INDICADORES PARA AVALIAÇÃO QUALITATIVA DE DADOS ABERTOS: inteligibilidade, operacionalidade e interatividade nos *datasets* do Governo Federal no Portal Brasileiro de Dados Abertos

INDICATORS FOR OPEN DATA QUALITATIVE ASSESSMENT: intelligibility, operability and interactivity of Federal Government datasets on the Brazilian Open Data Portal

Sivaldo Pereira da Silva[1]
Ana Thereza Nogueira Soares[2]
Daniel Jorge Teixeira Cesar[3]
Leon Eugênio Monteiro Rabelo[4]

RESUMO
Este trabalho tem como principal objetivo contribuir na construção de metodologias para avaliação qualitativa de Dados Abertos governamentais, tendo como princípio averiguar se cumprem com os devidos parâmetros normativos. Neste sentido, o *corpus* analisado se concentrou nos *datasets* do governo federal brasileiro publicados no Portal Brasileiro de Dados Abertos. Foram estudados 715 *datasets* de todos os órgãos com *status* de ministério publicados no Portal, o que implicou na checagem de um total de 2.743 arquivos (conteúdos-bases), dos quais foram efetivamente encontrados e analisados 1.860 arquivos. Metodologicamente, o estudo foi executado em duas frentes: a análise qualitativa externa dos arquivos, observando suas condições e padrões de publicação; e análise qualitativa interna, verificando os arquivos em si, abrindo-os e observando-os individualmente. A metodologia proposta consiste na verificação de três dimensões normativas: (a) inteligibilidade; (b) operacionalidade e (c) interatividade. O estudo demonstrou que grande parte dos *datasets* tem problemas de inteligibilidade; detectou-se que 21% dos conjuntos de dados anunciados como "publicados" não possuem, de fato, arquivos disponíveis para *download* (*links* quebrados). Outros problemas também foram detectados, tais como lacunas de dados, falta de padronização e formatos, e insuficiência nos indicadores de interatividade.
Palavras-chave: Governo aberto. Dados Abertos. Transparência digital. Democracia Digital.

ABSTRACT
This study has as main objective to contribute to the construction of methodologies for qualitative evaluation of Government Open Data, having as analytical principle to verify if it complies with the appropriate normative parameters. In this sense, the corpus of analysis focused on the Brazilian federal government datasets published on the Brazilian Open Data Portal. A total of 715 datasets from all agencies with ministry status were studied, which involved checking a total of 2,743 files ('content-bases'), of which 1,860 files were effectively found and analyzed. Methodologically, the study was carried out on two fronts: an external qualitative analysis of the archives, observing their conditions and publication standards and an internal qualitative analysis, verifying the archives themselves, opening and observing them individually. The proposed methodology consists of checking three normative dimensions: (a) intelligibility; (b) operability and (c) interactivity. The study showed that most datasets have intelligibility problems; it was found that 21% of datasets advertised as "published" do not, in fact, have files available for download (broken links). Other problems were also detected, such as data gaps, lack of standardization and formats, and insufficient interactivity indicators.
Keywords: Open government. Open Data. Digital transparency. Digital Democracy.

Artigo submetido em 11/05/2020 e aceito para publicação em 03/06/2020

1 Professor do Programa de Pós-Graduação em Comunicação. Universidade de Brasília, Brasil. ORCID: https://orcid.org/0000-0001-8767-7679. E-mail: sivaldop@gmail.com
2 Doutoranda no Programa de Pós-Graduação em Comunicação. Universidade de Brasília, Brasil. ORCID: https://orcid.org/0000-0001-9973-0246. E-mail: anatsoares@gmail.com
3 Doutorando no Programa de Pós-Graduação em Comunicação. Universidade de Brasília, Brasil. ORCID: https://orcid.org/0000-0002-5786-5588. E-mail: danieljtc@gmail.com
4 Doutorando no Programa de Pós-Graduação em Comunicação. Universidade de Brasília, Brasil. ORCID: https://orcid.org/0000-0001-5188-2363. E-mail: leon.rabelo@gmail.com

1 INTRODUÇÃO

Com a importância e crescimento das ações e programas de Dados Abertos em todo o mundo, há hoje diversas iniciativas de avaliação dessas políticas, seja através de estudos acadêmicos mais específicos, seja através da publicação de índices comparativos internacionais sobre a performance de governos ao redor do globo (SUSHA *et al*, 2015, SAFAROV *et al*, 2017; ZUIDERWIJK & JANSEEN, 2014; HUIJBOOM & BROEK, 2011) Porém, esses instrumentos avaliativos ainda estão em fase inicial de implementação ou em processo de amadurecimento e consolidação (KUCERA *et al.*, 2015). Neste sentido, diversas metodologias, instrumentos e indicadores estão sendo testados e estruturados para tentar cobrir a diversidade de aspectos que este fenômeno abarca. A análise qualitativa das bases de dados sob a ótica de sua qualidade para a apropriação social são aspectos ainda carentes de maiores estudos.

Diante deste cenário, a questão central sobre a qual este artigo versará pode ser sintetizada na seguinte indagação: *Quais dimensões, indicadores e procedimentos de avaliação devem ser levados em conta na análise da qualidade dos recursos publicados em uma política de Dados Abertos?* Como pano de fundo, isto está vinculado ao horizonte de cumprimento dos parâmetros normativos que viabilizam ou dificultam a apropriação social de Dados Abertos. Neste sentido, o estudo se coloca como uma proposta que visa colaborar com os esforços metodológicos para análise qualitativa das informações (arquivos e metadados) que compõem uma boa política de Dados Abertos, podendo ser aplicável a qualquer instância governamental ou até mesmo a entes não-governamentais. Para testar tal proposta, o artigo traz um estudo de caso bastante representativo: foram analisados todos os *datasets* publicados por órgãos do Governo Federal com *status* de ministério no Portal Brasileiro de Dados Abertos.

Definido tal horizonte, o trabalho seguirá dividido em duas partes. Primeiramente, serão discutidas premissas teóricas que envolvem o conceito Dados Abertos apontando seus elementos normativos que servirão de base para os indicadores qualitativos. Na seção seguinte, o foco será sistematizar a aplicação de parâmetros e indicadores para a avaliação qualitativa de Dados Abertos, tendo como estudo de caso todos os conjuntos de dados dos dezoito ministérios do Governo Federal que possuíam *datasets* publicados no Portal Brasileiro de Dados Abertos no início de 2018.

2 DADOS ABERTOS: DIMENSÕES CONCEITUAIS E NORMATIVAS

Diversas definições e caracterizações de Dados Abertos têm sido publicadas em diferentes estudos, documentos de organizações civis, de organismos multilaterais ou em leis sobre o tema, geralmente

atreladas ao que vem sendo chamado de movimento de Governo Aberto (CHUNA, 2010; MCDERMOTT, 2010, MEIJER e HILLEBRANDT, 2012, JANSSEN *et al*, 2012, WIRTZA e BIRKMEYERA, 2015, SAFAROV *et al*, 2017; BRASIL, 2011, 2016; OECD, 2016, 2017, ONU, 2017). De forma prática, Dados Abertos significam recursos contendo informações que o Estado devolve para o cidadão na forma de bases de dados digitais - arquivos estruturados sobre os mais diversos temas de interesse público - para o livre reuso. Indivíduos, grupos de cidadãos, organizações, jornalistas, empresas, centros de pesquisa etc. podem se apropriar desses dados e produzir conhecimento ou aplicações úteis como estudos, reportagens, relatórios, *softwares* cívicos, APPs de serviço etc. (ATTARD *et al*, 2015, GURSTEIN, 2011, RUIJER, 2017).

Comum a qualquer concepção contemporânea de Dados Abertos estão três características importantes a serem destacadas. Primeiro, Dados Abertos não são apenas "dados abertos para o uso público" e sim "dados abertos para o uso público de forma automatizada, estruturada e dinamizada por vias digitais":

> É importante notar a distinção entre dados públicos e dados abertos. Enquanto os dados públicos são disponibilizados gratuitamente à população em geral, não são necessariamente abertos. Um arquivo contendo documentos legais é um exemplo extremo de dado público que não é aberto. Mesmo livremente acessíveis, imagine o esforço demandado para identificar e localizar um documento específico. Por outro lado, se tal dado é digitalizado e disponibilizado online de maneira padronizada (e indexada), então este dado público é também aberto (ATTARD *et al*, 2015, p. 04)[5]

Por isso, um livro em papel (que contém grande volume de dados) disponível para o público numa biblioteca não é considerado um "dado aberto", pois seu formato (analógico, não estruturado para o ambiente digital) viola boa parte dos princípios normativos que definem Dados Abertos. Assim, quando falamos de Dados Abertos estamos falando em formatos digitais estruturados.

A finalidade dos Dados Abertos é múltipla, isto é, não está vinvulada apenas a questões políticas específicas, mas também a propósitos de utilidade pública. Por isso, embora Dados Abertos sejam hoje potentes e importantes mecanismos para a transparência pública, o conceito não restringe a sua função apenas à "transparência". Dados Abertos podem ter diversas finalidades (incluindo transparência). Podem ser apropriadas pelos usuários na forma de um aplicativo, por exemplo, que ajude a melhorar o trânsito. Num outro exemplo, um cidadão pode desenvolver um aplicativo para acompanhar e fiscalizar os gastos públicos sobre trânsito na sua cidade. No exemplo 1, o objetivo não é transparência e sim um serviço de utilidade pública (melhoria do tráfego de carros na cidade). Já no exemplo 2, o objetivo trata de transparência, pois os dados são utilizados para aumentar a capacidade do cidadão de ver, de compreender, de acompanhar ações do Estado quanto ao gasto de recursos públicos.

5 Tradução própria do original em inglês.

Terceiro, na conceituação de Dados Abertos há o pressuposto de que são bens públicos e, por isso, o Estado tem a obrigação de publicá-los. Historicamente, isso tem seus precedentes no campo da epistemologia e depois ganhou impulso concreto em iniciativas como redes de estudo sobre fenômenos naturais globais (principalmente a partir dos anos de 1990), antes de migrar de forma mais robusta para o campo governamental (principalmente a partir dos anos 2000). Na Sociologia da Ciência, Robert Merton já chamava a atenção na década de 1940 para os benefícios da abertura do conhecimento científico (CHIGNARD, 2013, COELHO, 2017, SÁ e GRIECO, 2016), apontando a importância de se preservar e compartilhar dados comuns suscetíveis a livre troca, capazes de aumentar a sua apropriação e gerar novos conhecimentos:

> A concepção institucional de ciência como parte do domínio público se relaciona com a necessidade de comunicação de descobertas. O Sigilo é a antítese desta norma; comunicação total e aberta é a prática. A pressão para difundir resultados é reforçada pelo objetivo institucional de avançar as fronteiras do conhecimento e pelo incentivo ao reconhecimento que é, claramente, contingente à publicação. Um cientista que não comunica suas importantes descobertas à comunidade científica - portanto, um Henry Cavendish - torna-se alvo de respostas ambivalentes [...] A característica comum das ciências é refletida pelo reconhecimento dos cientistas sobre sua dependência em relação à herança cultural sobre a qual realizam diferentes afirmações. (MERTON, 1973, p. 274) [6]

Abertura significava estabelecer parâmetros comuns razoáveis para compartilhar informação capazes de descentralizar análises e gerar avanços em termos de produção de conhecimento. Isso alcança os governos pelo fato da máquina estatal ser historicamente uma fábrica de produção de dados. Em convergência com as potencialidades das tecnologias digitais e às demandas por mais transparência governamental, surgem o que chamamos hoje de políticas de Dados Abertos.

Se observarmos mais atentamente essas características discutidas até aqui, na forma de premissas, é possível perceber que a política de Dados Abertos é, por natureza, normativa. Traz em seu bojo deontologias oriundas de diversas áreas como Teoria Política, Ciências da Computação; Ciência de Dados; Administração Pública e Comunicação Social, além da influência das demandas provenientes de atores como organizações civis, *hacktivistas*, jornalistas e pesquisadores.

