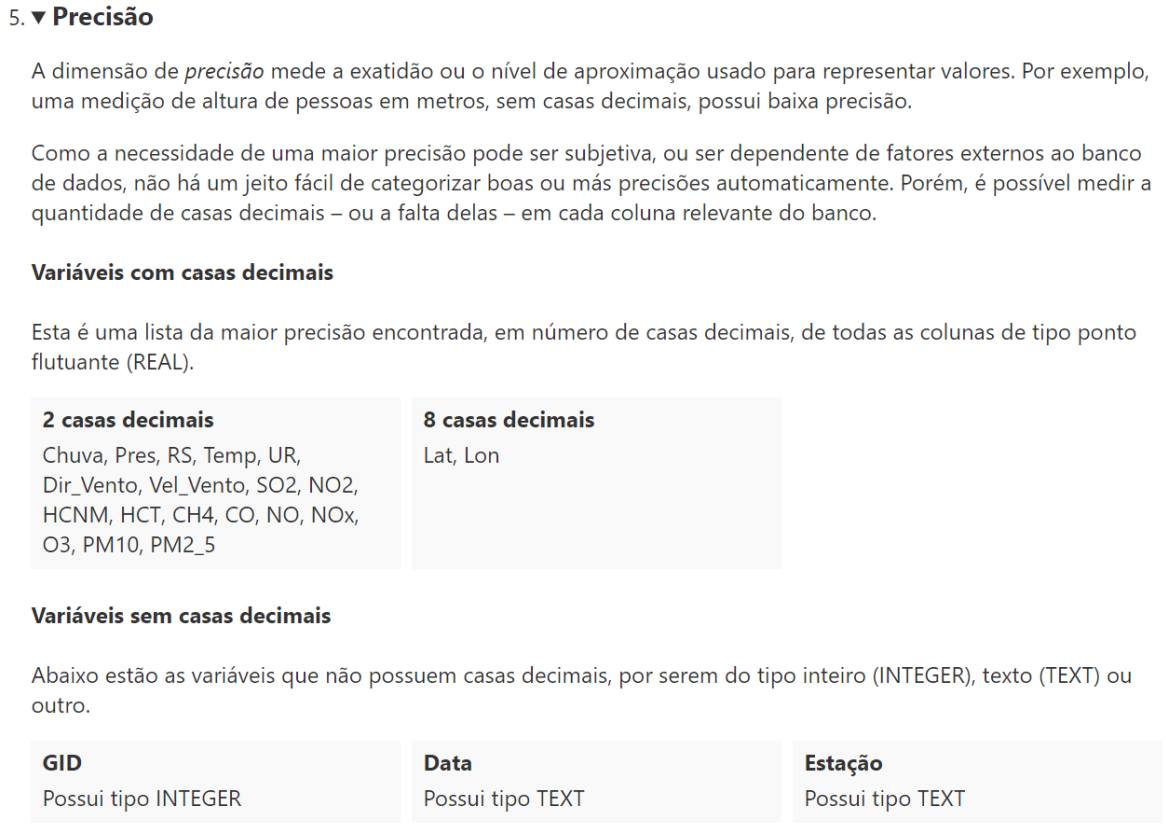
O conjunto de dados de monitoramento do ar é mais uniforme no que tange à precisão de seus valores; todos, com exceção dos campos de latitude e longitude, possuem valores

Figura 8 – Análise da dimensão de atualidade sobre dados de qualidade do ar



com precisão de até 2 casas decimais (Figura 9). Por sua vez, o *dataset* de pré-sal possui uma precisão mais distribuída, com número de casas decimais variando entre somente 1 em alguns campos (como grau API do óleo), a até 5 em outros (como percentual de metano); vide Figura 10.

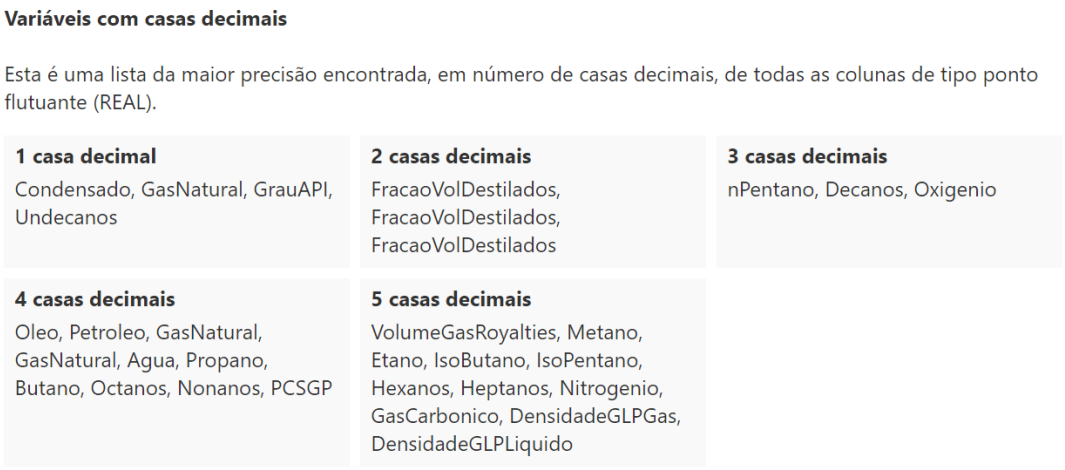
Figura 9 – Análise da dimensão de precisão sobre dados de monitoramento do ar



4.2.7 Compreensibilidade

Similarmente à precisão, a compreensibilidade é uma dimensão de fator subjetivo. Não é possível determinar programaticamente se um dado é compreensível ou não por um usuário humano. E, novamente, limitações sobre a análise serão explicitadas e discutidas em uma seção posterior.

