UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

SHIRLEI LÚCIA OLIVEIRA DO CARMO

CompOD: Framework de conformidade LGPD para dados abertos

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. Cláudio Fernando Resin Geyer
Orientador

Prof. Dr. Julio Cesar Santos dos Anjos
Co-orientador

Porto Alegre, Fevereiro de 2024

CIP - Catalogação na Publicação

DO CARMO, SHIRLEI LUCIA OLIVEIRA
CompOD: Framework de conformidade LGPD para dados abertos / SHIRLEI LUCIA OLIVEIRA DO CARMO. -- 2024.
114 f.
Orientador: Claudio Fernando Resin Geyer.

Coorientador: Julio César Santos dos Anjos.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2024.

1. OPEN DATA . 2. PERSONAL IDENTIFIABLE INFORMATION. 3. DATA PROTECTION . 4. LGPD COMPLIANCE. 5. DATA SHARING. I. Geyer, Claudio Fernando Resin, orient. II. dos Anjos, Julio César Santos, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

“Se eu enxerguei mais longe,
foi porque me apoiei nos ombros de gigantes.”
— Sir Isaac Newton

AGRADECIMENTOS

Durante vários meses a seção de agradecimentos continha somente a seguinte frase: *"Agradeço a todos..."*. Ao término de um árduo trabalho, é importante avaliar o caminho percorrido e relembrar dos momentos em que alguém dedicou parte de seu tempo para dividir ideias e compartilhar sonhos. Nem todos os momentos serão relembrados certamente. Porém, como relâmpagos numa noite de chuva, surge na memória alguma cena para relembrar uma situação vivida.

A primeira coisa que pensei foi agradecer a Deus, que me deu o sopro da vida, por ter conseguido chegar ao fim dessa etapa, e sem o qual nada faria sentido. Logo em seguida, pensei em quem sempre me apoiou e esteve ao meu lado em todos os momentos, meu porto seguro: minha família. Especialmente minha querida Mãe Nara — a mulher mais forte e de maior *fé* que Deus me oportunizou conhecer — que incondicionalmente me deu forças em todos os momentos difíceis. Aos meus amados irmãos Sheila e Júnior, por deixarem meus dias mais leves, e me mostrarem a beleza nas adversidades. E também, aos meus pequeninos amados: Diogo, Cecília e Donatela.

Ao meu orientador Prof. Dr. Cláudio Geyer, e co-orientador Prof. Dr. Júlio Anjos queria agradecer a paciência e a dedicação dispensada durante estes anos todos. A vossa aposta na ideia de estudar o *Open Data* foi fundamental para materializar este trabalho. Obrigada por não me deixarem desistir, mesmo em meio as incertezas de uma pandemia (COVID-19), do medo, perdas, e ansiedade que toda sociedade sofreu e sofre como consequência. Obrigada por não desistirem da educação, ciência e pesquisa, pois, sabemos que no Brasil muito pouco são valorizadas. O vosso papel é um pilar fundamental para construção de uma sociedade civilizada e sábia. Agradeço à todo time de professores e funcionários do Instituto de Informática da UFRGS. Em fim, *"agradeço a todos..."*

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ACM	Association for Computing Machinery
API	Application Programming Interface
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma Separated values
DAAP	Diretoria de Análise de Políticas Públicas
DCAT	Data Catalog Vocabulary
FGV	Fundação Getulio Vargas
GDPR	General Data Protection Regulation
GODI	Global Open Data Index
GPS	Global Positioning System
IA	Inteligência Artificial
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
ML	Machine Learning
NLP	Natural Language Processing
NRT	Near Real Time
LAI	Lei de Acesso à Informação
LGPD	Lei Geral de Proteção de Dados
LOD	Linked Open Data
OAD	Open Access Data
ODP	Open Data Platform
OGD	Open Government Data
OGP	Open Government Partnership
OGPD	Open Government Partnership Declaration
OGPL	Open Government Platform
OKFN	Open Knowledge Foundation

OWL	Ontology Web Language
PDF	Portable Document Format
PII	Personal Identifiable Information
RDF	Resource Description Framework
SLR	Systematic Literature Review
SVM	Support Vector Machine
TIC	Tecnologia da Informação e Comunicações
URL	Uniform Resource Locator
WWW	World Wide Web
W3C	World Wide Web Consortium