A apropriação social de Dados Abertos é uma condição inerente ao próprio conceito de Dados Abertos. Estar aberto implica em *estar acessível* (no sentido mais amplo do termo) ao público, uma diretriz que permeia todo o conceito de Governo Aberto e as políticas de Dados Abertos. É preciso que os dados "conversem" com os seus usuários estabelecendo uma relação baseada em informação, comunicação e interação. As bases de dados abertas não são apenas informações públicas lançadas

6 Tradução própria do original em inglês.

online e lá inertes: são informações públicas que demandam estruturações, explicações e integração com o usuário para serem efetivamente apropriadas. Não basta apenas o livre reuso (no sentido solitário do termo) e sim, o livre, efetivo e dinâmico reuso. Para isso é preciso criar pontes neste relacionamento entre bases de dados e seus usuários (finais e intermediários).

Por estar normativamente guiada por estes princípios, a política de Dados Abertos é, naturalmente, objeto de avaliações. Surgiram assim diversas abordagens que buscaram medir o desempenho de bons indicadores neste campo, geralmente na forma de índices comparativos, com algum tipo de classificação (*ranking*) entre as políticas de Dados Abertos. Podemos citar quatro índices hoje vigentes e conhecidos publicados por diversas organizações: Open Data Barometer (ODB), Global Open Data Index (GODI), Open Data Inventory (ODIN) e OurData Index (Open-Useful-Reusable Data Index). No geral, os índices trabalham com indicadores que tentam avaliar questões como: os temas sobre os quais os dados versam (Saúde, Segurança, Finanças governamentais, poluição etc.); os parâmetros técnicos gerais que garantem a abertura dos dados para leitura automática por máquinas (se estão em arquivos com formatos condizentes); se não existem barreiras para o livre reuso (licenças livres; não restrições ao acesso); se os países possuem ações concretas de incentivo à apropriação social de Dados Abertos; se possuem legislação condizente; se produzem dados com qualidade para o livre reuso e, em alguns casos, se desenvolvem ações de participação.

Embora os índices sejam relevantes e cumpram seu papel de pressão por políticas mais consistentes de Governo e Dados Abertos, do ponto de vista metodológico, importante ressaltar que há hoje uma lacuna comum nessas avaliações: o baixo aprofundamento em indicadores que tratem da qualidade dos *datasets* observando-os "por dentro", isto é, analisando a viabilidade prática das bases de dados do ponto de vista do usuário. Há uma carência de avaliações mais qualitativas que se aproximem das reais condições de uso para a apropriação social de Dados Abertos.

Análises qualitativas mais detalhadas da funcionalidade das bases de dados em si têm sido realizadas de modo ainda incipiente no âmbito acadêmico, principalmente na forma de artigos (BATINI *et al* 2009, ZUIDERWIJK e JANSSEN, 2015, OSAGIE, 2017, MÁCHOVÁ e LNĚNIČKA, 2017) e parcialmente em alguns dos índices anteriormente citados. Persiste, assim, uma carência e ao mesmo tempo e uma tendência: compreender a qualidade dos dados para a sua potencial apropriação é hoje um elemento fundamental em qualquer avaliação de política de Dados Abertos; Para tanto, é preciso levar em conta os elementos normativos que definem boas práticas e, ao mesmo tempo, observar a qualidade dos recursos (arquivos) sob o ponto de vista da sua usabilidade.

2.1 Três pilares normativos *versus* duas dimensões de análise

Importante observar que uma boa política de Dados Abertos não consiste apenas na disposição de arquivos *online*. Como foi colocado anteriormente, a apropriação social dos dados publicados é um elemento fundamental para que o ciclo da política de Dados Abertos se concretize. A simples publicação de conteúdo em uma página na internet está muito aquém das exigências que compõem o conceito de Dados Abertos. A ideia de Dados Abertos pressupõe um arranjo que prevê, incentiva e possibilita a apropriação pelo usuário. Isso passa pela publicação ordenada e qualificada de conteúdo, visando o seu reuso automatizado por algoritmos ou *softwares*. Para que isso ocorra, os arquivos precisam ter elementos internos devidamente qualificados e, ao mesmo tempo, elementos externos que possibilitam sua compreensão, contextualização para o efetivo reuso.

Indo nesta direção, Silva e colegas (2020) propõem que esta análise qualitativa seja mais ampla, apontando assim duas dimensões de análise que envolvem a publicação dos *datasets*:

> [...] (a) análise qualitativa externa [...] recai sobre as informações que estão no entorno da publicação (metadados, significados, interação etc.) e (b) análise qualitativa interna [...] se concentra na estrutura da publicação (sobretudo na operacionalidade e adequação técnica dos arquivos em sua estrutura interna) (SILVA et al 2020, p. 7)

Essas duas dimensões (ou frentes de análise) são úteis pois dizem respeito ao arranjo sobre o qual uma boa política de publicação de Dados Abertos está assentada, o que envolve tanto elementos pertinentes aos arquivos em si quanto à sua publicação em plataformas ou *websites*. Do ponto de vista metodológico, esse conjunto de indicadores está baseado em princípios, hoje amplamente aceitos, que delimitam o conceito Governo aberto e Dados Abertos:

> Todos esses indicadores [...] foram construídos levando em conta os aspectos normativos que estão nas bases da literatura sobre dados abertos e da observação empírica acerca dos problemas que podem facilitar ou dificultar a apropriação do ponto de vista do usuário. Neste sentido, uma boa síntese da base normativas nas quais os indicadores estão apoiados são os 8 princípios que os dados abertos devem seguir, na perspectiva da ideia de Governo Aberto. Os dados devem ser: (1) completos; (2) primários; (3) atuais; (4) acessíveis; (5) processáveis por máquina; (6) com acesso não discriminatório; (7) com formatos não proprietários; (8) livres de licenças. Outros princípios também foram posteriormente adicionados a esta lista: (9) *online* e gratuitos; (10) permanentes; (11) confiáveis; (12) com presunção de abertura; (13) documentados; (14) seguros para abrir; e (15) projetado com participação pública. Os indicadores utilizados no presente estudo atravessam boa parte desses princípios (SILVA et al 2020, p. 10)

Deste modo, os indicadores que constituem as duas dimensões de análise devem estar integrados e percebê-los (como duas faces de uma mesma moeda) evita que a avaliação seja incompleta[7].

7 Por exemplo, um governo pode publicar *datasets* de Dados abertos com alto grau de qualidade técnica interna dos arquivos (bem estruturados, sem lacuna de dados, com dados padronizados, em formatos adequados etc.) mas com péssimo entorno de publi-

Adicionalmente, para avançarmos com maior clareza e objetividade em direção a uma metodologia integrada e robusta de análise qualitativa de Dados Abertos, propomos que tais dimensões analíticas devem ser pensadas como indicadores que sustentam três pilares normativos: (a) inteligibilidade, (b) operacionalidade e (b) interatividade.

A *inteligibilidade* consiste na propriedade dos recursos em serem compreensíveis onde o "aberto" pressupõe que o dado deve ser acessível cognitivamente para o usuário se aproximando da noção de transparência, ou seja, aberto para o seu entendimento e reuso. Mesmo que um conjunto de informações seja publicado em arquivos estruturados seguindo todos os requisitos técnicos exigidos, se o usuário não tiver informações que proporcionem a apreensão do significado do conteúdo isso torna a publicação inútil do ponto de vista prático, pois inviabiliza a sua real apropriação[8].

A *operacionalidade* diz respeito à propriedade dos recursos em termos de funcionalidade no seu processo de apropriação, livrando o usuário de barreiras, constrangimentos ou custos desnecessários de reuso. Em termos práticos, significa analisar o quanto um recurso (arquivo) está operacional para a livre apropriação. Isso inclui a avaliação de existência ou não de *links* "quebrados" que impossibilitem o acesso real aos arquivos; se os arquivos, uma vez baixados, não estão corrompidos, incompletos ou com erros de estruturação; se foram publicados em formatos adequados etc.

Por fim, a *interatividade* se refere à dimensão mais comunicacional que envolve os recursos. Significa avaliar o quanto um *dataset* pode ser objeto de diálogo, de *feedback* informativo, interação social ou discussão pública. Isso envolve elementos como a existência e funcionamento de canais de comunicação do agente público para esclarecer dúvidas do usuário; a existência de espaços de discussão sobre os *datasets*; a efetividade de *feedback* informativo da curadoria dos dados quando acionada etc.

Dito isto, propõe-se uma metodologia de avaliação de Dados Abertos que leve em conta as duas dimensões de análise (interna e externa) e que sejam ao mesmo tempo capazes de responder aos três pilares normativos (inteligibilidade, operacionalidade e interatividade) que estão nas bases deontológicas de uma boa política de Dados abertos. O Quadro 1 traz uma visualização dos principais indicadores, seus significados e como se posicionam neste sistema de avaliação proposto:

cação (sem metadados, sem informações sobre a atualização das publicações, sem informações sobre os seus significados, sem cartilha ou manual para apropriação ou sem qualquer outro dado relevante que ampliem os significados para reuso dos arquivos). Neste caso hipotético (mas não improvável) a política de Dados abertos estaria deficiente.

8 Por exemplo, se uma variável que compõem a coluna de uma planilha for um código ou palavra sem significado intuitivo, todos os dados desta coluna não podem ser utilizados com segurança; ou se a unidade de medida utilizada em determinada informação é vaga, não é possível compreender o que significa os números a que esta se refere etc.

Quadro 1: Indicadores, dimensões de análise e pilares normativos

	Indicadores qualitativos externos	
	Indicador	**O que identifica ou verifica?**
INTELIBILIDADE	Temporalidade	Identifica a temporalidade do *dataset*; data de criação o tempo-base de atualização dos dados. Este indicador repercute na inteligibilidade do *dataset* pois arquivos publicados sem informações sobre data de criação ou atualização afeta a capacidade cognitiva do usuário em contextualizá-lo historicamente.
	Parametrização	Verifica se o *dataset* possui alguma informação auto-instrutiva capaz de explicar seus parâmetros viabilizando assim o seu uso na prática. Isto inclui explicação sobre suas variáveis, metodologias ou outras informações técnicas que vão além da rotulação. Exemplos: manual de dados, cartilha ou arquivo similar com esta função.
	Rotulação	Verifica o nível de inteligibilidade do *dataset* em termos de clareza sobre seu conteúdo, isto é, identifica se possui um rótulo; algo breve, porém capaz de servir como uma etiqueta informativa de entrada que traga detalhes suficientes para o usuário antever o que o arquivo contém sem precisar abri-lo.
	Visualização	Verifica se o *dataset* possui algum tipo de recurso que possibilite ao usuário ter uma visão panorâmica dos dados sem necessariamente necessitar abrir e tratar arquivos. Essa visualização pode aparecer na forma de tabelas em HTML, infográficos, mapas e outros recursos similares. Importante lembrar que a visualização deve ser compreendida como uma forma suplementar de inteligibilidade na qual todo dado correspondente também esteja paralelamente disponibilizado em formatos estruturados para leitura de máquina.
INTERATIVIDADE	Discursividade	Verifica se o *dataset*/recurso é apoiado por algum fórum ou espaço público para discussão e compartilhamento de experiências e usos. Neste indicador, para que seja compreendido como parte da política de Dados Abertos é necessário que a iniciativa esteja fomentada pelo Estado (ou, pelo menos, claramente apoiada por organismos estatais).
	Comunicação	Verifica se o *dataset*/recurso possui algum mecanismo ativo de *feedback* informativo entre a curadoria e o usuário e a efetividade de resposta em caso de demanda específica do usuário.
	Indicadores qualitativos internos	
OPERACIONALIDADE	Disponibilidade	Verifica se o recurso possui algum erro que inviabilize a sua acessibilidade como *link* quebrado; arquivo corrompido, isto é, se o arquivo de fato está disponível para *download* de modo concreto.
	Formatação	Identifica se os recursos estão publicados em formatos estruturados legíveis por máquinas; que tipos de formatos; se são formatos não-proprietários que garantam a devida abertura para a apropriação social sem restrições de uso.
	Padronização	Verifica se o conteúdo do *dataset* apresenta qualidade em termos de padrão quanto à morfologia das informações, homogeneidade quanto às nomenclaturas evitando grafias divergentes para uma mesma informação; verifica se o arquivo não possui "sujeiras de, dados" (isto é, resquícios de informação que atrapalha a leitura padronizada de informação); trechos ilegíveis ou erro de estruturação que compromete o padrão esperado de leitura por máquinas.
	Integridade	Verifica se o recurso respeita o princípio da completude dos dados, isto é, se não apresenta lacuna (*missing*) de informação em suas células internas.
	Opacidade	Verifica se o arquivo possui variáveis opacas, isto é, colunas devidamente nomeadas e transparentes quanto ao seu real significado para que o usuário possa fazer o correto uso dos dados a que se refere.