Figura 10 – Análise da dimensão de precisão sobre dados de pré-sal



Dito isso, a avaliação realizada pela plataforma se refere a informações provenientes dos metadados. Especificamente, a presença ou ausência de unidade de medida e descrição referente ao campo verificado. Na Figura 11, os dados de monitoramento do ar são analisados e aqueles que não possuem unidade de medida ou descrições associadas são enfatizados.

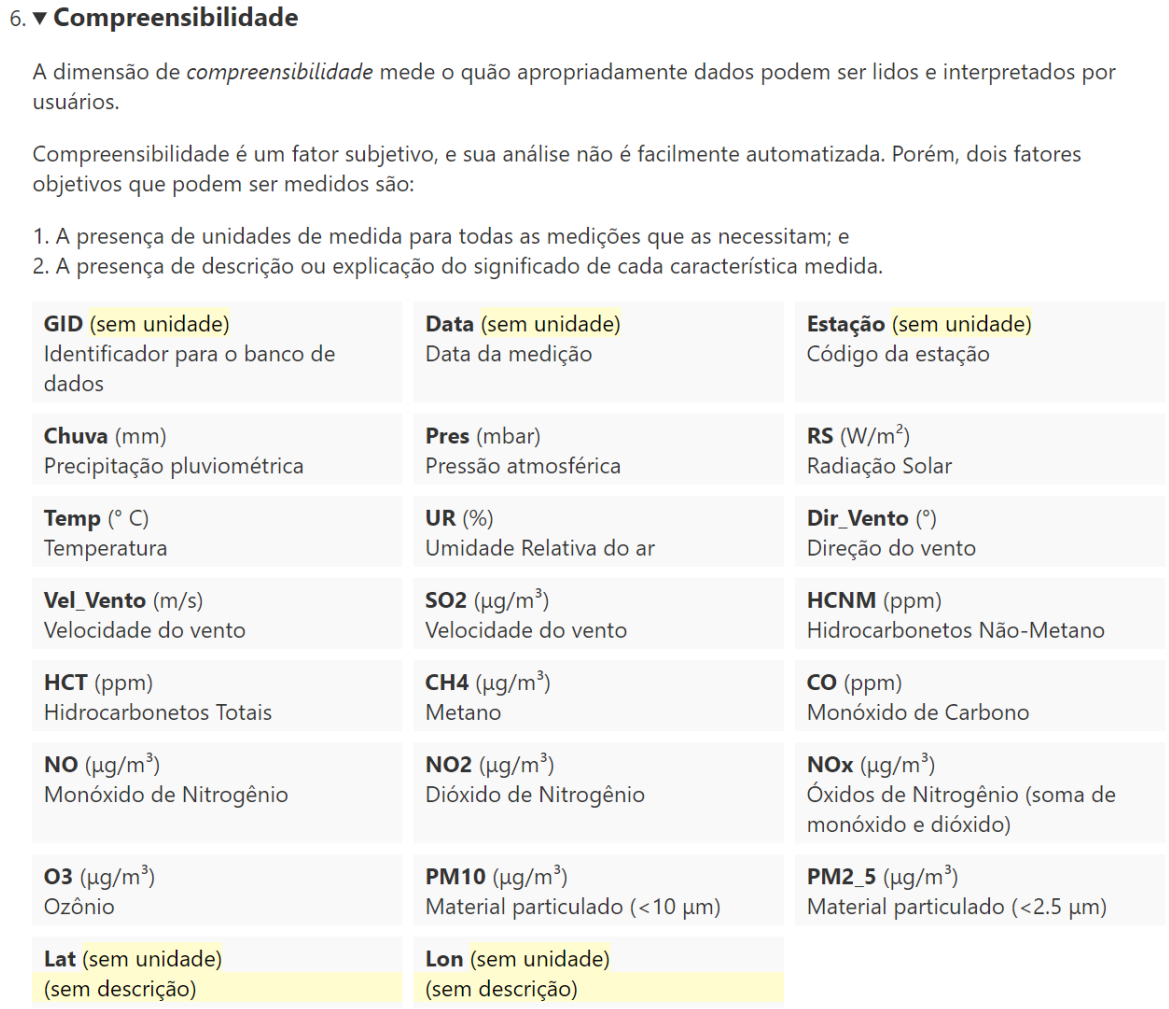
4.3 LIMITAÇÕES

A plataforma, contudo, não é perfeita, como esperado. Dentre as limitações, uma das mais claras é a necessidade de uma documentação de metadados externa. Não é possível para uma plataforma agnóstica de dados ter conhecimento inerente sobre as limitações necessárias para valores arbitrários. Em termos gerais, a problemática reside no desconexo entre a interpretação algorítmica de dados – como meros valores, numéricos ou não – e a interpretação humana de dados – como representando fatos ou atributos, limitados de um jeito ou de outro na vida real. Para sanar o problema, pelo menos em parte, uma ajuda "externa", em forma de um arquivo *machine-readable* contendo esses metadados, é imprescindível; caso contrário, a análise automática seria ainda mais limitada.