LISTA DE SÍMBOLOS

$\sum \frac{a}{b}$ Somatório do produtório

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Dados abertos são um conceito atribuído ao compartilhamento de dados com qualquer pessoa. Além de serem acessados, esses dados podem ser manipulados e redistribuídos. Esta é uma tendência global que incentiva a transparência dos governos e das entidades nas suas transações, além de fornecer à sociedade conhecimento sobre dados relevantes em áreas como infraestrutura, saúde, gastos públicos e meio ambiente. Diversas iniciativas discutem atualmente a importância dos dados abertos, seja para a sociedade ou para uso e suporte em várias áreas, como em tecnologia da informação, durante o treinamento de inteligência artificial que necessitam de um grande volume de dados para operar com precisão. O uso otimizado e intercambiável de dados abertos entre organizações pode levar à chamada inovação aberta, que pode ser entendida como a utilização não apenas de dados internos das organizações, mas também de dados externos para cruzar informações e gerar sistemas e soluções mais completos e inovadores. Sabendo da relevância desse tipo de dados e de seus desafios, foi desenvolvido um framework de análise da dados abertos e verificação de sua conformidade com a lei brasileira de proteção de dados, a LGPD. A partir deste framework, foi realizado um estudo quantitativo e qualitativo de dados abertos no Brasil, e verificada a sua conformidade com a referida lei brasileira de proteção de dados, de modo a avaliar a saúde dos dados abertos disponíveis no Brasil, sendo este o diferencial da solução proposta. A metodologia utilizada nesta dissertação compreende uma análise automatizada de portais de dados abertos brasileiros por meio do sistema de gerenciamento de dados CKAN, utilizado para publicação e compartilhamento de dados abertos no mundo todo. O escopo do estudo abrange todos os estados e instituições governamentais que possuem portais de dados abertos e que são expostos por meio do CKAN. A análise quantitativa validou 1.817 conjuntos de dados, em 19 portais de dados, e verificou que, assim como estudos da literatura mencionam, as principais lacunas nos dados abertos são: dados desatualizados, ausência de metadados para suportar sua utilização e reutilização, especificação de licença de uso imprecisa e heterogeneidade no modo de exposição dos dados entre portais. O módulo de verificação de conformidade com a LGDP analisou 11.154 recursos, contidos em 812 datasets dos estados brasileiros, e o resultado mostrou que, em quase todas as regiões analisadas, havia conjuntos de dados com informações pessoais (como CPF) expostas. Logo, apesar do grande potencial dos dados abertos, os obstáculos também se mostraram desafiadores, e as ferramentas de controle e fiscalização necessárias.

Palavras-chave: Dados abertos; plataforma de dados abertos; framework de dados abertos; lgpd; compartilhamento de dados pessoais; informações de identificação pessoal; leis de proteção de dados; lei geral de proteção de dados pessoais.

ABSTRACT

Open data is a concept attributed to sharing data with anyone. Besides being accessed, this data can be manipulated and redistributed. This is a global trend that encourages transparency by governments and entities in their transactions, in addition to providing society with knowledge about relevant data in areas such as infrastructure, health, public spending and the environment. Several initiatives currently discuss the importance of open data, whether for society or for use and support in different areas, such as artificial intelligence training or models that use machine learning and require a large volume of data to operate accurately. The optimized and interchangeable use of open data between organizations can lead to so-called open innovation, which can be understood as the use not only of organizations' internal data, but also of external data to cross-reference information and generate more complete and innovative systems and solutions. Knowing the relevance of this type of data and its challenges, a quantitative and qualitative study of open data in Brazil was performed, in addition to the development of a prototype that validated the compliance of this data with the Brazilian data protection law, LGPD, this being the differential of the proposed solution. The goal is to assess the health of open data available in Brazil. The methodology used in this article comprises an automated analysis of Brazilian open data portals through the CKAN data management system - used in open data portals around the world for publishing and sharing open data. The scope of the study covers all states and government institutions that have open data portals and are exposed through CKAN. Quantitative analysis validated 1,817 datasets across 19 data portals. It was found that, as studies in the literature mention, the main points of improvement in open data are: outdated data, lack of metadata to support the use and reuse of these, inaccurate usage license specification, and heterogeneity in the way data is exposed between portals. The LGDP compliance verification module analyzed 11,154 resources, contained in 812 datasets from Brazilian states, the result showed that in almost all regions analyzed, there were data sets with personal information (such as CPF) exposed. Therefore, despite the great potential of open data, the obstacles also proved to be challenging, and control and inspection tools are necessary.

Keywords: open data; open data platform; open data framework; lgpd framework; data sharing; personal identifiable information; data protection law.

1 INTRODUÇÃO

Produção de conhecimento através de dados é uma disciplina difundida com frequência na atualidade, uma vez que, na época atual, crescentes são os mecanismos e fluxos capazes de coletar e disponibilizar variados tipo de dados, com finalidades específicas, e em diferentes contextos diários de uma população. Os benefícios da exploração desses dados podem ser tanto em favor da própria pessoa que gera os dados, quanto em estudo e aplicação em soluções que atinjam mais pessoas em uma comunidade.

Os benefícios da análise de dados na área da saúde incluem, mas não se limitam, a detecção de doenças nos estágios iniciais e, com isso: a otimização das doses de medicamentos evitando efeitos colaterais; fornecimento de medicina eficiente baseada em composições genéticas; monitoramento dos sinais vitais do paciente para fornecer atendimento proativo, sendo isto possível através de análise de dados de pacientes que já sofreram dos mesmos sintomas, ajudando o médico a receitar medicamentos eficazes; entre outros (ARCHENAA; ANITA, 2015).