Fonte: adaptado e complementado pelos autores, a partir da proposta inicial de Silva et al 2020

Nota-se no quadro que a inteligibilidade e interatividade aparecem circunscritas à análise qualitativa externa dos arquivos enquanto o pilar da operacionalidade se posiciona na análise interna dos arquivos. Esta vinculação é uma tendência majoritária que caracteriza a atual conjuntura de publicação dos dados, mas não deve ser compreendida necessariamente como rígida nestes termos. Por exemplo, é possível que elementos de interatividade e inteligibilidade sejam inseridos no interior dos arquivos através de codificações suplementares capazes de serem lidas por algoritmos/*softwares* (como notações em RDF, no caso de dados conectados). De todo modo, esses elementos embora estejam no interior dos arquivos podem ser considerados como metadados ou dispositivos que dizem respeito ao entorno do conteúdo e não ao conteúdo do arquivo em si.

3 PROCEDIMENTOS METODOLÓGICOS

O principal objetivo deste artigo é avançar no desenvolvimento de metodologia de análise qualitativa dos recursos de uma boa política de Dados Abertos e demonstrar sua aplicabilidade em um estudo de caso representativo. Nesta seção, os principais elementos que compõem esta proposição metodológica serão detalhados bem como os procedimentos de coleta e análise.

3.1 *Dataset* x recursos

Primeiramente, convém especificar o que está sendo chamado neste trabalho de "*dataset*" e "recurso" pois serão dois termos bastante citados na avaliação. Um *dataset* é um "conjunto de dados" que estão relacionados a uma mesma temática. Por exemplo, *dataset* sobre indicadores de mortalidade infantil que traz diversas informações sobre esse tema. Dentro dos *datasets* existem os recursos. Um recurso é uma unidade de arquivo que compõe um *dataset*. Assim, um *dataset* (ou conjunto de dados) pode ser composto por vários recursos (arquivos) que, por sua vez, podem ser publicados em diversos tipos de formatos (como HTML, CSV, XML, JSON, PDF etc). A proposta de análise estabelecida neste artigo avalia tanto o conjunto de dados (*dataset*) e suas informações gerais quanto a especificidade de seus recursos.

3.2 Coleta de dados e metadados

Na aplicação da planilha e seus indicadores, o procedimento de coleta de dados foi dividido em duas partes. Na Parte I - referente à análise qualitativa externa - o foco foi a captação de meta-

informações sobre os *datasets*, consideradas relevantes para cruzar com indicadores qualitativos. Ou seja, analisou-se primeiramente as informações acerca dos arquivos (data de postagem; agentes curadores; se o *dataset* possui rotulação explicando seu conteúdo; se possui manual ou cartilha para o reuso etc.) e não os arquivos em si. Já na Parte II, a análise incide especificamente sobre os arquivos (recursos) que compõem o *dataset* (análise qualitativa interna).

Neste estudo de caso, a coleta referente à primeira parte da Planilha se deu basicamente a partir da coleta manual de informações publicadas no Portal Brasileiro de Dados Abertos, especificamente nas páginas dedicadas aos *datasets*. Já na segunda parte o procedimento metodológico consistiu na abertura dos arquivos para analisá-los por dentro, testando-os, observando suas estruturas, formatos e buscando identificar possíveis erros ou disfunções, seguindo a ideia de "conteúdo-base" proposta por Silva e colegas (2020), como veremos a seguir.

3.3 Conteúdo-base

Tendo em vista que um *dataset* pode ser constituído por diversos arquivos (recursos), por uma questão de recorte, adotou-se o método de análise do que chamamos de "conteúdo-base" ou "recurso-base" dos *datasets*: consiste em um pacote de informação específica que pode aparecer - a mesma informação - em diferentes tipos arquivos (SILVA *et al*, 2020). Por exemplo, um *dataset* pode ter cinco recursos (arquivos) publicados respectivamente nos formatos CSV, XML, JSON, PDF e TXT porém, imaginemos que cada um desses arquivos tenha o mesmo conteúdo (exemplo: a lista com o gasto em Recursos Humanos na área de Saúde de um Ministério). Neste caso, o *dataset* possui cinco recursos totais, porém apenas um conteúdo-base publicado (ou seja, um conteúdo-base publicado em cinco formatos diferentes).

Feita esta definição, tendo em vista que a ideia de Dados Abertos exige a existência de arquivos estruturados[9] para leitura automática a ênfase analítica recairá naturalmente sobre este tipo de arquivo. Os arquivos "fechados" (isto é, não estruturados, como PDF, DOCX, JPG etc), por não serem considerados Dados Abertos, foram apenas catalogados a título de registro (não foram nem poderiam ser analisados por dentro). Como um *dataset* pode ter vários tipos de arquivos estruturados (através do qual o mesmo

9 Arquivos não-estruturados são recursos publicados sem uma estrutura adequada para leitura automática de algoritmos, isto é, leitura automática por máquinas (exemplo: uma página de HTML, um PDF, um DOCX, um ODT, uma imagem como JPG... são arquivos digitais mas não estão estruturados para leitura automática (mesmo que um PDF ou um arquivo de Word traga uma tabela ela não é legível de modo automático por um algoritmo) . Quando chamamos de "estruturado" significa que é um arquivo legível por *softwares* de bases de dado (Exemplo: CSV, JSON, XLS, XML etc.) e quando chamamos de não-proprietários significa que não são de propriedade privada de uma empresa, são de livre uso (Exemplo, o formato XLXS é de propriedade da Microsoft; já o formato CSV é código livre pois não possui proprietário.

conteúdo é publicado) e como seria pouco útil analisar o mesmo conteúdo em diferentes arquivos estruturados, o estudo tomou o seguinte procedimento proposto por Silva e colegas (2020): (a) analisou apenas um dos arquivos estruturados de um mesmo conteúdo-base; (b), criou uma hierarquia de escolha desses arquivos na qual se tomou como prioridade a análise do arquivo no formato CSV como arquivo-padrão[10]; (c) caso o conteúdo-base não tenha sido publicado em formato CSV, a escolha do arquivo seguiu a seguinte hierarquia: JSON e XML. Caso não haja nenhum dos formatos previstos, a análise recairá no arquivo estruturado existente tomando em ordem alfabética (ordenado pelo nome da extensão) [11].

Do ponto-de-vista metodológico, a análise de um arquivo estruturado para cada conteúdo-base é seguro pois (a) conceitualmente, apenas arquivos estruturados podem ser considerados compatíveis com a concepção de Dados Abertos e relevantes para análise (pois somente formatos legíveis por máquinas cumprem um dos requisitos da concepção de Dados Abertos) e; (b) tecnicamente, a publicação de um mesmo conteúdo em diversos formatos tem como base um arquivo original isto é, um arquivo inicial é convertido pelo curador em outros formatos. Neste caso, analisar dois ou mais arquivos em formatos diferentes referentes a um mesmo conteúdo-base significa encontrar os mesmos erros replicados, o que seria redundante e contraproducente, não acrescentando nova informação para além da análise de apenas um arquivo que por si só já é representativo dos demais.

4 ESTUDO DE CASO: ANÁLISE QUALITATIVA DOS *DATASETS* NO PORTAL DADOS.GOV.BR

Em linhas gerais, a síntese metodológica proposta possibilita termos uma visão geral tanto descritiva quanto ao contexto de publicação de uma política de Dados Abertos quanto uma visão mais específica sobre a qualidade dos arquivos. Isso se torna hoje imprescindível pois análises superficiais sobre tal política pode apenas esconder seus reais problemas e não possibilitam o desenvolvimento de indicadores capazes de acompanhar a sua evolução (ou involução). Nesta seção, traremos os dados da aplicação destas diretrizes metodológicas em um estudo de caso bastante representativo, avaliando

10 O arquivo CSV é considerado por diversas normatizações, guias e organizações como W3C um formato-base no qual o dado deve ser disponibilizado por ser um formato leve, simples, universal e não proprietário. Isso não significa que o *dataset* precisa ter apenas formatos CSV: significa que este é considerado um formato-base exigido em qualquer *dataset*, ainda que ela seja enriquecido com outros formatos mais avançados como por exemplo arquivos com notações RDF para websemântica.

11 Como procedimento-padrão adotado foram analisados todos os conteúdos-bases que compõe os *datasets* no limite de 15 recursos (a partir do 16º recurso, por ordem listada no portal, a análise dos arquivos subsequentes deixou de ser realizada). Isso foi necessário devido ao grande volume de arquivos em alguns poucos *datasets* tipificados como "*outliers*". Considerou-se o método seguro e representativo pois 95,5% dos *datasets* estudados possuem até 15 recursos. Ou seja, em apenas 5,5% dos *datasets* a análise dos recursos não foi total (ainda assim, foram analisados pelo menos 15 recursos desses conjuntos de dados maiores).

todos os órgãos ministeriais do governo federal brasileiro e seus recursos publicados no Portal de Dados Abertos, principal instrumento da política brasileira de Dados Abertos.

A análise proposta neste artigo recaiu sobre todos os *datasets* dos órgãos do Poder Executivo (administração direta) com *status* de ministério publicados no portal de Dados Abertos (www.dados.gov.br) até o dia 23 de janeiro de 2018. Embora entre 2018 e a data de publicação deste artigo o país tenha passado por uma série de mudanças administrativas (com fusões e extinção de ministérios) os dados continuam relevantes, tanto do ponto de vista da aplicabilidade da metodologia proposta (que pode ser reproduzida em avaliações longitudinais) quanto na forma de registro histórico dos problemas que envolvem a publicação de recursos no Portal Brasileiro de Dados Abertos, uma vez que boa parte das questões apresentadas continuam as mesmas, embora sob a tutela de outros ministérios. Nestes termos, foram analisados 715 *datasets* de 18 órgãos com *status* ministerial (ver Quadro 1 e Gráfico 1), que apontavam para 2.743 arquivos (conteúdos-bases), sendo a análise específica interna de 1.860 arquivos de fato disponíveis.

Quadro 1 – Volume de *datasets* e recursos por Ministério (em ordem alfabética)

Ministério (ou órgão equivalente, em ordem alfabética)	**Quant. de *datasets***
Ministério da Agricultura, Pecuária e Abastecimento	9
Ministério da Ciência, Tecnologia, Inovações e Comunicações	19
Ministério da Cultura	15
Ministério da Defesa	17
Ministério da Educação	49
Ministério da Fazenda	117
Ministério da Indústria, Comércio Exterior e Serviços	22
Ministério da Justiça	44
Ministério da Saúde	136
Ministério da Transparência e Controladoria-Geral da União	19
Ministério das Relações Exteriores	6
Ministério do Desenvolvimento Social	36
Ministério do Esporte	5
Ministério do Planejamento, Desenvolvimento e Gestão	76
Ministério do Trabalho	6
Ministério do Turismo	31
Ministério dos Transportes	28
Secretaria de Governo	80

Fonte: a partir das publicações em Dados.gov.br com elaboração própria (N = 715)

No gráfico 1 temos esta lista ordenada por volume. Até a data da coleta, os dois ministérios com maior volume de *datasets* publicados foram o Ministério da Saúde (com 136) e Ministério da Fazenda (com 117). Se somarmos a estes o terceiro e quarto colocados (respectivamente Secretaria de Governo, com 80 *datasets* e Ministério do Planejamento com 76) temos um grupo de quatro ministérios se destacam com um volume de *datasets* representando 65% do total.

