# ANÁLISE DA QUALIDADE DOS DADOS PÚBLICOS DE VACINAÇÃO DA COVID-19 NO BRASIL

**Arthur Ricardo de P. Oliveira, Beatriz Marmo Scaglione, Mário Olimpio de Menezes**

Faculdade de computação e Informática – Universidade Presbiteriana Mackenzie
São Paulo– SP – Brasil

{31824617,31811884}@mackenzista.com.br,
1146066@mackenzie.com.br

**Resumo.** *Atualmente, a divulgação de dados abertos governamentais (Open Government Data) vem aumentando, os órgãos públicos criam e divulgam uma grande variedade de dados sobre a sociedade. Com a pandemia, a disponibilidade de dados precisos e de boa qualidade são essenciais para auxiliar no controle da COVID -19 e na orientação de medidas de saúde pública. Entretanto, se dados de baixa qualidade forem utilizados podem acabar afetando a qualidade das medidas e orientações. Sendo a qualidade dos dados um fator determinante para a maior assertividade das ações públicas, o objetivo deste trabalho foi avaliar a qualidade dos dados públicos de vacinação contra a COVID-19 disponibilizados pelo Governo Brasileiro. As análises feitas neste artigo utilizam dimensões de qualidade de dados para indicar problemas de qualidade de dados.*

*Palavras-chave: Dados Abertos Governamentais, Políticas Públicas, Qualidade de Dados, Transparência Governamental, Dados Abertos, Dados de Saúde, Vacina, COVID-19.*

***Abstract.*** *Nowadays, the disclosure of government data (Open Government Data) has been increasing, so quality data is essential for better decision making and analysis, in every sector. With the pandemic, the availability of precise and good quality data is essential to aid the control of the COVID-19 and to guide public health measures. However, if low quality data is used, this can affect the quality of health measures and guidelines. Considering data quality as a determinant factor for the better assertiveness of public actions, this paper's main objective was to evaluate the quality of the COVID-19 vaccination public data provided by the Brazilian Government. The analysis done in this paper utilized data quality dimensions to indicate data quality problems.*

*Key-words: Open Government Data, Public Policies, Data Quality, Government Transparency, Open Data, Health Data, Vaccine, COVID-19.*

## 1. Introdução

Uma série de movimentos de dados abertos surgiu ao redor do mundo, com destaque para as iniciativas de Open Data (abertura de dados) promovidas pelo ex-Presidente dos Estados Unidos da América, Barack Obama, em 2009, e a Parceria de Governo Aberto, feita em 2011, como uma possibilidade de fornecer aos cidadãos e entidades interessadas informações governamentais (SILVA, 2020).

Dados abertos governamentais (open government data, OGD) são dados públicos, publicados na Web em formato aberto, estruturado e compreensível logicamente, de modo que qualquer pessoa possa livremente acessar, reutilizar, modificar e redistribuir, para quaisquer finalidades, estando sujeito a, no máximo, exigências de creditar a sua autoria e compartilhar sob a mesma licença (POSSA

De acordo com Quarati e De Martino (2019), a disseminação dos dados abertos governamentais é considerada a força motriz do crescimento econômico e social, além de ser um fator essencial para a publicidade das ações governamentais. Porém, a alta disponibilidade de dados não garante a qualidade das informações presentes n

A divulgação de dados pode mostrar que a qualidade dos dados sobre os quais são tomadas decisões importantes é ruim. O sucesso na abertura de dados governamentais requer mais do que a simples provisão de acesso aos dados. Também são necessários o aprimoramento da qualidade das informações governamentais, a criação e a institucionalização de uma cultura de governo aberto e o fornecimento de ferramentas e instrumentos com os quais os dados serão utilizados (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

Neste contexto, a qualidade dos dados tem sido um conceito estudado e aplicado em vários campos, dentre os quais a saúde. O Brasil dispõe de uma ampla rede de Sistemas de Informação em Saúde (SIS) de contexto nacional, com a maioria de suas informações disponíveis na Internet (Departamento de Informática do SUS DATASUS) (PICCOLO,2018).

O DATASUS surgiu em 1991, com a criação da Fundação Nacional de Saúde, e já desenvolveu mais de 200(duzentos) sistemas que auxiliam diretamente o Ministério da Saúde no processo de construção e fortalecimento do Sistema Único de Saúde (SUS) (PICCOLO,2018).

As informações disponibilizadas pelo DATASUS podem servir para subsidiar análises objetivas da situação sanitária, tomadas de decisão e elaboração de programas de ações de saúde. Neste sentido, a garantia da qualidade dos dados é condição essencial para a análise objetiva da situação sanitária, assim como para a tomada de decisões e para a programação de ações de saúde (PICCOLO,2018).

Disponibilizar dados abertos em tempo real, permite que pesquisadores de diferentes formações utilizem diversos métodos analíticos para construir evidências de forma rápida e eficiente (COSTA-SANTOS, 2021).

Os sistemas de vigilância epidemiológica precisam ser projetados tendo a qualidade dos dados como alta prioridade e, assim, promovendo, em vez de depender dos esforços dos usuários para garantir a qualidade dos dados (COSTA-SANTOS, 2021).

Com a pandemia da Covid-19, a disponibilidade de dados precisos e de qualidade são essenciais para orientar as medidas e políticas de saúde pública. Além disso as vacinas são consideradas uma das conquistas médicas mais importantes que salvam e melhoram a vida de milhões de pessoas em todo o mundo a cada ano (GIANFREDI, 2021).

Dados de imunização de alta qualidade facilitam a gestão, planejamento financeiro e capacidades de previsão de vacinas dos programas nacionais de imunização (PNI). Um pré-requisito para dados de qualidade é um sistema de informação que monitore a administração de vacinas e facilite a agregação e análise das informações de cobertura. Desde 2002, o Grupo Consultivo Técnico (GCT) da Organização Pan-Americana da Saúde (OPAS) sobre Doenças Preveníveis por Vacina para a Região das Américas emitiu recomendações para que os países melhorem a qualidade de seus dados de imunização e sistemas de informação (TRUMBO, 2018).

A partir disso, essa pesquisa se propôs a fazer uma análise e levantamento da qualidade dos dados públicos de vacinação contra COVID-19, disponibilizados pelo Governo Federal através do DATASUS, pois acredita-se que dados de boa qualidade podem contribuir com o aumento da transparência governamental, melhor organização e qualidade na criação e manutenção de políticas públicas.

## 2. Referencial Teórico

### 2.1. Dados Abertos

Tecnicamente, por dados abertos (open data) entendem-se os dados que qualquer pessoa pode livremente utilizar, reutilizar e redistribuir, estando sujeito, no máximo, à exigência de creditar a sua autoria à fonte original e de compartilhar sob a mesma licença em que foram apresentados (OPEN KNOWLEDGE FOUNDATION, 2014).