Para além disso, há também o caso de detecção de problemas em dados textuais. Textos podem conter informações arbitrárias – como nomes, identificadores, descrições, URLs, HTML, Tweets, para listar alguns exemplos – e estas não têm como ser facilmente distinguidas de forma agnóstica. Para um computador, não é trivial diferenciar entre uma entrada textual válida e uma inválida, sem uma limitação extrema de escopo, enumerando toda entrada possível.

Quando se trata de identificadores (como nomes de países) é trivial restringir o escopo de valores plausíveis a uma lista bem definida (como a lista de todos os nomes de países existentes), e tratar quaisquer valores não presentes como incorretos. Contudo, nem toda medição é bem comportada dessa maneira. A ISO/IEC 25012 (2008) cita como contra-

Figura 11 – Análise da dimensão de compreensibilidade sobre dados de monitoramento do ar



6. ▼ **Compreensibilidade**

A dimensão de *compreensibilidade* mede o quão apropriadamente dados podem ser lidos e interpretados por usuários.

Compreensibilidade é um fator subjetivo, e sua análise não é facilmente automatizada. Porém, dois fatores objetivos que podem ser medidos são:

1. A presença de unidades de medida para todas as medições que as necessitam; e
2. A presença de descrição ou explicação do significado de cada característica medida.

GID (sem unidade) Identificador para o banco de dados	**Data** (sem unidade) Data da medição	**Estação** (sem unidade) Código da estação
Chuva (mm) Precipitação pluviométrica	**Pres** (mbar) Pressão atmosférica	**RS** (W/m^2) Radiação Solar
Temp (° C) Temperatura	**UR** (%) Umidade Relativa do ar	**Dir_Vento** (°) Direção do vento
Vel_Vento (m/s) Velocidade do vento	**SO2** (µg/m^3) Velocidade do vento	**HCNM** (ppm) Hidrocarbonetos Não-Metano
HCT (ppm) Hidrocarbonetos Totais	**CH4** (µg/m^3) Metano	**CO** (ppm) Monóxido de Carbono
NO (µg/m^3) Monóxido de Nitrogênio	**NO2** (µg/m^3) Dióxido de Nitrogênio	**NOx** (µg/m^3) Óxidos de Nitrogênio (soma de monóxido e dióxido)
O3 (µg/m^3) Ozônio	**PM10** (µg/m^3) Material particulado (<10 µm)	**PM2_5** (µg/m^3) Material particulado (<2.5 µm)
Lat (sem unidade) (sem descrição)	**Lon** (sem unidade) (sem descrição)	

exemplo de uma boa acurácia sintática a inclusão acidental de um nome de indivíduo, "Mary", como "Marj". A detecção de tal erro não é trivial, pois "Marj" também pode ser um nome válido; isto é, não é plausível criar uma "lista de nomes aceitáveis" sem excluir nomes possíveis. Portanto, a plataforma não tem maneira de validar a acurácia sintática do que pode ser chamado de "entradas textuais irrestritas".

De modo geral, todas as limitações se dão devido ao fato de que uma grande parte da determinação da qualidade de um dado são subjetivas, ou dependem de fatores externos ao banco de dados. A requisição por um arquivo de metadados tornou possível a análise de mais dimensões do que seriam possíveis dado somente um *dataset* descontextualizado. Ainda assim, uma grande parte das dimensões (9 de 15, ou 60%) não puderam ser automaticamente analisadas por motivos descritos em mais detalhes no Apêndice A (p. **??**), mas que em geral se resumem à falta de contextualização dos dados e subjetividade das interpretações por parte do algoritmo.

5 CONCLUSÃO

Concluímos que a análise frequente da qualidade de dados é imprescindível, em especial para sistemas complexos e cruciais – como bancos de dados governamentais e de grandes empresas. Adicionalmente, em muitos casos, não é viável que tal aferição seja realizada manualmente. Dessa maneira, utilizamos esse trabalho para avaliar se seria possível ou eficaz automatizar o processo de análise de qualidade de dados.

Apesar do escopo limitado, dado que o projeto somente possuía a proposta de prova de conceito e prototipagem, confirma-se o potencial da automatização, mesmo que parcial, do processo de aferição de qualidade dados. Uma ferramenta desenvolvida para esse nicho seria amplamente benéfica para a área.

Assim como Pipino, Lee e Wang (2002) e Eckerson (2002), concluímos que não é possível nem desejável ter uma solução única e rígida para a análise automatizada de qualidade de dados – em especial dado o requisito de ser agnóstica a dados. Muitas das dimensões que se deseja medir são subjetivas – como compreensibilidade que, por definição, depende da compreensão (ou falta dela) por parte do usuário.