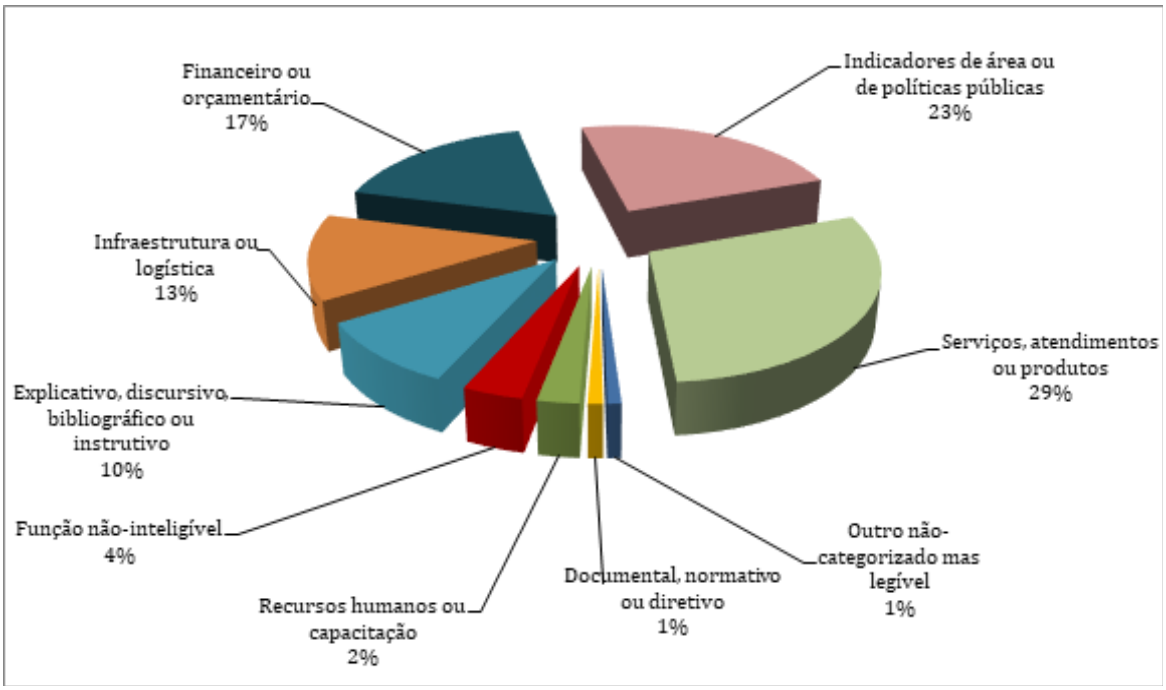
Gráfico 1 – Volume de *datasets* e recursos por Ministério

Fonte: a partir das publicações em Dados.gov.br com elaboração própria (N = 715)

Por outro lado, quatro ministérios tiveram menor volume publicações: Ministério da Agricultura, Pecuária e Abastecimento (com 9); Ministério do Trabalho (6); Ministério das Relações Exteriores (6) e Ministério do Esporte (5). Embora possa ser plausível que alguns ministérios tenham menor volume de dados a serem disponibilizados, o volume muito ínfimo desses quatro ministérios não pode ser justificado pela sua natureza e sim por haver uma postura menos ativa no cumprimento das diretrizes de uma boa política de Dados Abertos.

Ainda numa análise mais descritiva e exploratória, também foram identificados os tipos de ênfases temáticas que os recursos (arquivos) mais tratavam. No gráfico da Gráfico 2 temos um panorama desta distribuição:

Gráfico 2 – Abordagem temática dos *datasets* (por ocorrência de categorização)



Fonte: elaboração própria (N = 3483)[12]

Os dados dizem respeito ao total de abordagens temáticas que cada recurso (arquivo) se refere, sendo que um mesmo arquivo poderia ter mais de uma abordagem simultaneamente[13]. Deste modo, os números demonstram que três categorias temáticas foram as mais detectadas nos arquivos: (1) Serviços, atendimento ou produtos; (2) Indicadores de área ou políticas públicas e (3) Informações financeiras ou orçamentárias.

Após esse quadro geral sobre o *corpus* estudado, iremos agora tratar mais especificamente das duas frentes de análises (externa e interna) e buscaremos avaliar, em paralelo, a desenvoltura dos pilares normativos que atravessam tais dimensões.

4.1 Análise qualitativa externa

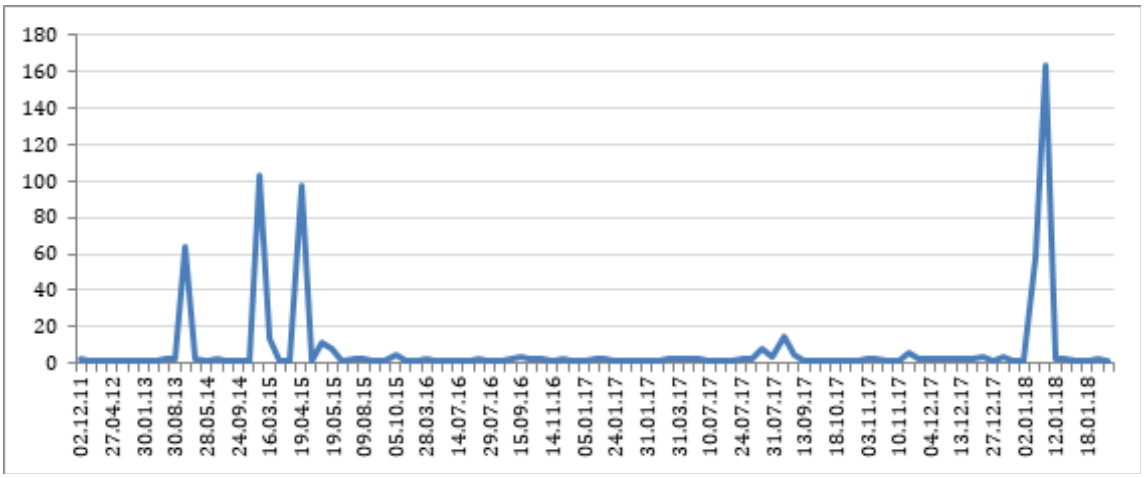
Um primeiro indicador da análise qualitativa externa diz respeito à temporalidade. Há basicamente duas formas de se analisar a temporalidade de um *dataset*: o *timing* da sua criação e o ritmo de

12 Aqui trata-se não do total de *datasets* mas sim de cada arquivo que compõe o *dataset*, por isso o N é bem mais elevado que 715. Adicionalmente, convém explicar que N neste gráfico trata do total de categorizações temáticas identificadas nos arquivos, lembrando que um mesmo arquivo pode ter mais de uma ênfase temática.

13 Por exemplo, um arquivo que trazia o custo financeiro de determinado serviço público e o detalhamento da oferta e gasto com esse serviço por região tem uma abordagem temática que trata simultaneamente de finanças e serviços, recebendo assim duas categorias de abordagens temáticas.

sua atualização. Com base no registro de criação dos *datasets* o Gráfico 3 traz uma visualização da cronologia da publicação no portal de Dados Abertos:

Gráfico 3: Data de criação dos *datasets* conforme registro no Portal de Dados Abertos



Fonte: elaboração própria (N = 703[14])

Nota-se, pelo próprio desenho do gráfico, que a publicação dos conjuntos de dados se dá na forma de "ondas agudas" e não em pequenas dosagens distribuídas ao longo do tempo. Nesta amostragem, grande parte dos dados foi publicada em quatro dias durante os últimos quatro anos, respectivamente: 08 de janeiro de 2014; 05 de novembro de 2014; 19 de abril de 2015 e 08 de janeiro de 2018 (esses últimos com o maior volume publicado). Esses picos de publicação ocorrem justamente por duas razões principais: (a) primeiramente, por ainda estarmos num período inicial de publicação dos primeiros volumes de Dados Abertos, isso faz com que haja um "acumulado" histórico de dados sendo publicado de uma só vez, incluindo a preparação e conversão dos dados já existentes para os formatos adequados; (b) segundo, isso também é impulsionado pelos prazos aos quais os órgãos estão submetidos em seus PDAs (Plano de Dados Abertos), que foi um instrumento bastante utilizado para pressionar as publicações nesta fase inicial da construção de uma política de Dados Abertos. Hipoteticamente, a tendência com o tempo é que esta política amadureça e se crie rotinas nos órgãos possibilitando que as publicações ocorram de maneira mais espraiada e cotidiana e com menos volumes de dados represados a serem publicados.

14 O volume de *datasets* listados para análise foi 715, porém, o N neste gráfico tem volume um pouco menor (especificamente 703) pois 12 *datasets* não possuíam informação sobre data de criação no momento da coleta.

Sobre a atualização, a Tabela 1 traz uma visão geral do ritmo de renovação (ou alteração) dos *datasets*, tomando como referência o dia no qual cada *dataset* foi analisado, contabilizando a partir daí o tempo transcorrido desde a última atualização registrada:

Tabela 1 – Registros de atualização do *dataset*

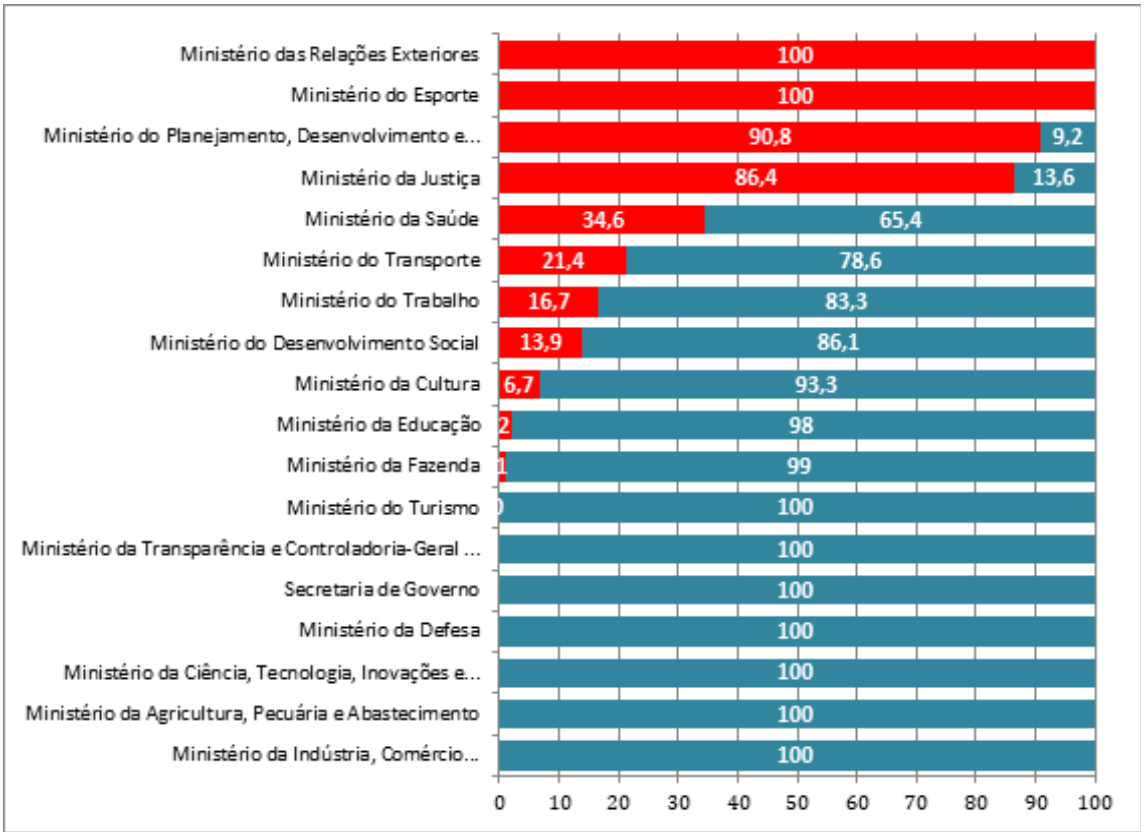
Tempo transcorrido desde a última atualização	%
De 1 a 6 dias atrás	7,1
De 7 a 15 dias atrás	3,2
De 16 a 29 dias atrás	3,8
De 30 a 39 dias atrás	5,1
De 40 a 59 dias atrás	5,5
De 60 a 89 dias atrás	4,4
Entre 3 meses e 5 meses	12,0
De 6 a 11 meses	10,1
1 ano ou acima	34,7
Sem atualização ou não identificada	14,1

Fonte: Elaboração própria (N = 715)

Primeiramente, a tabela demonstra que 14,1 % dos conjuntos de dados não possuem informação sobre última atualização. Além disso, o ritmo de atualização preponderante no Portal é acima de três meses (em 56% dos *datasets*). Atualizações semanais são escassas (menos de 10%). Embora este quadro já aponte que há uma deficiência no indicador de temporalidade quanto à atualização, é preciso aprofundar pesquisas e ponderar que tipos de conteúdos de fato requerem atualização mais curta e que tipos deve-se considerar normal que tenha prazos mais longos de *upgrade* devido à própria natureza do dado que eventualmente exige mais tempo para ser consolidado, a ponto de não ser considerada uma deficiência.