Para satisfazer essa classificação, o dado deve estar disponível por inteiro, em formato conveniente e modificável e por um custo razoável de acesso e reprodução. Deve ser fornecido sob termos que permitam sua utilização, reutilização (incluindo o cruzamento com outros conjuntos de dados) e redistribuição, não havendo discriminação de áreas de atuação, pessoas, grupos ou finalidades.

Em geral, essas características são comportadas por dados (incluindo, mas não se limitando a, textos, planilhas de dados, transcrições e gravações audiovisuais, etc.) representados em meio digital, estruturados em formato aberto, processáveis por máquina, referenciados na web e disponibilizados sob uma licença aberta que permi'a sua livre utilização, implementação ou cruzamento (OPENGOVDATA, 2007).

O portal OpenDATASUS (https://opendatasus.saude.gov.br) é uma ferramenta do Ministério da Saúde para divulgação de bases dados abertos sobre saúde para a sociedade.

Uma iniciativa que visa melhorar a qualidade de dados abertos é GO FAIR, que define princípios que melhoram a qualidade de dados e seu uso. Seu objetivo é implementar os princípios "FAIR" aos dados, tornando-os mais fáceis de encontrar, acessíveis, interoperáveis e reutilizáveis (Findable, Acessible, Interoperable e Reusable – FAIR). Estes princípios tornam o uso de dados e o seu entendimento mais fácil, e, portanto, são importantes para dados públicos abertos (WILKINSON, 2016).

### 2.2 Categorias e dimensões de qualidade dos dados

O termo qualidade de dados refere-se tanto às características associadas a dados de alta qualidade quanto aos processos usados para medir ou melhorar a qualidade dos dados (DAMA, 2017).

Os dados são de alta qualidade na medida em que atendem às expectativas e necessidades dos consumidores de dados. Ou seja, se os dados estão aptos para os fins a que pretendem aplicá-los. São considerados de baixa qualidade se não forem adequados para esses fins. A qualidade dos dados depende, portanto, do contexto e das necessidades do consumidor de dados (DAMA, 2017).

A qualidade de dados é um conceito multidimensional e é por isso que os padrões de qualidade de dados devem incluir uma série de características de qualidade que incorporam o conceito de "adequação ao uso". Esses elementos dos padrões de qualidade de dados devem ser considerados e equilibrados no desenho, implementação e validação dos processos e procedimentos de gerenciamento de dados (JESILEVSKA, 2017).

Atualmente, as atividades e a tomada de decisões em uma organização e em nível de país são baseadas em dados estatísticos e informações obtidas a partir da análise de dados desses dados. A análise de dados oferece várias possibilidades para a construção de processos confiáveis e precisos para tomada de decisão. Dados de baixa qualidade podem implicar em muitas consequências negativas, por exemplo, a má qualidade dos dados aumenta os custos operacionais, uma vez que tempo e outros recursos são gastos para detectar e corrigir erros (JESILEVSKA, 2017).

Para começar, erros de dados que não são identificados e corrigidos podem ter impactos econômicos e sociais extremamente negativos em uma organização (BALLOU, 2004; STRONG-WANG, 1996).

Para melhorar a qualidade dos dados, bem como avaliar o nível atual de qualidade dos dados, o efeito das iniciativas de qualidade dos dados deve ser medido. Vários autores apontam que: "Só o que pode ser medido pode ser melhorado" (STRONG-WANG, 1996).

Uma dimensão de qualidade de dados é um recurso ou característica mensurável dos dados. O termo dimensão é usado para fazer a conexão com as dimensões na medição de objetos físicos (por exemplo, comprimento, largura, altura). As dimensões de qualidade de dados fornecem um vocabulário para definir os requisitos de qualidade de dados. A partir daí, eles podem ser usados para definir os resultados da avaliação inicial da qualidade dos dados, bem como a medição contínua. Para medir a qualidade dos dados, uma organização precisa estabelecer características que sejam importantes para os processos de negócios (que valem a pena ser mensuradas) e mensuráveis. As dimensões fornecem uma base para regras mensuráveis, que devem estar diretamente conectadas a riscos potenciais em processos críticos (DAMA, 2017).

De acordo com o framework de Strong-Wang (1996), que possui um foco nas percepções de consumidores de dados, são descritas 15 dimensões entre 4 categorias de qualidade dos dados como intrínseca, acessibilidade, contextual e representacional.

O quadro 1 apresenta a compilação do conjunto de categoria qualidade de dados, suas dimensões e descrições. O quadro 2 apresenta as dimensões de qualidade dos dados e suas descrições, segundo o framework de Strong-Wang (1996), já o quadro 3 apresenta as dimensões de qualidade dos dados e suas descrições, segundo o livro DAMA (2017).

**Quadro 1. Categorias de Qualidade de dados, suas respectivas dimensões e descrições, com base em Strong, Lee e Wang.**

| Categoria de Qualidade de Dados | Dimensões de Qualidade de Dados | Descrição |
|---|---|---|
| Intrínseca | Acurácia, Objetividade, Credibilidade, Reputação | A categoria intrínseca diz respeito a qualidade já existente na informação de forma natural. |
| Acessibilidade | Acessibilidade, Segurança de Acesso | Por fim, a categoria de acessibilidade diz respeito ao modo de acesso à informação e é avaliada a partir da usabilidade, segurança, entre outras dimensões. Esta categoria mais avalia o sistema onde a informação está armazenada e seu modo de armazenamento e disponibilização do que a informação em si. |
| Contextual | Relevância, Valor-Adicionado, Temporariedade, Completude, Quantidade de dados | A categoria contextual apresenta a qualidade de uma informação quando é requisitada em determinado momento. Em dado contexto, a informação é avaliada de acordo com a importância que tem, o quão adequada é, o custo de armazenamento no tempo não utilizada e o tempo em que será útil para a corporação. |
| Representacional | Interpretabilidade, Facilidade de Entendimento, Representação Concisa, Consistência Representacional | A categoria de representação da informação diz respeito ao modo como ela deve estar quando disponível. Assim, a informação deve ser legível, de fácil compreensão e compatível. |

Fonte: Framework Strong-Wang(1996), tradução dos autores.