De acordo com (ZHANG; LV, 2021) as organizações governamentais da China buscam melhorar e construir serviços digitais de qualidade para seus cidadãos através de inovação viabilizada por dados. A otimização desses serviços é um ponto-chave para elevar a confiabilidade da população no governo, melhorar modelos operacionais, favorecer a eficiência na tomada de decisões internas e a transparência da informação.

A utilização de análise preditiva de dados tem impacto também na segurança nacional. Para (MONTASARI, 2023), ela possibilita diferentes formas de atuação da polícia em suas operações através da utilização de vastos conjuntos de dados que são produzidos a partir da atividade humana e que, em conjunto com dados históricos de crimes, podem prever onde e quando os crimes são prováveis de ocorrer. Ainda segundo o autor, utilizando esses conjuntos de dados, as agências responsáveis pela aplicação da lei também poderiam prever quem tem maior probabilidade de perpetrar crimes e quem tem maior probabilidade de ser vítima de um crime.

A massiva quantidade de dados disponíveis na atualidade beneficia, também, a pesquisa e a ciência, pois gera novas percepções e agrega valor em testes e teorias de pesquisadores, como é indicado por (LEUNG et al., 2020). Os autores citam, entre outros, os artigos de (LEUNG; CARMICHAEL, 2010), (FARIHA et al., 2013) e (LEUNG; MACKINNON; TANBEER, 2014), para fundamentar e descrever a técnica chamada de mineração de dados; citam os trabalhos de (BARI; SAATCIOGLU, 2018), (AHN et al., 2019), (LEUNG et al., 2020) para descrever sobre aprendizado de máquina; e, por fim, citam (LEUNG, 2018) para modelagem matemática e estatística. Segundo (LEUNG et al., 2020), todas as técnicas mencionadas anteriormente neste parágrafo podem ser aplicadas a serviços da vida real e/ou para o bem social, como por exemplo, em 2020, a utilização abundante de dados epidemiológicos da doença respiratória aguda grave (conhecida por

coronavírus 2019 ou COVID-19), ter ajudado pesquisadores, epidemiologistas e formuladores de políticas a obter uma melhor compreensão da doença e, por sua vez, a encontrar maneiras de detectá-la, controlá-la e combatê-la (MELLO et al., 2020).

Como visto, grande é o potencial e importantes são os benefícios provenientes do uso de técnicas de extração e exploração de dados para geração de conhecimento através de dados abertos. Através da exploração de dados, diversas áreas como tecnologia, pesquisa e ciência, saúde, governamentais entre outras podem se beneficiar.

Apesar da notória importância da prática mencionada, durante a leitura é possível prever a existência inerente de alguns desafios técnicos, como a complexidade na coleta, tratamento e disponibilização de dados, e também subjetivos de maior importância, como consistência, veracidade, disponibilidade e privacidade da informação. O manuseio de dados, especialmente em grandes volumes, é objeto de estudo de pesquisa para diversos grupos, como o NIST Big Data Public Working Group (GRADY et al., 2014) que discute, entre outros tópicos, o estado da arte em tecnologias que melhoram a segurança, a privacidade de dados em si, e as preocupações com privacidade enquanto se busca derivar conhecimento dos dados.

Considerando o manuseio de dados abertos, a sua comprovada importância para a sociedade, ciência e pesquisa, além dos desafios inerentes a essa prática, é preciso tratar do que motivou este estudo. Como motivador para a pesquisa, esteve o foco na sociedade e sua evolução a partir de tecnologias para dados abertos, considerando a cautela, responsabilidade e fiscalizações necessárias para o seu processamento e exposição na internet. Dada a dimensão e relevância do manejo adequados de dados (especialmente dos dados abertos) foi desenvolvido um framework de apuração da qualidade de dados abertos, e verificação de sua conformidade em relação a leis de proteção e privacidade de dados pessoais, sendo está ultima parte, o diferencial do framework proposto.

1.1 Motivação

O Programa de Desenvolvimento das Nações Unidas (UNDP) fez um estudo (UNDP, 2023) em que explana que o Conselho de Direitos Humanos da ONU e a Assembleia Geral afirmaram que "os mesmos direitos que as pessoas têm offline também devem ser protegidos online" (ASSEMBLY, 2016). No entanto, o estudo afirma que permanecem muitas questões sobre as consequências sociais desta transformação digital e o seu impacto nos direitos humanos. À medida que as sociedades se tornam mais dependentes das tecnologias digitais, a proteção dos direitos humanos é cada vez mais crítica, tal como a utilização destas tecnologias no interesse público.