Um segundo indicador externo aos arquivos, que também diz respeito à inteligibilidade, refere-se à rotulação. Os resultados demonstraram que 73% dos *datasets* possuíam algum tipo de rotulação válida capaz dar ao usuário uma síntese acerca do que se trata. Este índice é razoável, pois demonstra que boa parte dos *datasets* são devidamente rotulados. Porém, é preciso não esquecer que há 1/3 (um terço) de arquivos que não possuem uma auto-explicação, sustentando assim, à primeira vista, pouca inteligibilidade no nível inicial de contato com o usuário. E também devemos ponderar, ao observarmos caso a caso, que há ministérios cujo desempenho neste indicador está bem abaixo da média geral, conforme aponta o Gráfico 4:

Gráfico 4 - Existência de rotulação nos *datasets* por Ministério
(em termos percentuais quanto ao volume específico de cada órgão)

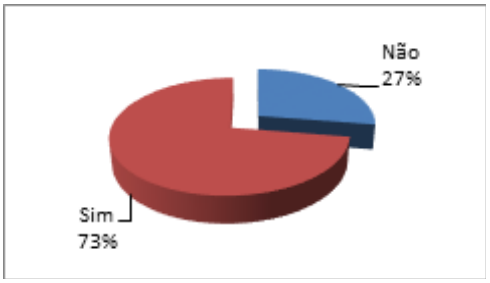


Fonte: elaboração própria (N = 715)

Como demonstra o gráfico 4, o Ministério das Relações Exteriores (Itamaraty) e Ministério do Esporte, além de terem os menores volumes de recursos publicados, sustentam também pior desempenho no indicador de rotulação (nenhum de seus poucos *datasets* possui rotulação). Outros dois ministérios também apontam índices bem baixos neste indicador, como o Ministério do Planejamento (com 90,8 % dos *datasets* sem rotulação) e o Ministério da Justiça (com 86,4 %). Por outro lado, há órgãos com praticamente todos os *datasets* devidamente rotulados como o Ministério da Educação, Ministério da Fazenda, Ministério do Turismo, Ministério da Transparência, Secretaria de Governo, Ministério da Defesa, Ministério da Ciência, Tecnologia, Inovações e Comunicações e Ministério da Agricultura.

Gráfico 5 - Comparativo sobre a existência de rotulação ou parametrização nos *datasets*

Possuem Rotulação



(N = 715)

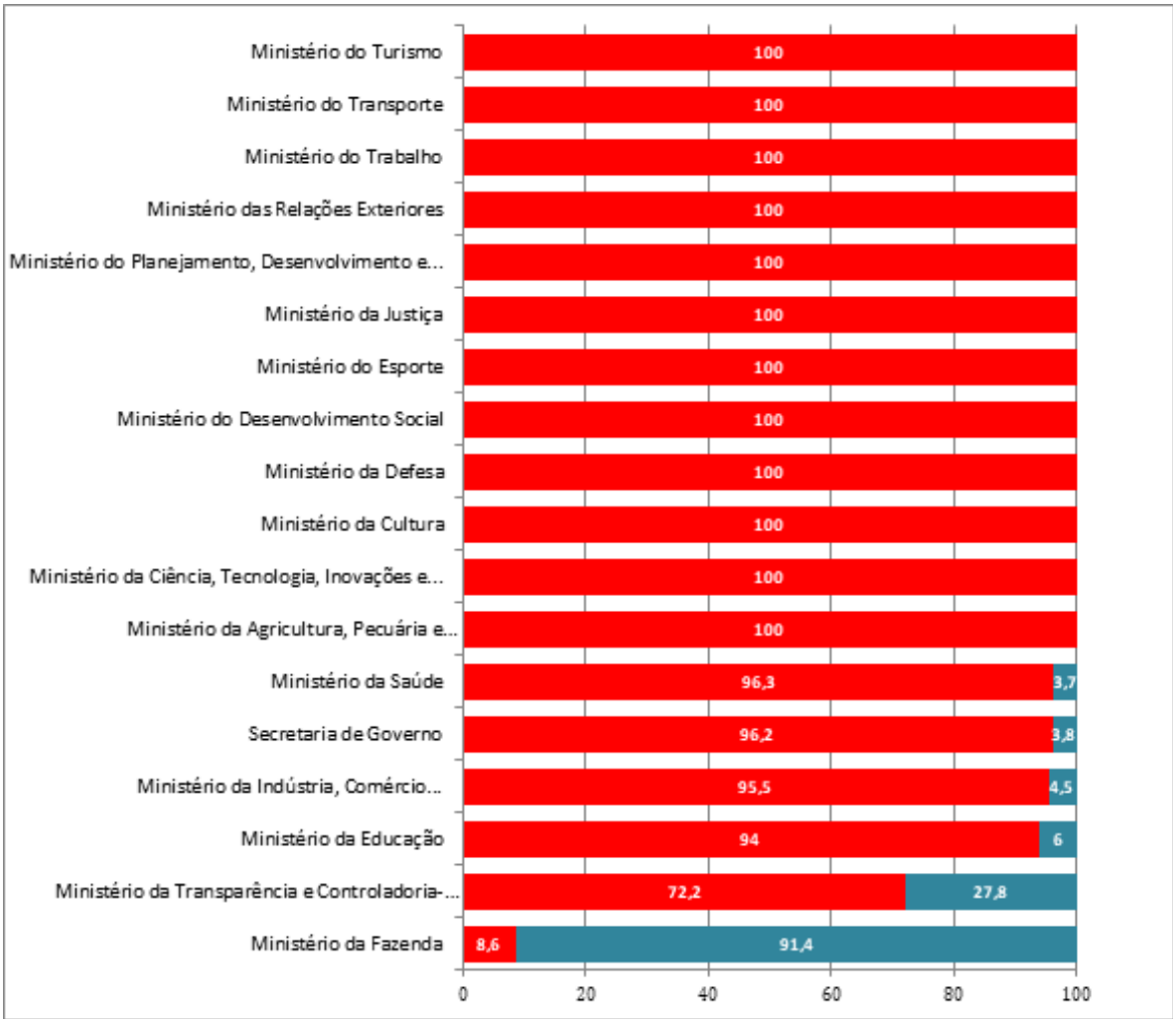
Possuem parametrização (Cartilha/Manual)

(N = 715)

Fonte: elaboração própria

Ainda como demonstra o Gráfico 5, no segundo gráfico temos a proporção de *datasets* que possuem parametrização, isto é, existência de algum tipo de manual/cartilha ou texto explicativo que possibilite maior segurança e qualidade no processo de entendimento pelo usuário acerca dos parâmetros dos dados, o que significam suas variáveis e a metodologia de mensuração de suas métricas. Neste indicador, é possível notar uma performance bem mais frágil em termos de inteligibilidade: apenas 17% dos *datasets* apresentaram este tipo de conteúdo, o que dificulta o reuso e a capacidade do usuário em compreender mais a fundo as variáveis, as métricas e categorias que compõem aquele conjunto de informações. Também ocorrem, entre os ministérios, desempenhos ainda mais heterogêneos e díspares. Como é possível visualizar no Gráficos 6, há percentuais que são, na verdade, "zero" para a maioria dos ministérios (sendo basicamente dois ministérios responsáveis por praticamente pela ocorrência deste indicador):

Gráfico 6 - Existência de parâmetros nos *datasets* por Ministério
(em termos percentuais quanto ao volume específico de cada órgão)

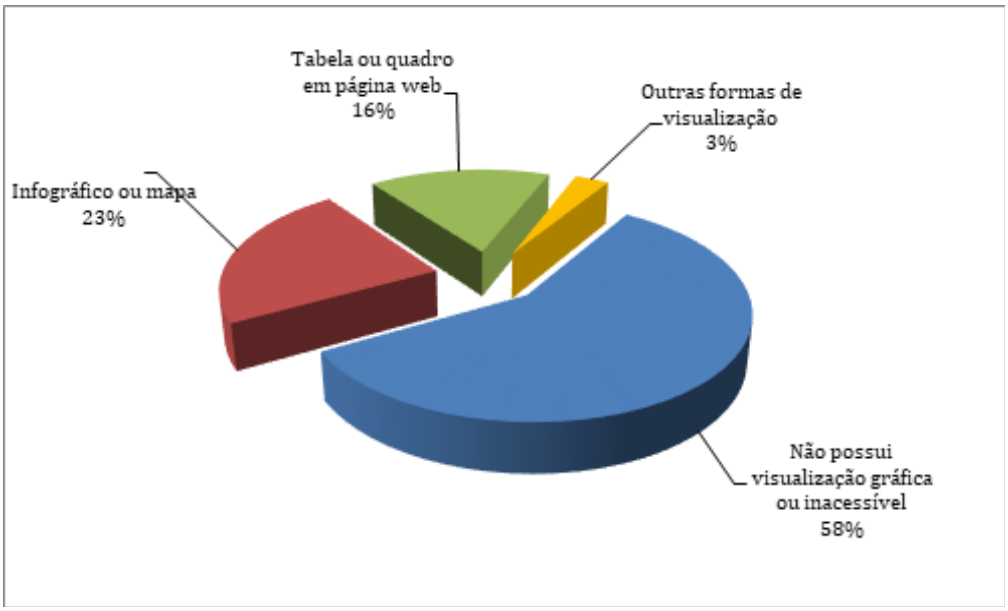


Fonte: elaboração própria (N = 715)

Na realidade, quase a totalidade dos ministérios não possui informações que tragam parâmetros explicativos sobre os dados publicados. A publicação de *datasets* devidamente parametrizados na amostra está concentrada basicamente no Ministério da Transparência com 27,8% de seus *datasets* com parâmetros publicados e, principalmente, o Ministério da Fazenda que possui o melhor desempenho neste indicador, com 91,4% de parametrização em seus *datasets*.

Um outro indicador de inteligibilidade analisado diz respeito à existência de visualização dos dados no próprio *link* dos recursos, que possibilita ao usuário ter uma visão panorâmica sobre os dados que pretende se apropriar. O Gráfico 7 traz os resultados da análise:

Gráfico 7 - Percentuais de *datasets* que contém alguma forma de visualização gráfica



Fonte: elaboração própria. (N = 715)

Mais da metade (58%) não possui qualquer forma de visualização. Quando há a ocorrência deste tipo de indicador, a forma mais comum é infográfico ou mapa (com preponderância para o primeiro) existente em 23% dos *datasets* seguida por tabela ou quadro em página HTML, presente em 16%. A visualização ajuda no processo de inteligibilidade dos dados por propiciar uma visão rápida e global do conteúdo dos arquivos auxiliando assim na tomada de decisões sobre o potencial reuso dos dados publicados.