**Quadro 2. Dimensões de Qualidade de dados e suas respectivas descrições, com base em Strong, Lee e Wang.**

| Dimensão de Qualidade de Dados | Descrição |
|---|---|
| Acurácia | Até que ponto os dados são corretos, confiáveis e certificados como livres de erros. |
| Objetividade | Até que ponto os dados são imparciais (sem preconceitos) e imparciais. |
| Credibilidade | Até que ponto os dados são aceitos ou considerados verdadeiros, reais e confiáveis |
| Reputação | A medida em que os dados são confiáveis ou altamente considerados em termos de sua fonte ou conteúdo. |
| Acessibilidade | Até que ponto os dados estão disponíveis ou podem ser recuperados com facilidade e rapidez. |
| Segurança de Acesso | Até que ponto o acesso aos dados pode ser restrito e, portanto, mantido em segurança. |
| Relevância | Até que ponto os dados são aplicáveis e úteis para a tarefa em questão. |
| Valor-Adicionado | Até que ponto os dados são benéficos e oferecem vantagens de seu uso. |

| | |
|---|---|
| Temporariedade | Até que ponto a idade dos dados é apropriada para a tarefa em questão. |
| Completude | A extensão em que os dados são de amplitude, profundidade e escopo suficientes para a tarefa em mãos. |
| Quantidade de dados | A extensão em que os dados estão disponíveis de várias fontes de dados diferentes. |
| Interpretabilidade | A extensão em que os dados estão em linguagem e unidades apropriadas e as definições de dados são claras. |
| Facilidade de Entendimento | Até que ponto os dados são claros, sem ambiguidade e facilmente compreendidos |
| Representação Concisa | A extensão em que os dados são representados de forma compacta sem serem esmagadores (ou seja, breves na apresentação, mas completos e diretos ao ponto). |
| Consistência Representacional | A medida em que os dados são sempre apresentados no mesmo formato e são compatíveis com os dados anteriores. |

Fonte: Framework Strong-Wang(1996), tradução dos autores.

**Quadro 3. Dimensões de Qualidade de dados e suas respectivas descrições, com base no livro DAMA-DMBOK.**

| **Dimensão de Qualidade de Dados** | **Descrição** |
|---|---|
| Completude | A proporção de dados armazenados *versus* o total (100%) possível. |
| Singularidade | Nenhuma instância de dado deverá ser armazenada mais de uma vez com o mesmo identificador. |
| Pontualidade | O grau no qual os dados representam a realidade temporal do ponto consultado. |
| Validação | Medida para avaliar se os dados seguem a sintaxe proposta no dicionário. |
| Acurácia | O grau no qual os dados representam corretamente o objeto do mundo real que representam. |
| Consistência | Se há diferença quando comparamos representações de algo com sua definição. |

Fonte: Livro DAMA-DMBOK(2017), tradução dos autores.

## 3. Materiais e Métodos

### 3.1 Dimensões de qualidade dos dados

Após a revisão bibliográfica, foram definidos os critérios e dimensões de qualidade dos dados baseados no framework de Strong-Wang e nos conceitos presentes no livro DAMA-DMBOK.

Segundo Jesilevska (2017), as 10 dimensões mais significantes para analisar a qualidade dos dados em uma pesquisa cientifica são: Objetividade, Completude, Representatividade, Acurácia, Qualidade da metodologia, Realidade, Coerência, Acessibilidade, Utilidade e Interpretabilidade.

Dessa forma as dimensões de qualidade que iremos utilizar para avaliar a qualidade dos dados, suas descrições e respectivas métricas são:

Acessibilidade: Dimensão de qualidade de dados do framework Strong-Wang que avalia se o acesso aos dados tem algum problema. Para esta dimensão, o ideal seria que

o acesso aos dados seja permitido para todos, os dados possam ser encontrados e baixados facilmente e o dicionário de dados seja completo e descreva bem todos os campos de dados.

Consistência: Dimensão de qualidade do DAMA, utilizada para medir se não há diferença quando comparamos representações que querem dizer a mesma coisa com a sua definição. Neste caso, não deveria haver mais de dois termos para representar a mesma informação.

Completude: Dimensão de qualidade do DAMA, mede quantos campos de dados estão em branco comparado ao total de dados na base. A base não deve ter mais de 20% de seus dados em branco.

Interpretabilidade: Dimensão do framework Strong-Wang. Os dados devem ser entendidos e interpretados facilmente. Todos os campos das bases deveriam ser entendidos com facilidade.

Validação: Dimensão do DAMA, os dados devem seguir o formato definido no dicionário.

Escolhemos estas dimensões por serem dimensões importantes para o uso de dados abertos e que podem levar a decisões erronias se tiverem problemas, o que é grave como estamos tratando de dados de saúde.

### 3.2 Base de dados

As bases de dados analisadas neste trabalho são referentes a Campanha Nacional de Vacinação contra Covid-19 e estão disponíveis no portal OpenDataSUS (https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao).

Os registros de vacinação contra Covid-19 são disponibilizados sem identificação do cidadão, englobam o número de doses aplicadas, por UF e municípios, por um determinado período, por gênero, por faixa etária e por tipo de vacina.

Os dados estão contidos na Rede Nacional de Dados em Saúde (RNDS), e são disponibilizados desde o início da campanha, janeiro 2021, pela SI-PNI, e-SUS, APS e os sistemas próprios de estados e municípios que estão devidamente integrados com a RNDS.

No portal OpenDataSUS as bases são atualizadas diariamente e são disponibilizados o dicionário de dados, um manual para consumo da API, a API para acesso a todos os registros de vacinação, registros de vacinação de todos os estados divididos em 10 partes em arquivo csv e os registros de vacinação separados por cada estado, sendo que as bases de cada estado são separadas em 3 partes em arquivo csv.

Após a definição das dimensões que foram utilizadas no trabalho, foi feita a coleta das bases de dados dos estados do Acre, Alagoas, Santa Catarina e Distrito Federal.

O tamanho da base de registros de vacinação de todos os estados possui cerca de 200gb em csv. Por conta do tamanho das bases com todos os estados, o critério de seleção

definido foi analisar duas bases de dados de estados com um baixo IDH (AC e AL) e duas bases de estados com um alto IDH (DF e SC).

A base do estado do Acre possui cerca de 768mb, do Alagoas 2,8gb, do Distrito Federal 2,8gb e de Santa Catarina 7gb.

O Dicionário de dados das bases se encontra no quadro 4.