Os recentes desenvolvimentos das tecnologias da informação e comunicações (TIC) e dos poderosos algoritmos de inteligência artificial vão além do quadro legal existente, sendo necessário construir novos modelos que se concentrem na responsabilidade dos programas informáticos — que cada vez mais tomam decisões próprias — e na privacidade dos dados dos cidadãos, regulando a utilização não autorizada dos seus dados pessoais (SOKOLOVSKA; KOCAREV, 2018).

Sobre privacidade de dados, existem limites para a transparência, pois, como previsto na Legislação, nem todas as informações em posse ou tutela do Estado são de acesso público (BERTOLINI et al., 2022).

A Lei de Acesso à Informação Brasileira (LAI), ao garantir o acesso a informações públicas, não anula outras proteções previstas na legislação. Os principais casos de proteção de dados são: sigilos previstos em lei; segredo de justiça; e proteção de dados pessoais,

que são assegurados por dispositivos da lei geral de proteção de dados brasileira (LGPD) e têm restrições de acesso previstos na própria LAI.

O problema de pesquisa abordado é: no Brasil, o tratamento de dados pessoais, por pessoa jurídica de direito público — no contexto de dados abertos — está em conformidade com a LGPD? garantindo assim os direitos fundamentais de liberdade, privacidade e desenvolvimento da personalidade da pessoa natural?

Deste modo, o objetivo é realizar estudo sistemático de verificação de conformidade de dados abertos brasileiros com a LGDP. Para tal uma etapa prévia de análise da qualidade dos dados abertos no Brasil também foi performada, através de critérios definidos por iniciativas - globalmente aceitas - de controle e fomento de dados abertos, como a *Open Knowledge Foundation* (OKFN). O trabalho contribui com um framework, adaptável, de verificação de conformidade com leis de proteção de dados, não apenas brasileira (escopo deste trabalho), mas de qualquer portal de dados abertos que utilize CKAN em seus portais, e que tenham lei de proteção de dados definida, pois estes serão os parâmetros inseridos no modelo.

Em consequência, este trabalho contribui com meios para fiscalização à proteção e privacidade de dados pessoais em repositório de dados abertos governamentais no Brasil.

Para atingir esses objetivos, foi desenvolvido:

- o desenvolvimento de um framework para LGPD Compliance em dados abertos, cujo objetivo é verificar a conformidade LGPD em dados abertos, através das seguintes considerações da lei, no artigo 5º:
  - *"I - dado pessoal: informação relacionada a pessoa natural identificada ou identificável".* Ou seja, se os dados governamentais abertos publicados identificam ou trazem informações relacionadas a pessoas naturais;
  - *"XII - consentimento: manifestação livre, informada e inequívoca pela qual o titular concorda com o tratamento de seus dados pessoais para uma finalidade determinada"*;
- a análise quantitativa e qualitativa dos dados abertos no Brasil;
- a disponibilização do framework para fins de reprodutibilidade científica.

1.2 Organização do texto

Além das seções já apresentadas, este trabalho está organizado da seguinte forma: o Capítulo 2 de contextualização, que tem o objetivo de esclarecer o que são dados abertos, sua relevância, e interação em diferentes áreas; o Capítulo 3, de revisão da literatura, mostra os trabalhos relacionados em que este estudo se baseia, e expõe brevemente suas arquiteturas; o Capítulo 4 contém a estrutura da modelagem de dados utilizada na análise, o objetivo é mostrar o fluxo de desenvolvimento e execução dos passos utilizados no desenho do framework. A metodologia de desenvolvimento do framework é apresentada no Capítulo 5, com as tecnologias utilizadas, *endpoints*, as chamadas de API. Enfim, todas as implementações são explicadas através de diagramas e trechos de códigos. Também nele são explanados os resultados obtidos com os experimentos feitos. O Capítulo 6 apresenta a contribuição científica, os objetivos realizados, discute os desafios encontrados, e são debatidos oportunidades de evolução e limitações.

2 CONTEXTO

Neste capítulo será contextualizado o conceito de dados abertos, e será abordada sua aplicabilidade em diferentes meios e contextos, por exemplo: para órgão governamentais, qual a importância e uso de dados abertos? como estes órgãos podem se aperfeiçoar os utilizando?. Assim, os benefícios e desafios na utilização de dados abertos serão apresentados, baseados na literatura e em trabalhos relacionados. Por fim, serão apontadas as leis de acesso à informação e proteção de dados no Brasil.

2.1 Dados abertos

Abertura, por definição significa ausência de restrição ou sigilo, ou seja, é aquilo que é acessível. De acordo com (PETERS; BRITEZ, 2008), o conceito de abertura no que diz respeito à educação é anterior ao movimento de abertura que começa com software livre e código aberto em meados da década de 1980, com raízes que remontam ao iluminismo e que estão ligadas aos fundamentos filosóficos da educação moderna com os seus compromissos com a liberdade, cidadania, conhecimento para todos, progresso social e transformação individual.