Ainda na análise externa dos arquivos, os indicadores de discursividade e comunicação – vinculados ao pilar normativo da interatividade - apresentaram desempenho praticamente inexistente. Não há na estrutura de publicação do Portal Brasileiro de Dados Abertos mecanismos que possibilitem a interação efetiva, como o próprio portal explica:

> O Portal Brasileiro de Dados Abertos apenas organiza os dados abertos em um catálogo para fácil localização. Os responsáveis pelos dados são as organizações públicas que os coletam, gerenciam e publicam. Por isso cada uma dessas organizações que responde pelos seus próprios dados. Se você procura mais informações ou esclarecimentos sobre um conjunto de dados específico, procure na própria página do conjunto de dados a seção "Informações Adicionais". Nela há um campo "Autor". Esse é o responsável pelo conteúdo dos dados. Procure o nome do órgão que consta no valor desse campo para entrar em contato.[15]

15 Disponível em http://www.dados.gov.br/pagina/contato. Acesso 19 dez 2019.

Embora a referida tabela com "informações adicionais" conste o nome do departamento/órgão ou servidor responsável pela criação e também pela manutenção do conjunto de dados, nesta não há indicativos de canais de comunicação visíveis (a exemplo de e-*mail*, formulário, *chats* etc.) capazes de possibilitar a interação. A não disponibilidade de contato demandaria uma outra pesquisa específica por parte do usuário, extrapolando a análise do portal para descobrir como poderia se comunicar com os criadores/mantenedores dos dados. Por isso, por não haver tal mecanismo de contato exposto no entorno da publicação considerou-se inexistente, impossibilitando assim a aplicação de teste de responsividade[16] e outros instrumentos de coleta de dados correlatos[17]. Da mesma forma, no tocante ao indicador de discursividade, o Portal não traz informações sobre a existência de fóruns de discussão ou mecanismos similares de interação entre usuários[18].

4.2 Análise qualitativa interna

Na análise interna dos arquivos, um primeiro indicador a explorar refere-se a real disponibilidade dos arquivos. Neste item, analisamos se os *links* para os arquivos apresentados no Portal Brasileiro de Dados Abertos estão de fato acessíveis para *download*.

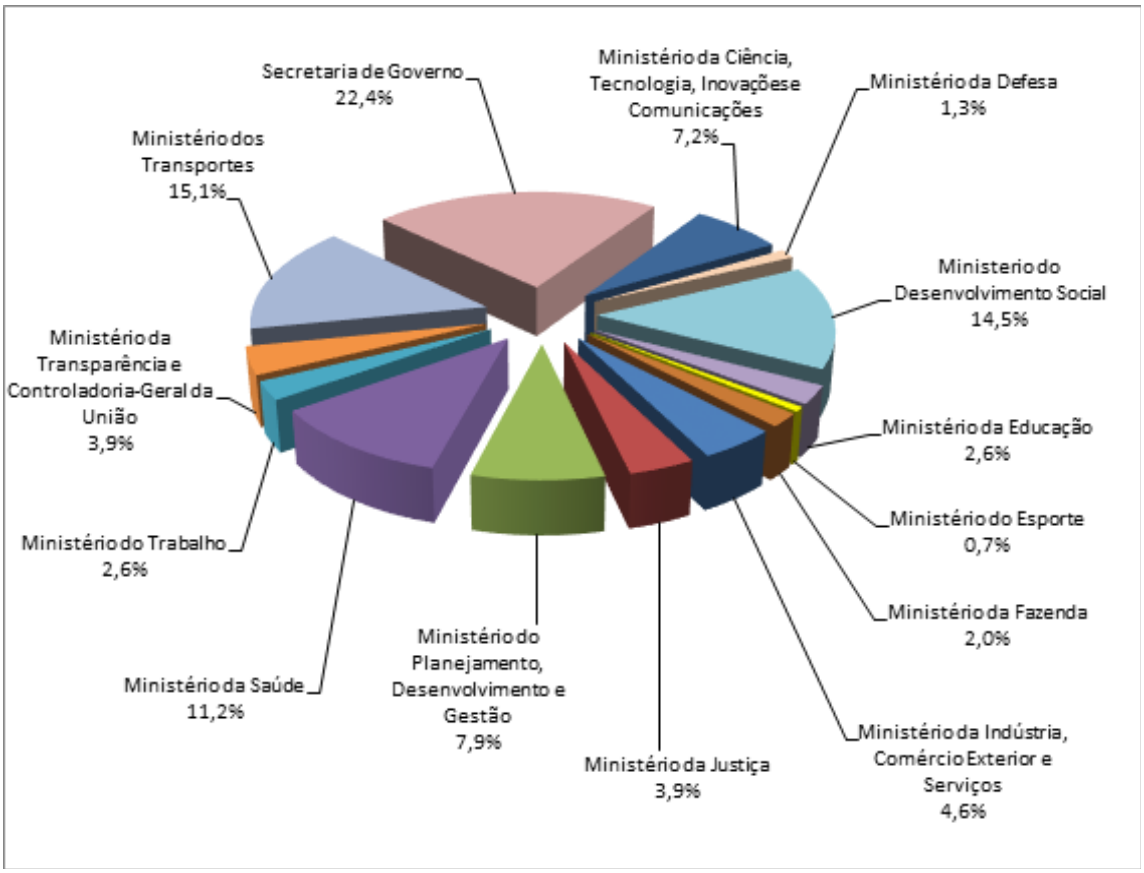
Detectou-se uma alta ocorrência de não-operacionalidade. Do total de 715 *datasets* 152 não possuíam arquivos para *download*, isto é, no momento em que o usuário tenta baixar os arquivos ocorrem erros como *links* quebrados ou endereço eletrônico não localizado pelo servidor-*web*. Isso significa que 21,2% dos *datasets* publicados não existem, na prática, em termos operacionais. Ao analisarmos este montante de *links* inoperantes, no Gráfico 8 temos a distribuição por Ministérios:

16 O teste de responsividade consiste no envio de pergunta simples (por exemplo, indagando sobre o significado de uma variável de um arquivo) direcionada ao mantenedor do *dataset* através do canal oficial informado na publicação (e-mail, formulário, chat etc.). O horário do envio é registrado, passando a se contabilizar o tempo de resposta do mantenedor e o grau de efetividade da resposta. O teste ajuda a verificar se o canal de comunicação informado de fato está operacional e gera interatividade satisfatória.

17 A título de nota metodológica, a interatividade deve ser considerada como um pilar que deve estar no horizonte mais amplo de desenvolvimento de metodologias que requer uma forma específica de análise, indicadores e instrumentos (como análise do conteúdo de conversação *online*; entrevistas com atores que fazem reuso de Dados abertos; entrevistas como gestores; teste de responsividade etc.).

18 Atualmente, existem fóruns ou grupos de discussão de usuários sobre Dados abertos (como *fanpages* do Facebook, grupos no Telegram etc.) mas que não constituem a política de publicação de Dados abertos pois são, na verdade, iniciativas independentes.

Gráfico 8 – Percentual de cada ministério no montante de *datasets* com *downloads* inoperantes



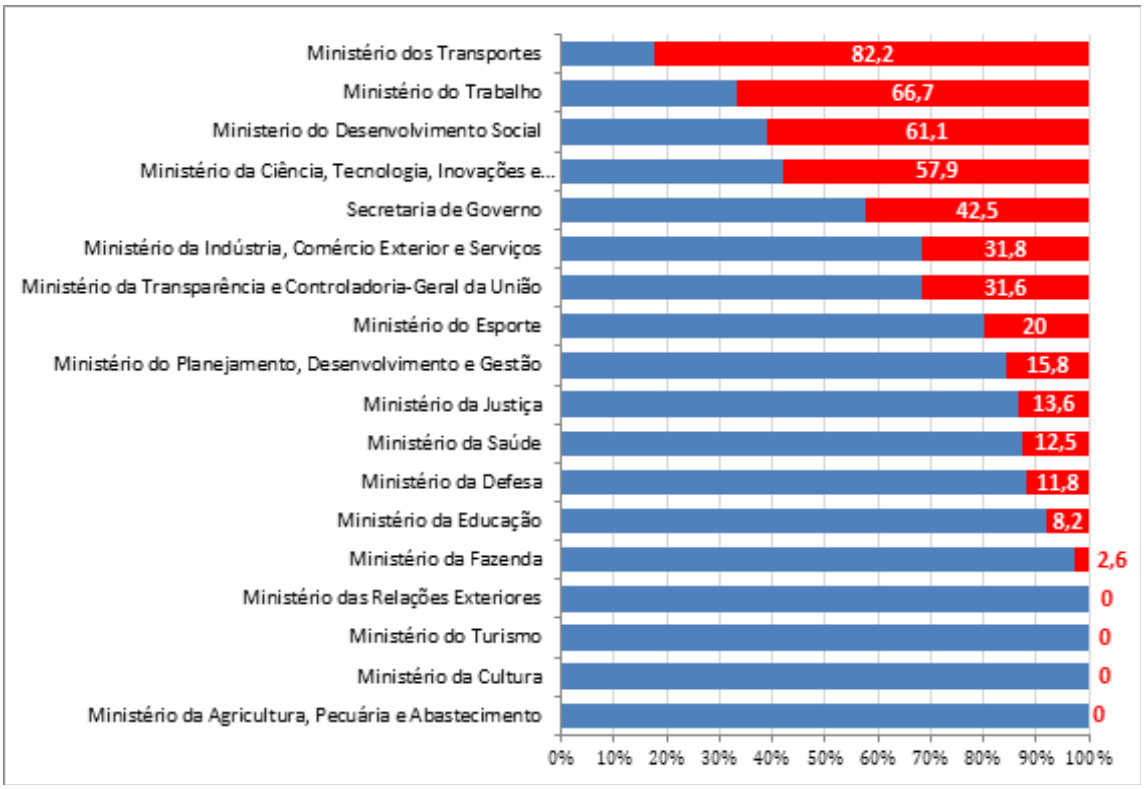
Fonte: elaboração própria. (N = 152[19])

Neste gráfico da (Figura 8) os percentuais não estão relativizados, isto é, trata-se do quanto cada ministério contribuiu para o total de *datasets* com *links* para arquivos inoperantes (quebrados) [20]. Numa perspectiva comparada, esses percentuais precisam ser dimensionados em relação ao volume de dados publicados por cada órgão, pois ministérios com grande volume de *datasets* tendem, naturalmente, a repercutir mais no percentual total de *links* inoperantes em termos absolutos. Por exemplo, embora o Ministério do Trabalho signifique apenas 2,6% dos *datasets* inoperantes, isso não quer dizer que é o órgão com menor volume proporcional de erros deste tipo (como veremos no próximo gráfico – Figura 9 – este ministério é o segundo com maior ocorrência proporcional de *datasets* inoperantes). O percentual é pequeno não porque este ministério tem baixa ocorrência de *links* inoperantes em relação aos seus arquivos próprios arquivos publicados, mas pelo fato do volume de *datasets* deste ministério ser bastante diminuto quando comparado a outros (apenas 6 *datasets*). Para termos uma visualização específica e proporcional acerca de quais ministérios tem maior ou menor percentual de ocorrência deste tipo de deficiência operacional, o Gráfico 9 traz essa visão mais relativizada:

19 Total de *datasets* cujos downloads para recursos (arquivos) estão inoperantes (links quebrados).

20 Os seguintes ministérios não aparecem no gráfico por não terem *datasets* com downloads inoperantes: Ministério da Agricultura, Pecuária e Abastecimento; Ministério da Cultura; Ministério das Relações Exteriores e Ministério do Turismo.