**Quadro 4. Dicionário de Dados das bases da Campanha Nacional de Vacinação contra Covid-19**

| Ordem | Campo | Descrição | Categoria |
|---|---|---|---|
| 1 | document_id | Identificador do documento | |
| 2 | paciente_id | Identificador do vacinado | |
| 3 | paciente_idade | Idade do vacinado | |
| 4 | paciente_dataNascimento | Data de nascimento do vacinado | |
| 5 | paciente_enumSexoBiologico | Sexo do vacinado | M = Masculino, F = Feminino |
| 6 | paciente_racaCor_codigo | Código da raça/cor do vacinado | 1; 2; 3; 4; 99 |
| 7 | paciente_racaCor_valor | Descrição da raça/cor do vacinado | 1 = Branca; 2 = Preta; 3 = Parda; 4 = Amarela; 99 = Sem informação |
| 8 | paciente_endereco_coIbgeMunicipio | Código IBGE do município de endereço do vacinado | |
| 9 | paciente_endereco_coPais | Código do país do endereço do vacinado | |
| 10 | paciente_endereco_nmMunicipio | Nome do município do endereço do vacinado | |
| 11 | paciente_endereco_nmPais | Nome do país de endereço do vacinado | |
| 12 | paciente_endereco_uf | Sigla da UF do endereço do vacinado | |
| 13 | paciente_endereco_cep | 5 dígitos para anonimizado e 7 dígitos para identificado | |
| 14 | paciente_nacionalidade_enumNacionalidade | Nacionalidade do vacinado | |
| 15 | estabelecimento_valor | Código do CNES do estabelecimento que realizou a vacinação | |
| 16 | estabelecimento_razaoSocial | Nome/Razão Social do estabelecimento | |
| 17 | estalecimento_noFantasia | Nome fantasia do estabelecimento | |
| 18 | estabelecimento_municipio_codigo | Código do município do estabelecimento | |
| 19 | estabelecimento_municipio_nome | Nome do município do estabelecimento | |
| 20 | estabelecimento_uf | Sigla da UF do estabelecimento | |
| 21 | vacina_grupo_atendimento_code | Código do grupo de atendimento ao qual pertence o vacinado | |
| 22 | vacina_grupo_atendimento_nome | Nome do grupo de atendimento ao qual pertence o vacinado | |
| 23 | vacina_categoria_codigo | Código da categoria | |
| 24 | vacina_categoria_nome | Descrição da categoria | |
| 25 | vacina_lote | Número do lote da vacina | |
| 26 | vacina_fabricante_nome | Nome do fabricante / fornecedor | |

| | | | |
|---|---|---|---|
| 27 | vacina_fabricante_referencia | CNPJ do fabricante / fornecedor | |
| 28 | vacina_dataAplicacao | Data de aplicação da vacina | |
| 29 | vacina_descricao_dose | Descrição da dose | |
| 30 | vacina_codigo | Código da vacina | |
| 31 | vacina_nome | Nome da vacina / produto | |
| 32 | sistema_origem | Nome do sistema de origem | |

Fonte: Portal OpenDATASUS

### 3.3 Manipulação e Análise dos dados

Para manipular e analisar as bases foi escolhida a linguagem Python 3 e o ambiente Jupyter Notebook, pois é uma linguagem de alto nível, atualizada, com grande variedade de bibliotecas e muito utilizada na manipulação de base de dados.

Dessa forma foi construído um script Python para manipular e analisar os dados. O script utiliza a biblioteca Pandas, uma biblioteca poderosa para manipulação e análise de dados, para fazer o import dos arquivos csv e unificar as 3 partes de cada base, pois as bases de cada estado são separadas em 3 partes.

Após obter a base unificada de cada estado é utilizado o pandas_profiling, uma biblioteca open source, disponível no Github que elabora uma análise exploratória dos dados.

É gerado um relatório em html, que consolida as informações da análise em 5 partes: Resumo, variáveis, interações, correlações, valores faltantes e amostra.

O resumo contém informações sobre cada variável das bases como: Quantidade de variáveis, quantidade de observações, quantidade de células vazias, porcentagem de células vazias, quantidade de linhas duplicadas, porcentagem de linhas duplicadas, o tamanho do dataset em memória, o tamanho de cada registro em memória e os tipos de variáveis.

A parte de variáveis analisa cada variável individualmente e mostra um relatório estatístico descritivo de cada uma delas, onde é possível observar todos os valores presentes em cada variável, valores de quantis, os valores mais comuns e sua representatividade e valores extremos, mínimo e máximo.

A parte de interações mostra a relação entre duas variáveis no quadro de dados como gráficos de dispersão.

A parte de correlações apresenta os coeficientes de correlação de Pearson, Spearman, Kendall e Phik entre cada variável.

A parte de dados faltantes apresenta os detalhes sobre dados faltantes de cada variável da base de dados.

A Parte de amostra apresenta as 10 primeiras e últimas linhas de cada base.

As figuras 1, 2, 3 e 4 apresentam os resultados do resumo gerado pela biblioteca ao analisar as bases de cada estado.