O acesso aberto (ou em inglês, *Open Access*) é um conceito originalmente percebido no meio acadêmico, sendo definido por (SUBER, 2012) como sendo uma literatura de acesso aberto digital, online, gratuita e livre da maioria das restrições de direitos autorais e licenciamento, que surgiu para designar autores e pesquisadores que compartilhavam o resultado de suas pesquisas de modo livre, visando beneficiar e atingir uma maior número de pessoas.

A *abertura* é um conceito que passou a caracterizar sistemas de conhecimento e comunicação, epistemologias, sociedade e política, instituições ou organizações e personalidades individuais (PETERS; BRITEZ, 2008). Para os autores, em essência, a abertura em todas essas dimensões refere-se a um tipo de transparência que é o oposto do sigilo e, na maioria das vezes, esta transparência é vista em termos de acesso à informação, especialmente em organizações, instituições ou sociedades.

Seguindo o movimento de abertura em diferentes áreas, existe a prática chamada *Dados Abertos*. De acordo com (TEMIZ et al., 2022), dados abertos são inspirados em práticas conhecidas como inovação aberta (HARHOFF; LAKHANI, 2016), e código aberto (UEDA, 2005), que se baseiam na ideia de que os dados devem ser gerados e compartilhados livremente. Além disso, seu potencial depende das características essenciais do dado: os dados são cumulativos e combinatórios, e o valor derivado deles está, muitas vezes, exponencialmente relacionado com o tamanho do conjunto de dados. Ao reunir dados abertos, as organizações podem criar valor a um nível que nenhuma organização conseguiria sozinha.

A frase que mais se aproxima dos dados abertos vem da versão de Budapeste, que diz que o termo "acesso aberto", na referida literatura, diz respeito a sua disponibilidade gratuita na Internet pública, permitindo que qualquer usuário leia, descarregue, copie, distribua, imprima, pesquise, ou crie links para os textos completos desses artigos, rastrei para indexação, transmita como dados para software ou use-os para qualquer outra finalidade legal, sem barreiras financeiras, legais ou técnicas que não sejam aquelas inseparáveis do acesso à própria internet (MURRAY-RUST, 2008).

Os benefícios da abertura de dados para a sociedade, que serão aprofundados nas próximas seções, são: conhecimento e fiscalização dos gastos públicos (MCDERMOTT, 2010), levando à transparência governamental e, ainda, utilização de dados abertos do governo para a melhora dos serviços públicos prestados, como visto em (TANG et al., 2023) (ALMUSALAMI et al., 2022); desenvolvimento de aplicações em prol da comunidade (CHEN; JAKUBOWICZ, 2015) (GUO et al., 2019) (ZHANG; YUE, 2016) (CONSOLI et al., 2023) (CABEZUELO, 2020); uso na ciência e pesquisa de modo a desenvolver soluções e/ou planejamentos sustentáveis como trabalhos de: (PAREJA-LORA et al., 2019) (LIN et al., 2023) (PEDDI; DASGUPTA; GAIDHANE, 2022) (HO et al., 2021) e (RENZI et al., 2023).

Em contrapartida, para haver utilização efetiva dos dados, os desafios da abertura de dados são relacionados à rara promoção e conscientização desses benefícios (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012) (GEBKA et al., 2019), e à desatenção sobre a importância da padronização na abertura dos dados — durante a disponibilização e compartilhamento — para serem reutilizáveis e intercambiáveis em diferentes oportunidades. Outras dificuldades percebidas em relação aos dados abertos na educação, são: dispersão, licenciamento pouco claro, padronização insuficiente de dados, falta de incentivos e infraestrutura para compartilhamento de dados (MACHADO et al., 2019).

Esses desafios confirmam as principais ideias promovidas pela Open Knowledge Foundation (OKFN), uma organização da sociedade civil apartidária e sem fins lucrativos, ao definir o conceito de "aberto": o conhecimento deve ser visto como um bem comum, sendo possível para qualquer pessoa usar e participar de sua construção; e, informatizados ou não, os sistemas devem ser “interoperáveis”, o que significa maximizar a sua capacidade de comunicar de forma transparente e de se conectar com outros sistemas (FOUNDATION, 2023).

Apesar dos grandes desafios, os dados abertos têm um grande potencial, e existem diversas iniciativas que procuram promovê-los, como a OKFN, que apoia e orienta a utilização e partilha de dados. Através da OKFN são incentivadas iniciativas de padronização na disponibilização de dados e, atualmente, na página da fundação é possível verificar o CKAN (sistema de gerenciamento de dados) como uma ferramenta adequada para compartilhamento, disponibilização e busca de dados abertos. No site do CKAN, por sua vez, é possível verificar que países como Estados Unidos, Canadá, Alemanha e Austrália já utilizam essa ferramenta em seus portais de dados abertos.