Gráfico 9 – Percentual de *datasets* com arquivos inoperantes por ministério



Fonte: elaboração própria.

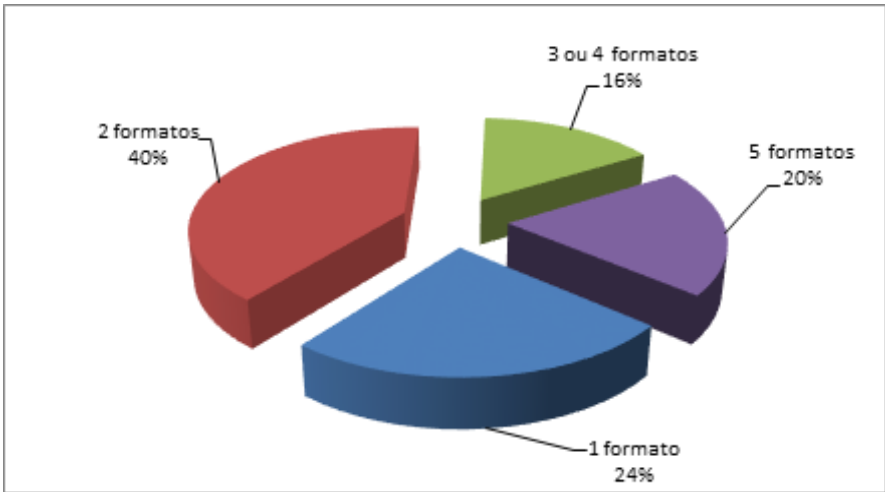
(N = 715, sendo que as barras gráficas em cada item trata do N total de *datasets* de cada ministério)

	Datasets com arquivos indisponíveis
	Datasets com arquivos existentes

Neste gráfico, o Ministério do Transporte é o órgão com maior percentual de *datasets* inoperantes, cerca de 82%. Em seguida temos o Ministério do Trabalho; Ministério do Desenvolvimento Social e Ministério da Ciência, Tecnologia, Inovações e Comunicação com pior desempenho neste indicador (todos acima ou em torno de 60% de *datasets* com arquivos indisponíveis para *download*). Por outro lado, não foram encontrados *datasets* com *downloads* inoperantes nos seguintes órgãos: Ministério da Agricultura, Pecuária e Abastecimento; Ministério das Relações Exteriores; Ministério da Cultura e Ministério do Turismo. Porém, mais uma vez, devemos também ponderar esse dado ao lembrarmos que os dois primeiros Ministérios (Agricultura e Relações Exteriores) têm baixa performance de publicação, isto é, possuem pequeno volume de dados publicados. De modo mais ponderado, o melhor desempenho deste indicador está no Ministério da Fazenda, que possui um volume relativamente alto de *datasets* (117), porém com um índice baixo de arquivos com *links* inoperantes (apenas 2,6 % de *downloads* inativos).

Um outro indicador na análise interna dos arquivos trata da formatação. Neste item, os resultados demonstraram que mais da metade dos *datasets* (64%) é composta por apenas dois formatos de arquivo (sendo 24% possuem apenas um tipo arquivo). Apenas 36% trazem maior diversidade neste item comportando três ou cinco formatos diferentes publicados, conforme detalha o Gráfico 10:

Gráfico 10 – *Datasets* por quantidade de formatos publicados



Fonte: elaboração própria.

N = 563 (excluídos *datasets* com arquivos inacessíveis)

Isso demonstra que uma parte significativa (mais da metade) dos *datasets* não possui pluralidade de formatos, o que pode representar pouca versatilidade em um horizonte que deve prever diferentes tipos de usuários e apropriações, impactando na sua operacionalidade.

O formato mais recorrente de arquivo é a extensão CSV, presente em 57,5% dos *datasets*, seguida de JSON, existente em 43,3 %; HTML que está presente em 29,3% dos *datasets*; PDF, em 25.6%; e XML, em 19,5 %. Formatos em RDF (que remetem a Dados Abertos Conectados[21]) são raros, apenas em menos de 1% dos *datasets*. Porém, neste indicador, o dado mais importante está no percentual de *datasets* compostos apenas por arquivos fechados. Cerca de 8,3% dos *datasets* com

21 Dados Abertos Conectados estão vinculados à concepção de *Linked Data* (em português, geralmente traduzido por "dados conectados" ou "dados ligados") que está relacionado à construção da websemântica. Na ideia de Dados Conectados os arquivos possibilitam a criação de uma *web* de dados, pois possuem informações sobre seus conteúdos (metadados) operando como "*links*" capazes de indexar termos ou temas sobre os conteúdos que os arquivos possuem, possibilitando assim o rastreamento desses temas (que podem ser assuntos, pessoas jurídicas, pessoas físicas etc.) em diversos arquivos dispersos em diferentes bases de dados. Já a sigla RDF significa *Resource Description Framework* (estrutura de descrição de recursos) trata de uma forma de notação utilizada na construção de Dados Conectados. Assim, a ideia de Dados Abertos Conectados seriam arquivos de Dados Abertos utilizando alguma forma de *Linked Data*.

arquivos existentes para *download* trazem apenas arquivos em formatos fechados (principalmente em PDF e HTML) não se encaixando assim no conceito de Dados Abertos, portanto, de difícil grau de operacionalidade.

Conforme foi apontado anteriormente, 21,2% dos *datasets* publicados (152 *datasets* de 715) não possuíam arquivos para *downloads* (*links* quebrados), por isso, inaptos para serem analisados por estarem inacessíveis. Além disso, 8,3% desses *datasets* com arquivo acessíveis (47 de 563) traziam apenas arquivos fechados (não-estruturados), por isso, inaptos para serem analisados internamente. Assim, o *corpus* de *datasets* com arquivos acessíveis e estruturados aptos para terem seus arquivos abertos e analisados internamento ficou com N = 516. Para seguirmos adiante com o desempenho de outros indicadores que tratam da análise interna dos arquivos, convém a partir de agora levarmos em conta este N de 516 *datasets* viáveis para serem de fato abertos e avaliados. Por isso, os próximos percentuais recairão sobre este *corpus* de *datasets* que de fato existem e tratam de arquivos estruturados[22].

Detectou-se que em 15,3% desse *corpus* de arquivos existentes há ocorrência de problemas quanto ao indicador de padronização. Problemas de padronização podem significar diversas disfunções que impedem a leitura automática por máquinas. Dentre elas, problemas como resquícios de notações que não tem valor efetivo em tabelas estruturadas dificultando o processamento automatizado[23]; grafias divergentes que consistem em dados de uma mesma categoria de informação grafados de formas distintas, quando os dados deveriam ser niformizados; trechos ilegíveis[24] e erros de estruturação de arquivos[25]. Esses elementos afetam a qualidade do *dataset* pois recursos com informações

22 Ou seja, dos 715 *datasets* iniciais que apontavam nas publicações para a possível existência de 2.743 arquivos, restou na prática 516 *datasets* que somaram, ao final, um montante de 1860 arquivos de fato aptos para serem estudados em suas estruturas internas.

23 São dados apresentados com valores fora de padrões e que dão um pouco mais de trabalho para o usuário conseguir gerar um gráfico ou fazer cruzamentos diversos. Vem na forma de palavras ou números que estão "sobrando" e que não trazem nenhum tipo de informação relevante e "despadronizam" os dados. Exemplo: 05- 10-2018 00: 00: 00: 00. Neste caso, o excesso de zeros sem valores reais atribuídos(minutos, segundos….) é despadronizante, pois tem valor vazio e dificulta a leitura automática de alguns *softwares* (usados por jornalistas de dados, por exemplo) ou exige que o usuário do *dataset* se faça a limpeza prévia dos dados antes de utilizá-lo ou ainda que se escreva mais linhas de códigos para contornar o elemento despadronizante.

24 São considerados trechos ilegíveis partes que significam problemas de leitura ou erros de caracteres especiais (por exemplo, &%%%#@+!$) que não formam palavras ou conjunto lógico de números Caracteres exógenos no meio da estrutura de palavras onde deveria haver acentuação ou cedilha (como PopulaÃ§Ã£o, MANUTEN&@*AO, AVIA%%O) não foram erros considerados nesta categoria pois são típicos problemas de codificação (geralmente UTF-8 , Latin-1 ou CP1252) e não uma disfunção do arquivo. De todo modo, no procedimento de abertura dos arquivos os pesquisadores e pesquisadoras foram treinados para abri-los de modo correto, respeitando a codificação adequada a fim de evitar que tais desvios de caracteres.

25 Erros de estruturação ocorrem principalmente em casos de arquivos CSV cuja quantidade de variáveis (colunas) identificadas na primeira linha é incompatível com a quantidade de resultados a partir da segunda linha do arquivo. Isso gera um erro no momento de abertura do arquivo, forçando o usuário a desistir de utilizar o arquivo ou, usuários mais experientes acabam gastando parte do tempo para identificar o problema e, uma vez identificado, excluir a coluna não nomeada ou nomeá-la por conta própria quando os dados apontam claramente do que se trata.

"despadronizadas" requerem maior energia do usuário em tratá-las, uniformizá-las para um efetivo processo de reuso. Em alguns casos, pode inviabilizar a apropriação da base de dados.

Outros dois aspectos de operacionalidade analisados consistem nos indicadores de integridade e opacidade. A integridade está diretamente vinculada à existência de lacuna de dados no interior dos arquivos, por exemplo, células de *dataframes* (planilha) com *missing*. Consideramos uma lacuna de dado quando o recurso possui claramente ausências de informações, principalmente em caso de tabelas com células vazias em alguma variável. Os resultados demonstram que 17,2% dos recursos que estavam disponíveis para *download* possuem lacunas de dados, isto é, possuem problemas de integridade do conteúdo. Em termos operacionais, células vazias de dados podem comprometer o uso de um *dataframe* pois, a depender da variável, a falta de informação pode enviesar o dado e impossibilitar resultados de análises completas. Em relação à opacidade, cerca de 10% dos arquivos existentes para a análise apresentaram problemas de variáveis sem uma definição clara sobre o seu real significado. Uma variável é opaca ocorre quando o recurso possui pelo menos uma de suas variáveis não inteligíveis para o usuário, ou seja, quando não é possível ter certeza sobre o que significa os valores de uma coluna, por exemplo, por não ser intuitiva ou por não haver explicações sobre o seu significado.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como principal objetivo contribuir para o amadurecimento, desenvolvimento e aplicabilidade de indicadores e metodologia para avaliação da qualidade de Dados Abertos visando seu efetivo reuso e apropriação social. Como estudo de caso, o artigo avaliou todos os *datasets* do governo federal brasileiro, em nível ministerial, publicados no Portal Brasileiro de Dados Abertos, atualmente a principal plataforma no país neste campo.

Do ponto de vista teórico, afirmamos que a concepção de Dados Abertos traz em suas bases três premissas importantes e que foram destacadas. Primeiro, Dados Abertos são informações completas acessíveis ao público de forma estruturada, automatizada e dinamizada por vias digitais. A finalidade dos Dados Abertos é múltipla, isto é, não está presa apenas a questões políticas específicas, mas também a questões de utilidade pública. Terceiro, há o pressuposto de que são bens públicos e, por isso, o Estado tem a obrigação de publicá-los. Tais dimensões são pressupostos teóricos que apontam para o caráter normativo dos Dados Abertos e também reforçam o papel da apropriação social: uma

condição inerente ao próprio conceito de Dados Abertos. Por estar normativamente guiada por estes princípios, demonstramos que a política de Dados Abertos é, naturalmente, objeto de avaliações e assim apontamos a necessidade de desenvolvimento de pesquisas capazes de avaliar a qualidade dos recursos publicados, tendo em vista que tal abordagem ainda é bastante incipiente.