A aplicação de uma ferramenta de análise, ao invés disso, seria plausível em sistemas que possuam outras formas de averiguação que cubram as dimensões de caráter subjetivo. Sua funcionalidade poderia ser similar ao uso de testes unitários, de escopo limitado e executados automaticamente a cada atualização do banco de dados. De qualquer maneira, a aplicação desenvolvida serve apropriadamente seu propósito esperado de prova de conceito e prototipagem, e satisfaz as expectativas definidas para o trabalho.

5.1 TRABALHOS FUTUROS

O código desenvolvido durante o projeto está disponível,[1] e poderá ser utilizado futuramente como base para outros trabalhos, seja diretamente via reutilização do código disponibilizado, ou como exemplo de decisões corretas e incorretas durante o desenvolvimento de uma nova aplicação semelhante.

Algumas limitações, como a impossibilidade de algumas averiguações de dimensões, se dão pela externalidade da plataforma dos sistemas que analisa; no caso de empresas ou organizações implementarem aplicações similares em suas infraestruturas, seria possível avaliar ainda mais dimensões de qualidade, como disponibilidade, portabilidade e recuperabilidade. Outras limitações, como a subjetividade de certas dimensões, poderiam concebivelmente ser resolvidas com uso de tecnologias de *machine learning* ou inteligência artificial.

[1] Disponível em: s.avl.la/tcc.

Finalmente, há possibilidade de melhorias para a plataforma. A restrição de intervalos possíveis para cada métrica usada no projeto é uma distribuição probabilística uniforme – isto é, um valor dentro do intervalo definido no arquivo de metadados é considerado tão possível quanto qualquer outro, enquanto que um valor fora do intervalo é inaceitável. A implementação de diferentes funções de distribuição – como a distribuição Normal (também conhecida como "de Gauss") ou a distribuição Poisson – poderia ser uma adição interessante para a plataforma. A menor *probabilidade* de um valor poderia ser sinalizada, criando assim maior nuance quando comparado à representação binária atual.

Como visto, devido à falta de aplicações similares no mercado, há estímulo societal e possivelmente também financeiro para a elaboração de novas ferramentas que cubram mais apropriadamente essas carências.

REFERÊNCIAS

BALLOU, D. P.; PAZER, H. L. Modeling data and process quality in multi-input, multi-output information systems. **Management Science**, v. 31, n. 2, p. 150–162, 1985. Disponível em: https://doi.org/10.1287/mnsc.31.2.150.

BATINI, C. et al. Methodologies for data quality assessment and improvement. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 41, n. 3, jul 2009. ISSN 0360-0300. Disponível em: https://doi.org/10.1145/1541880.1541883.

CASTELLS, M. **The Information Age: Economy, Society and Culture**. 2. ed. [S.l.]: Wiley-Blackwell, 2010. v. 1. ISBN 978-0-631-21594-3.

ECKERSON, W. W. **Data Quality and the Bottom Line: Achieving Business Success Through a Commitment to High Quality Data**. [S.l.]: The Data Warehousing Institute, 2002.

ISO/IEC 25012. 1. ed. Geneva, Switzerland, 2008. Disponível em: https://www.iso.org/standard/35736.html.

JESIļEVSKA, S. Data quality dimensions to ensure optimal data quality. **Romanian Economic Journal**, v. 20, n. 63, p. 89–103, mar 2017. ISSN 2286-2056. Disponível em: http://www.rejournal.eu/article/data-quality-dimensions-ensure-optimal-data-quality.

LAUDON, K. C. Data quality and due process in large interorganizational record systems. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 29, n. 1, p. 4–11, jan 1986. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/5465.5466.

LEI GERAL de Proteção de Dados Pessoais. 2018. Lei Nº 13.709, de 14 de agosto de 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 16 fev.2022.

NASA. Data from nasa rover's voyage to mars aids planning. 2013. Disponível em: https://mars.nasa.gov/news/1478/data-from-nasa-rovers-voyage-to-mars-aids-planning/. Acesso em: 26 fev.2022.

PIPINO, L. et al. Developing measurement scales for data-quality dimensions. **Information Quality: Advances in Management Information Systems**, M.E. Sharpe, Armonk, NY, USA, v. 1, p. 37–50, 2005.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 45, n. 4, p. 211–218, abr 2002. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/505248.506010.

RAFIQUE, I. et al. **Information Quality Evaluation Framework: Extending ISO 25012 Data Quality Model**. Zenodo, 2012. Disponível em: https://doi.org/10.5281/zenodo.1072956.

REDMAN, T. C. The impact of poor data quality on the typical enterprise. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 41, n. 2, p. 79–82, fev 1998. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/269012.269025.

SADIQ, S.; INDULSKA, M. Open data: Quality over quantity. **International Journal of Information Management**, v. 37, n. 3, p. 150–154, 2017. ISSN 0268-4012. Disponível em: https://www.sciencedirect.com/science/article/pii/S0268401216309021.