2.1.1 Dados abertos e a ciência

De acordo com (PAREJA-LORA et al., 2019), nas últimas décadas a comunidade científica tornou-se cada vez mais consciente da importância da abertura – para software (código aberto), para publicações (acesso aberto), para estruturar dados (conhecimento aberto) e para coletas de dados em geral (dados abertos). O autor afirma que a publicação de dados científicos, sob recursos abertos, se tornou rotina na investigação moderna, e que, movimento de dados abertos em linguística – bem como em todas as áreas de estudo

da ciência, computação e humanidades – se baseia em três motivações principais: (1) responsabilidade, (2) reprodutibilidade e (3) reutilização, como segue:

1. O processo científico – a geração de novas ideias, o estabelecimento e revisão de paradigmas de pensamento e metodologias científicas, e a sua documentação, disseminação e reflexão crítica – é impulsionado pela necessidade social, econômica e ecológica de compreender e desenvolver o nosso passado, presente e futuro. Neste sentido, a investigação científica acarreta um privilégio e uma responsabilidade *(PAREJA-LORA et al., 2019)*: quaisquer projetos são apoiados por financiamento público e, em troca, os seus resultados devem (e de fato são muitas vezes obrigados a) ser disponibilizados ao público. Nas últimas décadas, isto contribuiu para o aumento do acesso aberto em publicações científicas e, juntamente com ele, para o licenciamento de código aberto de códigos e dados científicos.

2. Outra motivação para a crescente importância dos dados abertos na investigação é inerente ao método científico: as hipóteses científicas devem ser testáveis, as teorias científicas devem ser verificáveis e os resultados publicados devem ser replicáveis. Para disciplinas baseadas em dados, como ramos empíricos da linguística, a verificação pressupõe a disponibilidade de dados empíricos, enquanto a *replicabilidade* requer acesso aos dados originais nos quais a investigação se baseia. O autor finaliza a fundamentação afirmando que a publicação sob uma licença de código aberto apresenta a menor barreira possível para a reutilização, acessibilidade e disseminação de dados de pesquisa.

3. Uma terceira motivação prática para publicar (e utilizar) dados científicos é o imenso esforço colocado na criação de tais recursos e os ganhos potenciais da partilha e *reutilização* de dados existentes. Em diversas áreas da linguística, diz respeito a dados primários, tais como gravações, transcrições e textos escritos; como exemplo extremo, línguas à beira da extinção e/ou faladas em áreas remotas do mundo.

A complementação de dados abertos na ciência, e nas demais áreas, é a iniciativa conhecida como dados abertos vinculados (ou LOD, da sigla em inglês Linked Open Data). Através desta prática, é possível combinar conjuntos de dados heterogêneos que estão em diferentes silos, e gerar percepções de conhecimento mais completas.Quando conjuntos de dados estão em silos, estes impõem limitações para responder a questões complexas e interdisciplinares que exigem o estabelecimento de ligações de associação entre entidades e conceitos semelhantes que estão presentes em diferentes conjuntos de dados. Para estabelecer tais associações, a informação correspondente deve ser representada num formato acessível e sem ambiguidades quanto à sua estrutura e semântica (FOTOPOULOU et al., 2016).

Um notório projeto que apoia a integração de informações através dos princípios de dados vinculados (em inglês, *Linked Data*) (HYLAND et al., 2014) (BERNERS-LEE, 2006), é o DBpedia. O projeto DBpedia aproveita essa gigantesca fonte de conhecimento extraindo informações estruturadas da Wikipédia e tornando essas informações acessíveis na Web (BIZER et al., 2009). Segundo os autores, a base de conhecimento resultante da DBpedia descreve atualmente mais de 2,6 milhões de entidades, incluindo 198 mil pessoas, 328 mil lugares, 101 mil obras musicais, 34 mil filmes e 20 mil empresas, e contém 3,1 milhões de links para páginas externas e 4,9 milhões de links RDF para outras fontes de dados da Web. Para eles, a base de conhecimento DBpedia tem diversas vantagens sobre as bases de conhecimento existentes: abrange muitos domínios, representa

um acordo real da comunidade, evolui automaticamente à medida que a Wikipédia muda, é verdadeiramente multilíngue e é acessível na Web.

Iniciativas como a "Nuvem de Dados Abertos Vinculados"(em inglês *LOD Cloud*) surgem, como o previamente mencionado DBpedia, e sua visualização — entre outros LOD Cloud datasets — pode ser acessado em diagramas de visualização online, como o disponível em https://lod-cloud.net/, que mostra datasets publicados no formado de dados vinculados. Em novembro de 2023, o diagrama continha 1.314 conjuntos de dados com 16.308 links, como pode ser visto na Figura 2.1.1 adapatada de (MCCRAE, 2017). Os trabalhos de (IM et al., 2014) e (MARTíN-MONCUNILL; ALONSO GAONA GARCíA; RAJABI, 2015) mostram como navegar em LOD clouds como o DBpedia.