Neste sentido, o artigo tomou como ponto de partida a análise qualitativa interna (no interior dos arquivos) e a análise qualitativa externa (no entorno da publicação do arquivo) aprimorando indicadores e cruzando-os com três pilares normativos que a publicação dos recursos deve cumprir: (a) inteligibilidade, (b) operacionalidade e (c) interatividade.

No tocante à inteligibilidade, a pesquisa analisou primeiramente o indicador de temporalidade e demonstrou que o ciclo de publicação dos *datasets* no Portal Brasileiro de Dados Abertos tem ocorrido de modo ondular-agudo com um ou dois dias por ano com grande volume de publicação. Referente à atualização, identificou-se que o ritmo de *update* em metade dos *datasets* (56%) no Portal é acima de três meses e a ocorrência de atualizações semanais são escassas (menos de 10%). Em relação ao indicador de rotulação, observou-se que 73% dos *datasets* trazem algum tipo de rotulação. Embora este índice seja razoável - pois demonstra que boa parte dos *datasets* são devidamente rotulados - chamamos a atenção para o fato de um terço não possuírem tal indicador, apresentando assim pouca inteligibilidade no nível inicial de contato com o usuário nesta parcela.

Além disso, apenas 17% da amostra de *datasets* tiveram performance positiva quanto ao indicador de parametrização, isto é, *datasets* com informação detalhada sobre seus parâmetros, tais como explicação sobre metodologia empregada e métricas, ou manuais, cartilhas e similares. Isso significa que uma característica preponderante em quase todos os recursos é a prevalência de um baixíssimo grau de elementos informativos mais aprofundados sobre os dados capazes de propiciar maior inteligibilidade ao usuário e um poder de reuso mais efetivo. Estas performances são heterogêneas, pois há ministérios nos quais nenhum *dataset* possui rotulação ou parametrização, enquanto outros possuem desempenho bem mais adequado neste indicador. Um último indicador de inteligibilidade analisado tratou da existência de elementos de visualização gráfica dos dados, detectado em 42% dos conjuntos de dados (os outros 58% não possui qualquer forma de visualização). Quando há a ocorrência deste tipo de indicador, a forma mais comum é infográfico ou mapa (com preponderância para o primeiro).

Quanto ao pilar da operacionalidade, detectou-se que cerca de 8% dos *datasets* com arquivos aptos para *download* não se enquadram no conceito normativo de Dados Abertos por serem conjuntos

de dados compostos tão somente por arquivos não-estruturados (como PDF, HTML, JPG etc), isto é, sem qualquer arquivo legíveis por máquinas de forma automática. Além disso, o fato de termos cerca de 21% dos recursos de toda a amostra totalmente inacessível (erros de *links* quebrados ou endereço eletrônico não localizado pelo servidor-web etc.) demonstra que há um alto grau de não-operacionalidade (em alguns ministérios, a taxa de *links* inoperantes ultrapassa os 60% chegando até a casa dos 80%).

Se extrairmos dos 715 *datasets* iniciais da amostra aqueles que são constituídos apenas por arquivos não-estruturado (8,3%) e também subtrairmos aqueles *datasets* que não possuem arquivos para download (21,2%) devido a *links* inoperantes, na prática, quase 30% dos *datasets* dos ministérios publicados no Portal Brasileiro de Dados Abertos não são inviáveis operacionalmente, ou seja, quase 1/3 (um terço) do que está anunciado como "publicado" não pode ser qualificado como Dados Abertos (ou por não existirem para o usuário, de fato, ou por se conterem apenas arquivos fechados).

Quando a análise recaiu apenas nos arquivos existentes, desse montante outros problemas de operacionalidade também foram detectados: cerca de 15% apresentaram problemas de padronização; em cerca de 17% desses existentes também apresentaram problemas de integridade de informação (lacunas de dados). Em relação ao indicador de opacidade, cerca de 10% desses arquivos apresentaram problemas de variáveis sem uma definição clara sobre o seu real significado. No caso de disfunções despadronizantes, isso aumenta o custo de apropriação pois demanda mais do usuário fazer correções e uniformizar os arquivos. No caso de problemas de integridade e opacidade, significa que há um vazio de informação para o usuário que pode inutilizar o recurso por comprometer a efetiva operacionalidade no ato de apropriação.

Na perspectiva mais ampla, ao combinarmos as ocorrências gerais de erros na análise interna dos arquivos existentes, convém destacar que cerca de 41% dos arquivos que compõem os *datasets* com recursos para *download* (sem levar em conta aqui os arquivos anunciados como disponíveis, porém inexistentes) apresentaram algum tipo de indicador que dificulta o seu reuso em termos operacionais .

Por fim, quanto ao pilar da interatividade este foi praticamente inexistente, pois não havia na estrutura do Portal, no momento da coleta de dados, elementos de discursividade ou comunicação claramente aportados e disponíveis para o usuário.

Diante desse quadro, os indicadores aplicados neste estudo foram capazes de configurar um panorama avaliativo que possibilita o mapeamento de problemas e a necessidade do desenvolvimento de boas práticas na melhoria da qualidade dos Dados Abertos. Tendo em vista que a concepção de

Dados Abertos pressupõe que são bens públicos, que devem ser disponibilizados de modo livre e estruturados, visando a sua apropriação dinâmica através de ferramentas digitais, a existência de barreiras técnicas detectadas na publicação dos recursos significa, na prática, o não cumprimento dessas premissas e o reforço das assimetrias de conhecimento entre os diversos atores-usuários, pois para driblar tais fragilidades é preciso que o usuário tenha capacidade de ultrapassá-las, o que requer arregimentar mais recursos humanos e *expertise* para rompê-las.

O estudo chama a atenção para a importância de se consolidar indicadores de caráter longitudinal que avaliem a qualidade dos *datasets* e buscar análises complementares, sobretudo estudos de recepção que possam avaliar como a apropriação social de Dados Abertos está ocorrendo na prática, levando em conta a ótica e o perfil dos diferentes tipos de usuários. São algumas das questões que estão vinculadas ao horizonte maior desta pesquisa e que precisam ser objeto de análise no desenvolvimento de uma avaliação efetiva e contínua das políticas de Dados Abertos nos próximos anos.

Agradecimentos

Registramos nossos agradecimentos aos demais colegas integrantes do Centro de Estudos em Comunicação, Tecnologia e Política (CTPol) da Universidade de Brasília (UnB), pela contribuição e apoio durante o processo de discussão e testes metodológicos que precederam este estudo. Em ordem alfabética: Diogo Campos, Ébida Santos, Eduardo de Lima Rodrigues, Luana Ferreira Alves, Maria Eduarda Gomes de Souza, Mariah Sampaio Luciano, Nicole de Faria Bartolini Mattiello, Pedro Ivo de Sá Guimaraes e Vivian Peron.

REFERÊNCIAS

ATTARD, J., ORLANDI, F., SCERRI, S., AUER, S. A systematic review of open government data initiatives. **Government Information Quarterly**, 32(4), 2015, pp. 399–418.

BATINI, C., CAPPIELLO, C., Francalanci , C., MAURINO, A. Methodologies for data quality assessment and improvement. **ACM Computing Surveys**, v.41, n.3, p.1-52, 2009.

BRASIL. **Lei de Acesso à Informação**. LEI Nº 12.527, de 18 de novembro de 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm, 2011, acesso 16 de janeiro 2018.

BRASIL **Decreto 8.777 de maio de 2016**. Institui a Política de Dados Abertos do Poder Executivo federal. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2016/Decreto/D8777.htm, 2016, acesso 16 de janeiro 2018.

CHIGNARD, Simon. **A brief history of Open Data**. Disponível em http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data, 2013.

CHUNA, Soon Ae *et al*. Government 2.0: Making connections between citizens, data and government. **Information Polity**, v.15, p. 1-9, 2010.

COELHO, G. Bandeira. **Sociologia do conhecimento e da ciência**: da sua emergência a Pierre Bourdieu. **Sinais**, v.21, n.2, p.266-294, 2017.

GURSTEIN, Michael. Open data: empowering the empowered or effective data use for everyone? **First Monday**, v.16, n.2, 2011.

HUIJBOOM, Noor; BROEK, Tijs Van den; Open data: an international comparison of strategies. **European Journal of ePractice**, v.12, p.4-16, 2011.

JANSSEN, Marijn, CHARALABIDIS, Yannis; ZUIDERWIJK, Anneke. Benefits, Adoption Barriers and Myths of Open Data and Open Government. **Information Systems Management**, v.29, n.4, p.258-268, 2012.

KUCERA, Jan et al: Methodologies and Best Practices for Open Data Publication. **Dateso 2015**, n.1.343, p.52-64, 2015. Disponível em: http://ceur-ws.org/Vol-1343/. Acesso em 07 jun 2020.

MÁCHOVÁ, Renata; LNĚNIČKA, Martin. Evaluating the Quality of Open Data Portals on the National Level. **Journal of Theoretical and Applied Electronic Commerce Research***,* v.12, n.1, p.21-41, 2017.

MCDERMOTT, P. Building open government. **Government Information Quarterly***,* v.27, p. 401-413, 2010.

MEIJER, A. J.; CURTIN, D., HILLEBRANDT, M. (2012). Open government: connecting vision and voice. **International Review of Administrative Sciences***,* v.78, n.1, p.10-29, 2012.

MERTON, Robert K. **The Sociology of Science**: theoretical and empirical investigations. Chicago e Londres: The University of Chicago Press, 1973.

OECD. **Government at a Glance 2017**. Paris: OECD Publishing, 2017. Disponível em http://dx.doi.org/10.1787/gov_glance-2017-en, 2017, Acesso em 03 julho 2018.

OECD. **Open Government:** The global context and the way forward 2016. Paris: OECD, 2016. Disponível em http://www.oecd.org/gov/open-government.htm, 2016.

ONU. Órgão das Nações Unidas. **Guide on Lessons for Open Government Data Action** Planning for Sustainable Development. Nova York: ONU, Department of Economic and Social Affairs. Disponível em http://workspace.unpan.org/sites/Internet/Documents/UNPAN97913.pdf, 2017. Acesso 13 de abril 2018.

OSAGIE, Edobor *et al*. Usability Evaluation of an Open Data Platform. Texto apresentado na **18th Annual International Conference on Digital Government Research**. Staten Island, 2017.

RUIJER, Erna *et al*. Open data for democracy: Developing a theoretical framework for open data use. **Government Information Quarterly**, v.34, p.45-52, 2017.

SÁ, C., GRIECO, J. Open Data for Science, Policy, and the Public Good. **Review of Policy Research**, v.33, n.5, p.523-543, 2016.

SAFAROV, I., GRIMMELIKHUIJSEN, S.G., MEIJER, A.J. Utilization of open government data: A systematic literature review of types, conditions, effects and users. **Information Polity**, v.22, n.1, p. 1-24, 2017.

SILVA, Sivaldo Pereira da; RABELO, Leon Eugênio Monteiro; SANTOS, Ébida Rosa dos;

LUCIANO, Mariah Sampaio F. Avaliando a política de Dados abertos no Legislativo brasileiro: análise qualitativa dos *datasets* da Câmara dos Deputados. **Revista Compolítica**, v.10, n.1, p. 137-160, 2020.

SUSHA, I., ZUIDERWIJK, A., JANSSEN, Marijn, GRONLUND, A. Benchmarks for Evaluating the Progress of Open Data Adoption: Usage, Limitations, and Lessons Learned. **Social Science Computer Review**, v.33, n.5, p.613-630, 2015.

WIRTZA, Bernd W., BIRKMEYERA, Steven. Open Government: Origin, Development, and Conceptual Perspectives. **International Journal of Public Administration**, 2015, pp. 1-16.

ZUIDERWIJK, Anneke; JANSSEN, Marijn. Participation and Data Quality in Open Data use: Open Data Infrastructures Evaluated. **15th European Conference on e-Government**, Portsmouth, UK, 2015.

ZUIDERWIJK, Anneke; JANSSEN, Marijn. Open data policies, their implementation and impact: A framework for comparison. **Government Information Quarterly**, v.31, n.1, p. 17-29, 2014.