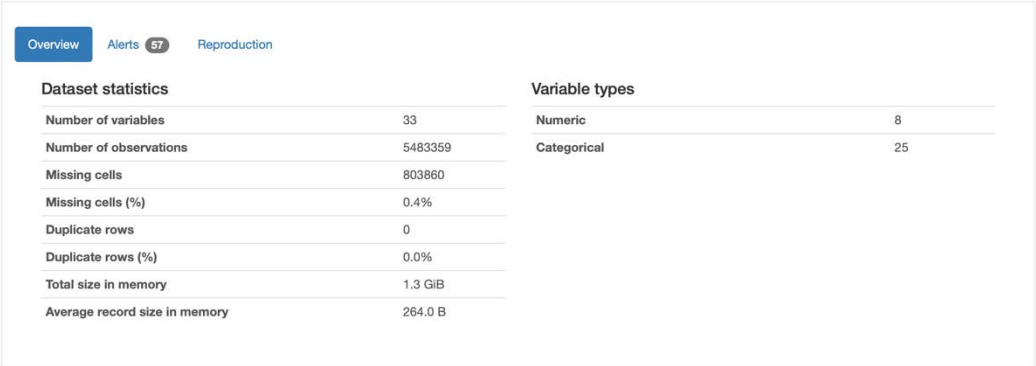


**Figura 1. Resumo da base do estado Alagoas a partir da análise gerada pelo *pandas_profiling***

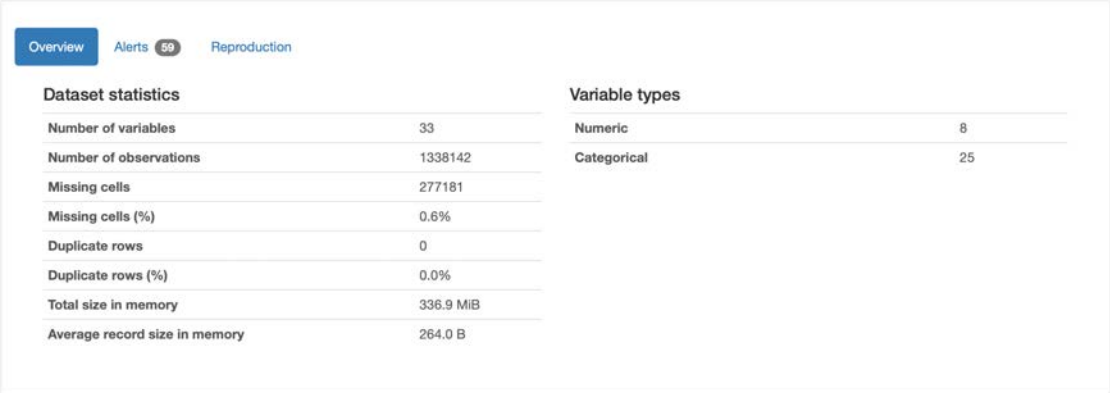


**Figura 2. Resumo da base do estado Acre a partir da análise gerada pelo *pandas_profiling***

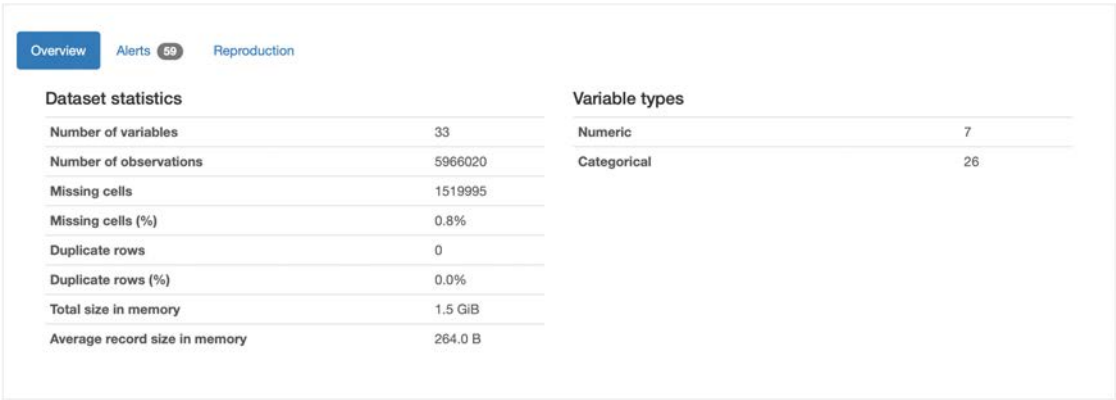


**Figura 3. Resumo da base do estado Distrito Federal a partir da análise gerada pelo *pandas_profiling***

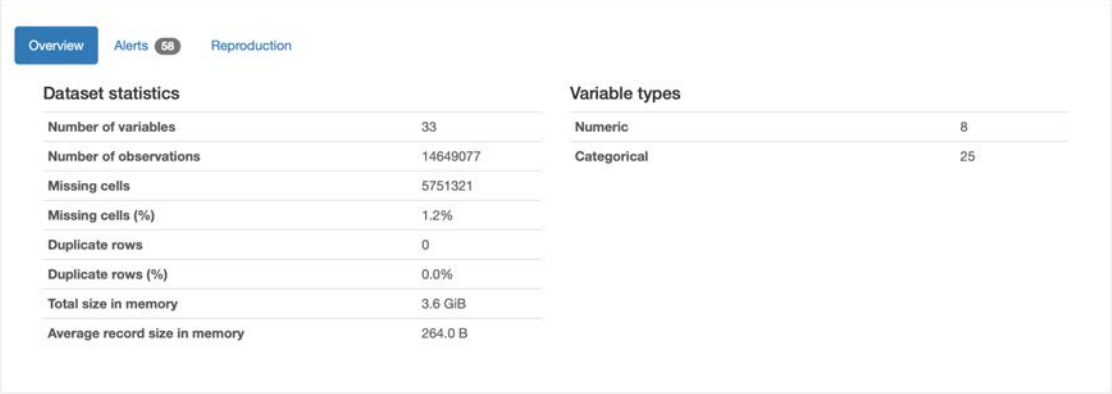


**Figura 4. Resumo da base do estado Santa Catarina a partir da análise gerada pelo *pandas_profiling***

Através dos relatórios gerados pelo *pandas_profiling* foi possível levantar as informações necessárias para o desenvolvimento da análise de qualidade dos dados das bases escolhidas.

Os resultados gerados pelo relatório foram consolidados e analisados em planilhas no programa Microsoft Excel, onde foi possível gerar as tabelas apresentadas na sessão de resultados.

Para analisar a dimensão de Acessibilidade, foi feito um levantamento das informações presentes no dicionário de dados e uma comparação com as informações presentes nas variáveis das bases de dados, além disso foi analisado a forma de como as bases são disponibilizadas no site do OpenDataSUS.

Para analisar a dimensão de Consistência, foi feito um levantamento do conteúdo das variáveis através do relatório gerado pelo pandas_profiling, onde foi possível observar a consistência das variáveis das bases de cada Estado.

Para analisar a dimensão de Completude, foi feito um levantamento da quantidade de dados faltantes de cada variável das bases a partir do relatório gerado pelo pandas_profiling.

Para analisar a dimensão de Interpretabilidade, foi feito um levantamento do conteúdo das variáveis, a partir do relatório gerado pelo pandas-profiling, e com isso foi possível observar a clareza das informações apresentadas.

Para analisar a dimensão de Validação, foi feito um levantamento do conteúdo das variáveis, a partir do relatório gerado pelo pandas-profiling, e uma comparação com a descrição dessas variáveis apresentadas no dicionário de dados para avaliar o quanto o conteúdo das variáveis era condizente com as descrições do dicionário.

## 5. Desenvolvimento e Resultados

As dimensões de qualidade de dados que foram utilizadas para avaliar as bases de dados escolhidas foram Acessibilidade, Completude, Consistência, Interpretabilidade e Validação. Estas dimensões foram escolhidas por serem importantes em dados públicos. As dimensões escolhidas são de diferentes autores e manuais, como mencionado no referencial teórico.

Os resultados encontrados para cada dimensão foram descritos com base nas métricas estabelecidas.

Acessibilidade: Falta de informação sobre as categorias no dicionário de dados: Há 32 campos e apenas 2 categorias estão definidas; somente nos campos sobre a raça e o sexo biológico do paciente há descrição da categoria e o que significa cada sigla nos campos de código.

Não há indicação de quais campos o preenchimento é obrigatório.

As bases não são divididas de acordo com quando são atualizadas; há informação de que as bases são atualizadas diariamente, porém não é possível baixar uma base de dados de uma data específica nem saber quais dados de uma base foram acrescentados em cada data.

Consistência: Nas bases analisadas, o campo "vacina_descricao_dose" utiliza mais de dois termos para definir categorias que parecem ser a mesma. Por exemplo, a base de Santa Catarina utiliza os termos "1º Reforço", "3ª Dose" e "Dose Adicional". A tabela abaixo mostra quais termos são utilizados neste campo em todas as bases analisadas.

**Quadro 5. Categorias encontradas no campo "vacina_descricao_dose"**

| Categorias | Base AC | Base AL | Base DF | Base SC |
|---|---|---|---|---|
| 1ª Dose | Utiliza | Utiliza | Utiliza | Utiliza |
| 2ª Dose | Utiliza | Utiliza | Utiliza | Utiliza |
| Reforço | Utiliza | Utiliza | Utiliza | Utiliza |
| 2º Reforço | Utiliza | Utiliza | Utiliza | Utiliza |
| Dose | Utiliza | Utiliza | Utiliza | Utiliza |
| Dose Adicional | Utiliza | Utiliza | Utiliza | Utiliza |
| Única | Não utiliza | Não utiliza | Utiliza | Utiliza |
| 2ª Dose Revacinação | Não utiliza | Utiliza | Utiliza | Não utiliza |
| 1ª Dose Revacinação | Não utiliza | Utiliza | Utiliza | Não utiliza |
| 3ª Dose | Não utiliza | Utiliza | Utiliza | Utiliza |
| 1º Reforço | Não utiliza | Não utiliza | Não utiliza | Utiliza |
| 3º Reforço | Não utiliza | Não utiliza | Não utiliza | Utiliza |
| Dose Inicial | Não utiliza | Não utiliza | Não utiliza | Utiliza |
| 3ª Dose Revacinação | Não utiliza | Não utiliza | Não utiliza | Utiliza |

Fonte: Autores

Completude: Através dos relatórios gerados pela biblioteca *pandas_profiling*, foi possível levantar as informações de completude da tabela.