WANG, R.; STOREY, V.; FIRTH, C. A framework for analysis of data quality research. **IEEE Transactions on Knowledge and Data Engineering**, v. 7, n. 4, p. 623–640, 1995. Disponível em: https://doi.org/10.1109/69.404034.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. **J. Manage. Inf. Syst.**, M. E. Sharpe, Inc., USA, v. 12, n. 4, p. 5–33, mar 1996. ISSN 0742-1222. Disponível em: https://doi.org/10.1080/07421222.1996.11518099.

XAVIER, F. C. **Qualidade de Dados: direito do cidadão e dever do Estado**. Tribunal de Contas do Estado de São Paulo, 2021. Disponível em: https://www.tce.sp.gov.br/publicacoes/artigo-qualidade-dados-direito-cidadao-e-dever-estado. Acesso em: 16 fev.2022.

YAU, N. **Visualize This: The FlowingData Guide to Design, Visualization, and Statistics**. [S.l.]: John Wiley & Sons, 2011. ISBN 978-1-118-14026-0.

APÊNDICES

APÊNDICE A – DESCRIÇÃO E JUSTIFICATIVA DO USO DE CADA DIMENSÃO DEFINIDA PELA ISO/IEC.

Este apêndice contém as 15 dimensões originais especificadas pelo ISO/IEC 25012 (2008), e uma descrição para cada uma justificando sua inclusão (ou não) como métrica para análise automática da plataforma.

A.1 DIMENSÕES ANALISADAS PELA PLATAFORMA

A.1.1 Acurácia (Accuracy)

A dimensão de acurácia mede a adequação de um dado a um domínio considerado correto. Ela é subdividida pela especificação em duas partes: acurácia sintática e acurácia semântica.

A acurácia sintática lida com erros no domínio sintático; isto é, o sentido (ou semântica) atribuído àquele valor não é diferente, mas seu registro foi incorreto. Por exemplo, a troca de um valor de idade de "23" para "230" acidentalmente é um erro sintático. Uma idade de "230" não é sintaticamente válida nesse contexto por estar fora de um intervalo esperado (e.g. 0–120).

A acurácia semântica, por outro lado, lida com problemas no sentido atribuído ao valor. Por exemplo, se, ao invés de "23", registramos "15", ainda estamos no domínio sintático ("15" está no intervalo anterior de 0–120), mas não estamos no domínio semântico ("15" não é equivalente a "23", e poderá ter repercussões por e.g. não ser acima da maioridade legal).

Na plataforma, a análise realizada é a de acurácia sintática; um arquivo externo de metadados é utilizado para determinar os intervalos esperados de cada dado medido. Dessa maneira, pode-se determinar se, por exemplo, uma medição de direção do vento está incorreta verificando sua presença no intervalo 0–360°. A análise de acurácia semântica é de cunho mais subjetivo e contextual – como instruir a um algoritmo que diga se "15" é uma idade semanticamente correta para o indivíduo, sem maiores informações? Assim, ela não é avaliada.

A.1.2 Completude (Completeness)

A dimensão de completude é uma das mais simples de se automatizar e compreender. Em suma, para um dataset ser considerado completo, ele deve possuir valores para todos os campos medidos. Por exemplo, um registro de clientes com campos de nome e endereço, se alguma das entradas não possuir ou nome, ou endereço, então o dataset não é completo; sua completude não é 100%.

A.1.3 Consistência (Consistency)

A dimensão de consistência mede a coerência e não contradição entre os dados de um conjunto. Como exemplificado pela especificação, uma entrada de banco de dados de um empregado cuja data de nascimento é depois da data de contratação possui uma má consistência.

Para a plataforma, é possível analisar a consistência de um conjunto de dados fazendo uso dos metadados externos. Neles, há cláusulas de restrição que indicam qual regra deve ser respeitada. Por exemplo, no dataset de monitoramento de qualidade de ar, há a seguinte restrição para as colunas "NO", "NO2" e "NOx": "NOx = NO + NO2". Dessa forma, valores que não respeitam a restrição são marcados como inconsistentes, e reportados ao usuário na plataforma.

A.1.4 Atualidade (Currentness)

A dimensão de atualidade lida com a idade de dados, e como isso afeta sua veracidade. Alguns dados – como, por exemplo, a profundidade de fossas submarinas – não se modificam frequentemente, e portanto provavelmente raramente ficarão desatualizados. Contudo, muitos outros necessitam de atualizações periódicas para não se tornarem inverdades ou perderem seu valor; cita-se alguns exemplos, em ordem crescente de frequência de desatualização: dados de conversão entre duas moedas, de temperatura, de endereços pessoais, de nomes de indivíduos, ou de nomes de ruas.

Para a testagem da plataforma, utilizou-se dois conjuntos de dados com taxas de atualização diferentes: dados horários de monitoramento de qualidade do ar, e dados mensais de pré-sal. Dessa forma, a dimensão de atualidade pôde comprovar se os dados realmente possuíam os intervalos de atualização declarados, ou não, calculando os desvios temporais entre cada par de medições.