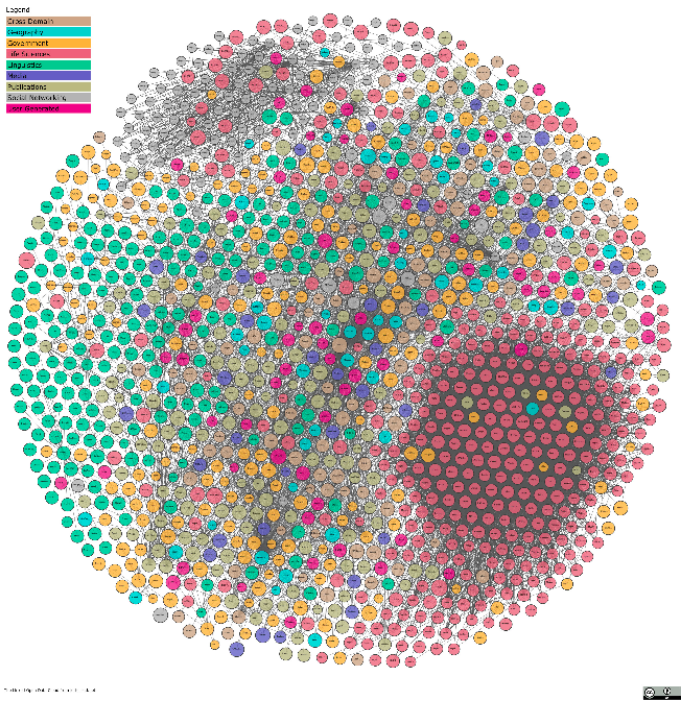


Figura 2.1: Diagram Linked Open Data Cloud (Adaptado de (MCCRAE, 2017))

2.1.2 Dados abertos e o governo

Dados Governamentais Abertos (da sigla em ingles *OGD*) é um conceito que visa envolver ao máximo os cidadãos na utilização ou reutilização de dados governamentais (RAHMATIKA et al., 2019). Os dados abertos do governo podem ser usados para ajudar o público a compreender melhor o que o governo faz e o seu desempenho, e para responsabilizá-lo por irregularidades ou resultados não alcançados (UBALDI, 2013). Além disso, (UBALDI, 2013) acredita na sensibilização do público para os programas e atividades governamentais através do aumento da transparência governamental, fornecendo a base para a participação e colaboração na criação de serviços inovadores e de valor agregado.

Diferentes abordagens são executadas para entender, aprimorar e incentivar o uso de OGD, como: (a) revisões sistemáticas de literatura, como as de (BACHTIAR; SUHARDI; MUHAMAD, 2020) e (ALI HASSAN; TWINOMURINZI, 2018), cujo objetivo é entender as tecnologias utilizadas no OGD, bem como as dificuldades de implementação; (b) modelos de validação de maturidade em implementações OGD (RAHMATIKA et al.,

2019); (c) metodologias de desenvolvimento de indicadores de dados abertos com a finalidade de medir a qualidade dos OGD como visto em (DOROBăț; POSEA, 2021); (d) proposição de métricas para melhorar a avaliação da qualidade dos dados (BOUCHE-LOUCHE; GHOMARI; ZEMMOUCHI-GHOMARI, 2022); (e) análise de uso de conjunto de dados OGD (ZAINAL et al., 2019); (f) e promoção do uso de LDO em OGD (FLEINER, 2018). Estas iniciativas tanto comprovam a relevância da prática, quanto demonstram os desafios da implementação, uma vez que a conclusão dos trabalhos citados, em sua maioria, indicou oportunidades de melhoria e sugestões de evolução em seus escopos de pesquisa.

No artigo de (QANBARI; REKABSAZ; DUSTDAR, 2015) é evidenciado o potencial da utilização de dados abertos governamentais em prol da sociedade, inclusive utilizando-os como serviço (DUAN et al., 2015). Os autores apresentam uma abstração que combina URL de dados governamentais e sua interface associada, chamada DCU, para permitir dados governamentais abertos como serviço (o DCU também incorpora a política governamental) e sua utilização em aplicativos por programadores (Figura 2.2 adapatada de (QANBARI; REKABSAZ; DUSTDAR, 2015)). Com base na DCU, os autores ampliam a discussão para a oportunidade de uso civil dos dados governamentais e suas potenciais vantagens, e, por fim, complementam que o objetivo é expor mecanismos que permitirão aos desenvolvedores publicar aplicativos em uma loja de aplicativos, além de conjuntos de dados governamentais.