Utilizando a métrica de Número de dados faltantes/Quantidade de colunas * Quantidade de linhas, é possível observar na tabela 1, que as bases possuem uma alta completude dos dados.

**Tabela 1. Valores de completude por base de cada estado**

| Estados | Número de linhas | Número de dados faltantes | % Valores completos |
|---|---|---|---|
| Acre | 1.338.142 | 277.181 | 99,4% |
| Alagoas | 5.483.359 | 803.860 | 99,5% |
| Distrito Federal | 5.966.020 | 1.519.995 | 99,2% |
| Santa Catarina | 14.649.077 | 5.751.321 | 98,8% |

Fonte: Autores

Entretanto ao olhar a completude por colunas, é possível identificar que algumas variáveis apresentam problemas de preenchimento.

Conforme mostra a tabela 2, as informações relacionadas a endereço, categoria e código de vacina representam os maiores percentuais de valores faltantes, de acordo com a métrica de Número de dados faltantes/Quantidade total de linhas.

Na base do estado do Acre as informações de endereço somam um percentual de 16% de valores faltantes, na base do Alagoas 8%, na base do Distrito Federal 17% e Santa Catarina 4%, este último apresenta uma concentração maior na variável de fabricante da vacina com 27% de valores faltantes.

Dessa forma se os dados de endereço forem utilizados para algum indicador, vão apresentar problemas em sua completude, afetando a qualidade da informação.

**Tabela 2. Quantidade/Percentual de valores faltantes por coluna das bases**

| Colunas | Acre | | Alagoas | | Distrito Federal | | Santa Catarina | |
|---|---|---|---|---|---|---|---|---|
| | Quantidade | Percentual | Quantidade | Percentual | Quantidade | Percentual | Quantidade | Percentual |
| document_id | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| paciente_id | 0 | 0,0 | 4 | 0,0 | 1 | 0,0 | 6 | 0,0 |
| paciente_idade | 0 | 0,0 | 4 | 0,0 | 1 | 0,0 | 6 | 0,0 |
| paciente_dataNascimento | 0 | 0,0 | 4 | 0,0 | 1 | 0,0 | 6 | 0,0 |
| paciente_enumSexoBiologico | 0 | 0,0 | 4 | 0,0 | 1 | 0,0 | 6 | 0,0 |
| paciente_racaCor_codigo | 0 | 0,0 | 4 | 0,0 | 4 | 0,0 | 6 | 0,0 |
| paciente_racaCor_valor | 0 | 0,0 | 4 | 0,0 | 4 | 0,0 | 6 | 0,0 |
| paciente_endereco_coIbgeMunicipio | 8.005 | 0,6 | 20.891 | 0,4 | 48.931 | 0,8 | 32.839 | 0,2 |
| paciente_endereco_coPais | 7.990 | 0,6 | 20.758 | 0,4 | 48.797 | 0,8 | 32.705 | 0,2 |
| paciente_endereco_nmMunicipio | 7.991 | 0,6 | 20.872 | 0,4 | 48.882 | 0,8 | 32.770 | 0,2 |
| paciente_endereco_nmPais | 7.990 | 0,6 | 20.758 | 0,4 | 48.797 | 0,8 | 32.705 | 0,2 |
| paciente_endereco_uf | 8.005 | 0,6 | 20.891 | 0,4 | 48.931 | 0,8 | 32.839 | 0,2 |
| paciente_endereco_cep | 167.593 | 12,5 | 329.391 | 6,0 | 795.722 | 13,3 | 455.119 | 3,1 |
| paciente_nacionalidade_enumNacionalidade | 612 | 0,0 | 11.021 | 0,2 | 2.671 | 0,0 | 5.706 | 0,0 |
| estabelecimento_valor | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| estabelecimento_razaoSocial | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| estalecimento_noFantasia | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| estabelecimento_municipio_codigo | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| estabelecimento_municipio_nome | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| estabelecimento_uf | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| vacina_grupoAtendimento_codigo | 0 | 0,0 | 1 | 0,0 | 2.410 | 0,0 | 0 | 0,0 |
| vacina_grupoAtendimento_nome | 0 | 0,0 | 1 | 0,0 | 2.410 | 0,0 | 623 | 0,0 |
| vacina_categoria_codigo | 32.526 | 2,4 | 171.104 | 3,1 | 232.230 | 3,9 | 513.642 | 3,5 |
| vacina_categoria_nome | 32.526 | 2,4 | 171.104 | 3,1 | 232.230 | 3,9 | 513.642 | 3,5 |
| vacina_lote | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| vacina_fabricante_nome | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| vacina_fabricante_referencia | 3.943 | 0,3 | 17.044 | 0,0 | 7.972 | 0,1 | 4.008.460 | 27,4 |
| vacina_dataAplicacao | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| vacina_descricao_dose | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| vacina_codigo | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| vacina_nome | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 |
| sistema_origem | 0 | 0,0 | 0 | 0,0 | 0 | 0,0 | 89.612 | 0,6 |
| Total | 1.338.142 | 20,7 | 5.483.359 | 15,0 | 5.966.020 | 25,0 | 14.649.077 | 39,0 |

Fonte: Autores

Interpretabilidade: Nos campos "vacina_grupoAtendimento_nome" e "vacina_categoria_nome", são descritos respectivamente o grupo de atendimento no qual pertence o vacinado e a descrição da categoria da vacina. Não há categorias destes campos no dicionário de dados. Os campos estão correlacionados, pois, para o mesmo paciente o campo vacina_grupoAtendimento_nome está preenchido com "Pessoas de 18 a 64 anos", e o campo vacina_categoria_nome, com Faixa Etária. Assim, afetando a clareza da informação.

Há também nestes campos o uso de termos parecidos para categorias diferentes, que demonstram falta de clareza na informação. O que pode parecer um problema de consistência, na verdade é um problema de interpretabilidadede e acessibilidade. Não são termos parecidos para definir a mesma coisa, mas sim termos parecidos para definir categorias diferentes, o que não foi explicado no dicionário

Os termos " Outros", " Outros Grupos" e " Outros Imunocomprometidos", utilizados no campo "vacina_grupoAtendimento_nome”, na verdade não são sinônimos para definir entradas que se categorizassem como "outros" em geral, e sim "outros" com base na categoria mais abrangente "vacina_grupoAtendimento_nome".