A.1.5 Precisão (Precision)

A dimensão de precisão lida com a possibilidade de discriminação de dados com valores similares. O registro de nomes próprios por somente sua primeira letra (e.g. "M" para "Matheus") possui uma baixa precisão, pois múltiplos valores inicialmente diferentes (e.g. "Matheus" e "Maria") se tornam indistinguíveis após o registro. A especificação dá como exemplo para esta dimensão o fato de que valores numéricos com 5 casas decimais permitem diferentes funcionalidades quando comparados a valores com 2 casas decimais.

Seguindo o exemplo da ISO/IEC, e dado que não é trivial analisar dados textuais, a dimensão de precisão é analisada na plataforma contando a quantidade de casas decimais disponíveis para valores ponto flutuante.

Não é também possível estabelecer uma corretude, ou ditar de maior qualidade valores com mais casas decimais, pois existe um equilíbrio subjetivo que deve ser calculado ma-

nualmente entre espaço de armazenamento gasto e benefício advindo de maior precisão. Em termos simples, não é o caso que uma maior precisão é sempre "melhor", mas sim que a precisão deve ser suficiente para as necessidades do sistema.

A.1.6 Compreensibilidade (Understandability)

A dimensão de compreensibilidade é bastante subjetiva; ela é definida, de modo geral, como "dados que podem ser lidos e interpretados pelos usuários apropriadamente". O exemplo dado pela ISO é a representação de um país por uma sigla (e.g. "BR"), ao invés de um código numérico arbitrário (e.g. "3").

Como não é possível pedir a um algoritmo que detecte a compreensibilidade de um certo dado diretamente, a interpretação utilizada para este trabalho foi em duas partes:

a) o dado a ser medido possui, ou não, uma unidade de medida definida; e

b) o dado a ser medido possui, ou não, uma descrição associada a ele.

Por exemplo, a coluna "NOx", dos dados de monitoramento da qualidade de ar, possui unidade de medida bem definida (i.e. $\frac{\mu g}{m^3}$) e uma descrição associada (i.e. "Óxidos de Nitrogênio"). Assim, o programa não considera que essa coluna possui uma má compreensibilidade. Colunas que não possuem unidade de medida definida ou descrição associada são destacadas como possuindo má compreensibilidade.

Essa análise, é claro, não é perfeita; a descrição presente pode não ser suficiente para transmitir o real propósito de um dado, e a falta dela não necessariamente significa que o dado não é facilmente compreendido. Porém, dentro das limitações de uma análise automática e agnóstica, ela é razoável, e serve aos nossos propósitos de validação.

A.2 DIMENSÕES SUBJETIVAS

A.2.1 Credibilidade (Credibility)

A dimensão de credibilidade refere-se a dados que são considerados verdadeiros e críveis. Isto é, tira o questionamento da qualidade do dado da plausibilidade do valor em si, e o passa para a organização ou entidade que registrou aqueles dados em primeiro lugar.

Credibilidade é um atributo totalmente subjetivo e dependente de contexto. Uma entidade renomada pode ter dados geralmente considerados críveis, mas que são enviesados em certos tópicos; e, paralelamente, uma entidade pouco conhecida pode ainda assim possuir dados 100% verificáveis e meticulosamente registrados. Dessa maneira, salvo a comparação com uma lista obrigatoriamente incompleta e imprecisa de "fontes confiáveis de dados", não é possível realizar uma análise automática dessa dimensão.

A exemplo, dados provenientes de agências governamentais relacionados ao monitoramento da qualidade do ar podem ser considerados críveis, pois possuem pouco impacto

na percepção pública e sua falsificação ou distorção não possuiria grande ganho político; por outro lado, dados de casos novos de uma epidemia em andamento possuem maior probabilidade de terem sido distorcidos, alterados ou falsificados para aliviar a percepção pública em relação ao governo. Ambos são dados publicados por entidades governamentais (que poderiam estar listadas como "fontes confiáveis"), mas possuem credibilidade em níveis diferentes.

A.2.2 Acessibilidade (Accessibility)

A dimensão de acessibilidade mede o nível ao qual dados podem ser acessados, em especial por pessoas com necessidades especiais. A especificação dá dois contraexemplos: dados guardados como imagens, impedindo portanto sua leitura por indivíduos com pouca ou nenhuma visão; ou dados guardados como arquivos de áudio, prejudicando aqueles com pouca ou nenhuma audição.

Mais uma vez, a análise automática dessa dimensão não é trivial. Fotografias paisagísticas e gravações de músicas instrumentais não possuem equivalente textual – ainda que possam ser descritas em forma de texto. Assim, existem variados casos em que é adequado ao banco possuir dados não textuais, ainda que estes sejam inacessíveis a pessoas com necessidades especiais.

Uma possível análise para a dimensão seria a verificação da presença, ou não, de uma descrição escrita para cada item não textual do conjunto de dados. Contudo, muitas plataformas automaticamente adicionam descrições ultra genéricas – como "Imagem" para imagens –, e sua presença não contribui para a acessibilidade do dado. Por isso, foi decidido que a dimensão seria melhor avaliada se implementada corretamente de alguma outra maneira.