Abaixo, no quadro 6, temos amostras da base do Acre mostrando esse problema:

**Quadro 6. Amostras de problemas de interpretabilidade**

| Código do grupo de Atendimento | Nome do grupo de atendimento | Categoria | Nome da categoria |
|---|---|---|---|
| 926 | Outros | 9 | Trabalhadores de Saúde |
| 999999 | Outros Grupos | 114 | Outros |

| | | | |
|---|---|---|---|
| 111 | Outros Imunocomprometidos | 1 | Comorbidades |

Fonte: Base de dados do Acre da Campanha Nacional de Vacinação contra Covid-19 - OpenDataSUS

Outro problema de interpretabilidade é o preenchimento de dados no campo "paciente_enumSexoBiologico" com a sigla "I", enquanto no dicionário só estão definidas duas siglas como categorias, "F" e "M". Portanto, não é possível ter certeza do significado desta sigla. O uso de categorias que não estão definidas no dicionário acontece também no campo “paciente_racaCor_valor” ; no dicionário, estão definidas as categorias "Branca", "Preta","Parda", "Amarela" e "Sem informação", mas nas bases foi encontrada a categoria "Indígena".

Validação: No campo "paciente_nacionalidade_enumNacionalidade", ao invés de ter o nome do país escrito, está escrito somente uma sigla, que não está definida no dicionário. De acordo com esta dimensão, isto é um problema, pois não é indicado que o preenchimento seria feito com uma sigla.

Abaixo, incluímos um quadro resumo com estes resultados:

**Quadro 7. Resumo dos resultados**

| Dimensão | Resultados |
|---|---|
| Acessibilidade | -Falta de informações no dicionário<br>-Sem divisões por data dos arquivos |
| Completude | -Todas as bases com completude acima de 90% |
| Consistência | -As bases utilizam termos diferentes para representar a mesma categoria |
| Interpretabilidade | -Falta de informações no dicionário e uso de termos diferentes dos descritos no dicionário |
| Validade | -Tipos de categorias diferentes das descritas no dicionário |

Fonte: Autores

## 5. Conclusões e Recomendações

Após analisar os resultados, é possível notar que em todas as dimensões de qualidade de dados escolhidas para a análise existem problemas; para algumas, isto é mais grave do que em outras. Por se tratar de dados públicos governamentais, principalmente as dimensões de acessibilidade e interpretabilidade deveriam ter resultados melhores, pois é importante que a sociedade possa acessar e entender estes dados com facilidade, para conhecer o que acontece.

Além disso, problemas em outras dimensões podem ser vistos como não tão importantes, mas acabam atrapalhando o foco de quem utiliza as bases, que podem perder o tempo que poderiam utilizar para fazer análises tentando entender os campos.

Há também que se levar em consideração que foram tratados dados de saúde, então alguns problemas podem levar a conclusões errôneas. Por exemplo, a dimensão de qualidade de dados consistência com problemas no campo que descreve a dose pode fazer com que se conclua que menos ou mais pacientes tomaram uma dose específica.

Um ponto interessante é que foram escolhidas bases de estados desenvolvidos e menos desenvolvidos, com base no IDH, esperando que estados mais desenvolvidos possivelmente tivessem acesso mais fácil à tecnologia, e, portanto, tivessem menos problemas nas bases, porém foi possível observar que este não é o caso, e em algumas dimensões, o resultado foi o contrário: na questão de consistência, por exemplo, é possível notar que na base do Acre (estado com IDH baixo) há uso de menos termos diferentes para descrever a mesma categoria do que na base de Santa Catarina (alto IDH); na questão de completude, dos estados analisados, os com maiores percentuais de completude são Acre e Alagoas.

Para que as dimensões de qualidade de dados tenham melhores resultados, é recomendável a quem produz e publica estes dados:

-Utilizar métricas de dimensões de qualidade de dados e aplicá-las em processos de controle, para garantir o acompanhamento e a medição da qualidade;

-Seguir o que dizem iniciativas como a FAIR;

-Verificar as bases, o portal e o dicionário e atualizá-los caso note-se algum problema;

-Consultar quem utilizará os dados previamente para sugestões do que precisam;

-Receber e considerar comentários dos utilizadores e consumidores das bases sobre a experiência de usabilidade e corrigi-las se necessário.

- É sugerido também que o dicionário de dados seja preparado com mais detalhes; muitos dos problemas encontrados nas dimensões escolhidas poderiam ter sido evitados se o dicionário estivesse mais completo.

Para possíveis trabalhos futuros com base neste, é sugerido que trabalhos que utilizarem estas bases sejam analisados para procurar problemas relacionados aos dados relatados pelos autores e que propostas com base neste artigo possam ser feitas para melhorar as qualidades de dados de outras bases de saúde.

## 6. Referências

ALBU, O. B.; FLYVERBOM, M. (2016) Organizational Transparency: Conditions and Consequences. Business & Society, v. 58, n. 2, p. 268-297, 2016. Disponível em: <https://doi.org/10.1177/0007650316659851>.Acesso em: 12 maio 2021.

BATINI, Carlo et al. (2009) Methodologies for data quality assessment and improvement. ACM Computing Surveys, Estados Unidos da América, v. 41, n.3, 2009. Disponível em: <https://doi.org/10.1145/1541880.1541883>. Acesso em: 9 dez. 2021.

BIANCONE, Paolo et al. (2019) Data Quality Methods and Applications in Health Care System: A Systematic Literature Review. International Journal of Business and Management, Canadá, v. 14, n.4, 2019. Disponível em: <https://doi.org/10.5539/ijbm.v14n4p35>. Acesso em: 9 dez. 2021.

Brasil. Presidência da República. Casa Civil. Subchefia para Assuntos Jurídicos. Lei n. 12.527, de 18 de novembro de 2011. Portal da Legislação, Brasília, nov. de 2011. Disponível em:<http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.html>. Acesso em: 12 maio 20

Brasil. Presidência da República. Casa Civil. Subchefia para Assuntos Jurídicos. Decreto n. 8.777, de 11 de maio de 2016. Portal da Legislação, Brasília, mai. de 2016. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.html>. Acesso em: 12 maio 20

CHEN, Hong; HAILEY, David; WANG, Ning; et al. A Review of Data Quality Assessment Methods for Public Health Information Systems. International Journal of Environmental Research and Public Health, v. 11, n. 5, p. 5170–5207, 2014.

DAMA International. DAMA-DMBOK: Data Management Book of Knowledge. Basking Ridge, New Jersey, Estados Unidos da América: Technics Publications, 2017.

DAWES, Sharon S.; VIDIASOVA, Lyudmila; PARKHIMOVICH, Olga.(2016) Planning and designing open government data programs: An ecosystem approach.Government Information Quarterly, sem local indicado, v. 33, n. 1, 29 Jan. 2016.Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0740624X1630003X>. Acesso em: 12 mai. 2021.