A.2.3 Eficiência (Efficiency)

A dimensão de eficiência refere-se ao nível ao qual dados podem ser processados e são performáticos quando usadas quantidades apropriadas de recursos. Como exemplo, uma medição de temperatura pode ser registrada de duas formas: como valor numérico (e.g. "30"), dado que algum metadado informe sua unidade de medida; ou como valor textual (e.g. "30 °C"). Em ambas as formas, é possível calcular, por exemplo, a média de temperatura dentre múltiplas entradas de um banco de dados. Contudo, enquanto que é trivial somar todos os valores numéricos de um banco, é necessário um esforço adicional para realizar o cálculo quando se lida com dados textuais: é preciso primeiro extrair o valor numérico do texto. Assim, o banco que utiliza valores numéricos para representar a temperatura possui maior eficiência quando comparado ao textual.

Similarmente à dimensão de precisão, o cálculo de eficiência é subjetivo e deve ser verificado individualmente para cada métrica medida. Contudo, diferentemente de precisão,

a eficiência não possui uma particularidade facilmente mensurável algoritmicamente. Um arquivo de metadados que informa que uma coluna seria mais eficiente se numérica já é um indicativo por si só, para seu criador, que a eficiência do banco está faltante.

A.3 DIMENSÕES DE RESPONSABILIDADE DO FORNECEDOR DOS DADOS

A.3.1 Conformidade (Compliance)

A dimensão de conformidade se dá na forma de dados que aderem a padrões, convenções e regulamentações em efeito. O maior exemplo disso são as preocupações de privacidade para o armazenamento de dados médicos pessoais; em certas legislações, é necessário uma proteção adicional que impeça o mau uso desses dados, por serem dados sensíveis.

Esta é a primeira de algumas dimensões cuja responsabilidade de manter recai sobre o fornecedor dos dados. Para que a plataforma tenha acesso aos dados, é necessário que eles estejam em conformidade com as regulamentações vigentes; caso contrário, o mero ato de utilizar os dados para análise já seria uma violação dessa dimensão. Assim, não há como averiguar programaticamente a conformidade dos dados; esta é uma análise compulsoriamente externa à plataforma.

A.3.2 Confidencialidade (Confidentiality)

A dimensão de confidencialidade lida com dados que só podem ser acessados por usuários autorizados. Um grande exemplo são dados de saúde de indivíduos, que em muitas legislações não podem ser divulgados ou acessados por ninguém sem consentimento do indivíduo a qual se referem. Outro exemplo significativo são senhas salvas em bancos de dados de serviços com autenticação.

Em muitos casos, a confidencialidade de um sistema é assegurada via criptografia, e outras técnicas de segurança de bancos de dados – como *hashing*, *salting* e *peppering*. De qualquer maneira, para que os dados sejam acessados pela plataforma, eles não podem estar protegidos; assim, uma análise automatizada do ponto de vista da plataforma não faz muito sentido. É necessário, ao invés disso, uma ferramenta de uso interno para assegurar a confidencialidade dos bancos de dados.

A.3.3 Rastreabilidade (Traceability)

A dimensão de rastreabilidade refere-se a informações de acesso e modificação de dados. O exemplo dado pela ISO fala de bancos de dados de administrações públicas e argumenta que, para uma boa rastreabilidade, a instituição deve manter logs sobre acessos a dados confidenciais.

Essa dimensão só pode ser medida no ponto de acesso dos dados – seja uma API, ou uma interface. Ela não consta em bancos de dados isolados, como os arquivos usados nesse projeto, então não é trivial analisá-la pela plataforma. Seria necessário uma ferramenta mais dedicada para esse propósito, e a disponibilização dessas informações através das plataformas que geram os dados – no nosso caso, o data.rio e o dados.gov.br.

A.3.4 Disponibilidade (Availability)

A dimensão de disponibilidade mede a possibilidade de dados serem acessados por usuários ou aplicações autorizadas, mesmo durante operações tradicionalmente limitantes, como *backups* ou acessos concorrentes. Refere-se comumente a essa dimensão como *uptime*. Por exemplo, se dados de um sistema são inacessíveis durante operações de backup, então o sistema não possui disponibilidade ideal.

Não é possível, utilizando dados isolados e separados de sua interface original, determinar a disponibilidade do sistema – afinal, eles não dependem mais do sistema original. Assim, a medição de disponibilidade só pode estar sob responsabilidade do sistema original.

A.3.5 Portabilidade (Portability)

A dimensão de portabilidade mede o nível ao qual dados podem ser instalados, substituídos ou movidos de um sistema para outro, porém ainda preservando a qualidade de dados pré-existente. Um exemplo simplório é o processo de fotocópia (ou a metonímia, Xerox); é um processo com perdas, pois uma fotocópia não possui a mesma qualidade da original. Assim, se uma fotografia não pode ser escaneada com alta fidelidade, somente fotocopiada, o sistema possui baixa portabilidade, pois perde-se qualidade para movê-la para outro sistema.

Os dados utilizados por esse projeto são todos de alta portabilidade – afinal, para serem utilizados pela plataforma, foi necessário movê-los de um sistema para outro. Dessa maneira, a medição da portabilidade de dados recai sobre a plataforma que disponibiliza os dados; se eles são introduzidos à plataforma, isso incorre que provavelmente são portáveis.

A.3.6 Recuperabilidade (Recoverability)

A dimensão de recuperabilidade se refere à preservação da operacionalidade do sistema, independentemente da situação em que o sistema se encontra – como, por exemplo, em eventos de falha física. Pode-se listar como alguns exemplos de ações pró-recuperabilidade backups e commits em sistemas de versionamento. Se um disco rígido queima, um sistema com boa recuperabilidade não deve perder dados.

Similarmente a algumas das dimensões listadas anteriormente, esta é uma dimensão que não pode ser mensurada pela plataforma. Seria necessário uma análise à parte e in-

terna, direcionada à averiguação da recuperabilidade do sistema. Por exemplo, realizando conferência de backups e instalando alguma forma de monitoramento do funcionamento de partes físicas.

APÊNDICE B – DESCRIÇÃO DO FORMATO DO ARQUIVO DE METADADOS

Este apêndice contém a formatação utilizada pelo arquivo de metadados. O arquivo utilizado pela implementação do trabalho possui formato JSON – mas sua implementação não seria significantemente diferente em outros formatos de notação estruturada.

A criação do arquivo de metadados não pôde ser automatizada, e portanto foi manual, tomando como base informações disponibilizadas pelos fornecedores de dados. Assim, foi a etapa mais laboriosa da importação dos dados.

B.1 ESTRUTURA BÁSICA

O arquivo possui uma propriedade raiz, "**config**", que possui em si todas as tabelas acessíveis do banco em forma de propriedades filhas (Código 1). Cada uma dessas propriedades filhas consiste em uma lista de objetos de metadados que descrevem, cada, características de um campo da tabela de dados.

Código 1 – Estrutura básica do arquivo de metadados em formato JSON

```
{
    "config": {
        "dados-horarios-ar": [ ... ],
        "dados-mensais-presal": [ ... ]
    }
}
```

B.2 DADOS CONTÍNUOS

O Código 2 exemplifica um dos objetos de metadados, referente ao campo de umidade relativa do ar ("UR") do conjunto de dados de monitoramento de qualidade do ar. Este dado é uma variável contínua, pois qualquer valor dentre um intervalo especificado é considerado correto.

Código 2 – Estrutura do metadado de “Umidade Relativa do ar”

```
{
    "name": "UR",
    "description": "Umidade Relativa do ar",
    "type": "range",
    "value": [0, 100],
    "unit": "%"
}
```

O significado de cada propriedade do metadado exemplificada no Código 2 está a seguir:

name O nome do campo presente no banco de dados, para que seja estabelecida a relação entre dado e metadado.

description Uma descrição para humanos do que o dado representa. Alguns campos possuem identificações pouco descritivas – como "UR" para umidade relativa do ar, ou "RS" para radiação solar. Assim, a descrição se faz necessária para melhor compreensibilidade do dado.

type Especifica se a variável é discreta ou contínua. Um tipo "*range*" indica que a variável é contínua, e possui um intervalo de valores possíveis.

value Os valores possíveis para a variável. Como ela é contínua, este possui os limites inferior e superior de valores para a variável.

unit Determina a unidade de medida do campo. Em alguns casos, como nomes identificadores, pode não haver unidade de medida associada, e este item deve possuir o valor "n/a" (vide Código 3).

B.3 DADOS DISCRETOS

O Código 3 exemplifica uma variável discreta, que possui lista de valores possíveis. Isso é especificado pelo valor "*list*"na propriedade "**type**". Diferentemente das variáveis contínuas, todo valor possível para dados discretos fica listado nos metadados. Qualquer valor que não esteja presente na lista é considerado inválido.

Código 3 – Estrutura do metadado de “Estado”

```
{
  "name": "Estado",
  "description": "Nome do estado de origem",
  "type": "list",
  "value": [ "Acre", "Alagoas", ..., "Sergipe", "Tocantins" ],
  "unit": "n/a"
}
```

B.4 DADOS IRRESTRITOS

O Código 4 possui um exemplo de dado irrestrito; isto é, dado cujo valor está sempre correto. Essa característica é indicada pelo valor "*any*"na propriedade "**type**". Dados irrestritos costumam ser dados textuais, como nomes, ou identificadores, como no exemplo.

Restringir possibilidades de valores seria excluir dados potencialmente acurados. Assim, não há restrição.

Código 4 – Estrutura do metadado de “Identificador”

```
  {
    "name": "GID",
    "description": "Identificador para o banco de dados",
    "type": "any",
    "unit": "n/a"
  }
```

B.5 DADOS COM RESTRIÇÕES DE CONSISTÊNCIA

Finalmente, o Código 5 exemplifica um dado com restrição de consistência. O campo "Petroleo" deve ser igual à soma (operação definida pela propriedade "**operation**") dos campos "Oleo" e "Condensado" (definidos pela propriedade "**columns**").

Código 5 – Estrutura do metadado de “Volume de petróleo”

```
{
    "name": "Petroleo",
    "description": "Volume de pertróleo (soma de óleo e condensado)",
    "type": "range",
    "value": [0, 65000],
    "unit": "bbl/dia",
    "consistency": {
        "operation": "+",
        "columns": ["Oleo", "Condensado"]
    }
}
```