FAGUNDES, Melissa Figueira; RIBEIRO JUNIOR, Divino Ignácio. (2020) Modelo baseado em Frictionless Data aplicado aos dados abertos governamentais. Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, v. 18, n. 00, 20 nov. 2020. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8661528.>. Acesso em: 12 mai. 2021.

GIANFREDI, Vincenza et al. Vaccine Procurement: A Conceptual Framework Based on Literature Review. Vaccines 2021, [s. l.], v. 9, 3 dez. 2021. DOI https://doi.org/10.3390/vaccines9121434. Disponível em: https://www.mdpi.com/2076-393X/9/12/1434#cite. Acesso em: 1 jun. 2022.

JANSSEN, Marijn; CHARALABIDIS, Yannis; ZUIDERWIJK, Anneke. (2012) Benefits, adoption barriers and myths of open data and open government. Information Systems Management, v. 29, n. 4, p. 258–268, 2012.

JAYAWARDENE, Vimuthki; SADIQ, Shazia; INDULSKA, Marta. (2015) An analysis of data quality dimensions. ITEE Technical Report, Austrália, n. 2013-01, 2015. Disponível em: <https://espace.library.uq.edu.au/view/UQ:312314>. Acesso em: 9 dez. 2021.

JESILEVSKA, Svetlana. Data Quality Dimensions to Ensure Optimal Data Quality. The Romanian Economic Journal, Romênia, p. 89-103, 2017. Disponível em: http://www.rejournal.eu/article/data-quality- dimensions-ensure-optimal-data-quality. Acesso em: 16 mar. 2022.

Lima, Claudia Risso de Araujo et al. (2009) Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde. Cadernos de Saúde Pública [online]. 2009, v. 25, n. 10, pp. 2095-2109. Disponível em: <https://doi.org/10.1590/S0102-311X2009001000002>. Acesso em: 9 dez. 2021.

LOURENÇO, Rui Pedro. (2015) An analysis of open government portals: A perspective of transparency for accountability. Government Information Quarterly, sem local indicado, v. 32, n. 3, 14 jun. 2015. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0740624X15000660?via%3Dihub>. Acesso em: 12 mai. 2021.

MACEDO, Dirceu Flávio; LEMOS, Daniela Lucas da Silva. (2021) Dados abertos governamentais: iniciativas e desafios na abertura de dados no Brasil e outras esferas internacionais. AtoZ, Paraná, v. 10, n. 2, 25 jan. 2021. Disponível em: <https://revistas.ufpr.br/atoz/article/view/77737>. Acesso em: 12 mai. 2021.

Marsh, R. (2005). Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. Database Marketing & Customer Strategy Management, 12(2), 105-112.

OpenDATASUS. Disponível em: https://opendatasus.saude.gov.br. Acesso em: 16 mai. 2022.

OPEN KNOWLEDGE FOUNDATION. Manual dos dados abertos: governos. São Paulo: Laboratório Brasileiro de Cultura Digital; W3C Brasil; CGI.br, 2011. Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acesso em: 12 abr. 2022.

OPENGOVDATA. The annotated 8 principles of open government data. [on line]. 2007. Disponível em <http://opengovdata.org/>. Acesso em: 12 abr. 2022.

PICCOLO, Daiane. (2018) QUALIDADE DE DADOS DOS SISTEMAS DE INFORMAÇÃO DO DATASUS: Análise crítica da literatura. Ciência da informação em revista. Maringá, 2018. 7 p. Disponível em: https://www.seer.ufal.br/index.php/cir/article/view/5387. Acesso em: 1 dez. 2021.

POSSAMAI, Ana Júlia. (2016) DADOS ABERTOS NO GOVERNO FEDERAL BRASILEIRO: desafios de transparência e interoperabilidade. 2016. Tese (Doutora em Ciência Política.) - Universidade Federal do Rio Grande do Sul, [S. l.], 2016. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/156363/001015755.pdf?sequence=1. Acesso em: 12 maio 2021.

QUARATI, Alfonso; DE MARTINO, Monica. (2019) Open government data usage: a brief overview. In: INTERNATIONAL DATABASE APPLICATIONS & ENGINEERINGSYMPOSIUM, 23., 2019, Atenas, Grécia. Proceedings [...]. Atenas, Grécia: Ideas, 2019. p.1-8. Disponível em: https://doi.org/10.1145/3331076.3331115. Acesso em: 12 maio 2021.

REN, Guang-Jie; GLISSMANN, Susanne. (2012) Identifying information assets for open data: the role of business architecture and information quality. In: INTERNATIONAL CONFERENCE ON COMMERCE AND ENTERPRISE COMPUTING, 14., 2012, Washington, Dc. Proceedings [...]. Washington, Dc: Ieee Computer Society, 2012. p. 94-100.

SILVA, Marcela. (2020) ABERTURA DE DADOS GOVERNAMENTAIS: Estudo da implementação e desempenho da Política de Dados Abertos no Poder Executivo Federal. Orientador: Caio César de Medeiros Costa. 2020. Dissertação (Mestre em Administração, área de concentração: Administração Pública) - Universidade de Brasília - UNB, Brasilia, 2020. Disponível em: <https://repositorio.unb.br/handle/10482/38803>. Acesso em: 12 maio 2021.

ŠLIBAR, B.; OREŠKI, D.; BEGIČEVIĆ REĐEP (2021), N.Importance of the Open Data Assessment: An Insight Into the (Meta) Data Quality Dimensions. SAGE Open, Estados Unidos da América, v. 11, n. 2, 2021. Disponível em: <https://doi.org/10.1177/21582440211023178>. Acesso em: 9 dez. 2021.

TRUMBO, Silas et al. Improving immunization data quality in Peru and Mexico: Two case studies highlighting challenges and lessons learned. ELSEVIER, [S. l.], p. 7674?7681, 7 nov. 2018. DOI https://doi.org/10.1016/j.vaccine.2018.10.083. Disponível em: https://www.sciencedirect.com/science/article/pii/S0264410X18314701. Acesso em: 1 jun. 2022.

VETRÒ, Antonio et al. (2016) Open data quality measurement framework: Definition and application to Open Government Data. Government Information Quarterly, sem local indicado, v. 33, n. 2, 20 fev. 2016. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0740624X16300132#!>. Acesso em: 12 mai. 2021.

Wang, Richard Y., Strong, Diane M. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, Oxfordshire, Reino Unido, 1996, vol.12, 1996. Núm. 4, pág. 5 – 33. Disponível em: https://www.jstor.org/stable/40398176?origin=JSTOR-pdf. Acesso em: 16 mai. 2022.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). Disponível em: https://doi.org/10.1038/sdata.2016.18. Acesso em: 16 mai. 2022.

WWWF. Open Data Barometer: Fourth Edition. (2019). Disponível em: <http://www.opendatabarometer.org/4thEdition/report>.Acesso em:12 dez. 2021.