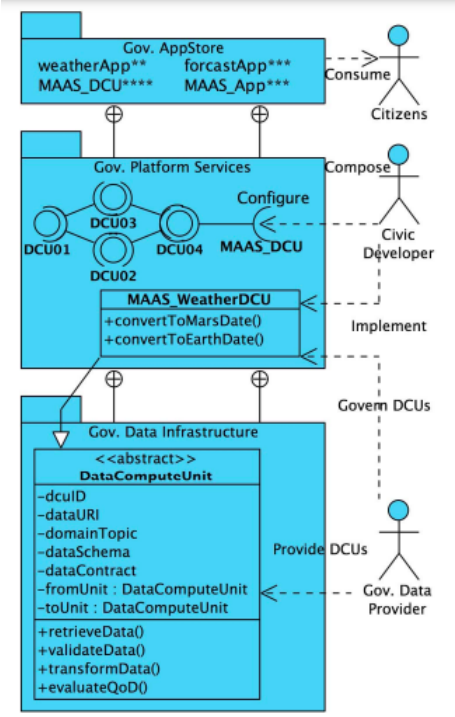


Figura 2.2: atores do GoDaaS e seus relacionamentos (Adaptado de (QANBARI; REKABSAZ; DUSTDAR, 2015))

A Forbes, renomada revista de negócios e economia, informou, em 2020, que o uso de dados abertos é essencial para projetos que buscam criar cidades inteligentes (ARBEX, 2020). Em suma, as cidades inteligentes podem ser entendidas como soluções tecnológicas — aplicadas nas cidades, áreas rurais e regiões vizinhas — a fim de construir ambientes abertos sustentáveis e ecossistemas de inovação orientados para os utilizadores

(DOMINGUE et al., 2011). Ainda segundo a revista Forber, Londres, país que criou o primeiro armazenamento de dados abertos do mundo (o London Datastore) e, também, pioneiro na adoção de cidades inteligentes, tinha, na época, o objetivo de sofisticar seu compartilhamento de dados públicos nos anos seguintes, com a finalidade de solucionar problemas gerados pelo crescimento populacional nos centros urbanos, como o mostrado por (CHANG; JANG, 2019) em seu artigo, e também por (MAYAUD; TRAN; NUTTALL, 2019).

Uma cartilha completa com detalhes sobre benefícios do *Governo Aberto*, e de iniciativas multilaterais internacionais, como o *Open Government Partnership (OGP)*, ou em português, *Parceria para Governo Aberto*, pode ser acessada em (BERTOLINI et al., 2022).

2.1.3 Dados abertos e organizações

Para empresas, públicas ou privadas, a abertura de dados e cooperação é vista de modo mais cauteloso devido à competitividade entre elas, por exemplo, no trabalho de (TEMIZ et al., 2022) é sugerido considerar o cenário do grande impacto ambiental negativo das compras online de roupas e moda. Segundo os autores, para alguns tipos de produtos, até 80% dos pedidos são devolvidos, muitas vezes porque os consumidores têm dificuldade em selecionar o tamanho certo quando fazem compras online. Ele explica que uma startup norueguesa mitigou o problema usando o histórico de compras de indivíduos, revendedores e marcas para criar gêmeos digitais para partes do corpo dos consumidores. A solução, segundo o artigo, aumentou significativamente a probabilidade de encomendar o tamanho certo, quando os clientes compram roupas que lhes cabem melhor. No entanto, os autores ressaltam que, apesar da simplicidade e conveniência da tecnologia, o sucesso dependeu da abertura e partilha de alguns dos dados das partes citadas por vários intervenientes ao longo da cadeia de valor, por vezes até com concorrentes.

Uma revisão sistemática da literatura foi feita por (ÇALDAğ; GöKALP, 2023) e mostrou resultados importantes sobre as barreiras que impedem a adoção e o uso de dados abertos em organizações. Segundo os autores, 3 dimensões dificultam a implementação dos dados abertos: (1) dimensão técnica, (2) dimensão organizacional e (3) dimensão ambiental. Alguns pontos serão citados a seguir, porém, todos os itens mencionados podem ser vistos na Figura 2.3 adaptada de (ÇALDAğ; GöKALP, 2023), e em detalhes no artigo dos autores.

1. Ausência de vantagem ou benefício relativo, pois, segundo o artigo, embora a adoção de dados abertos nas organizações tenha benefícios economicos, políticos, sociais e operacionais, o valor pouco claro dos conjuntos de dados apresenta uma barreira significativa; a complexidade dos conjuntos de dados, a má compreensão dos dados e a ausência de guias sobre acessibilidade aos conjuntos de dados são várias barreiras de complexidade técnica encontradas e a falta de compatibilidade/interoperabilidade entre conjunto de dados são alguns pontos citados;

2. A barreira de recursos financeiros nas organizações; a escassez de recursos humanos e de competências é vista como uma das principais barreiras à adoção do OAD no contexto da organização; estrutura organizacional engessada, afetando a inovação e cultura;

3. O trabalho de (ÇALDAğ; GöKALP, 2023) cita (KHURSHID et al., 2020), para fundamentar que fatores de desenvolvimento nacional podem ser barreiras à im-

plementação de dados abertos, pois os autores propõem que os níveis de educação dos cidadãos, o produto interno bruto, o índice de inovação global e o tamanho da população da cidade podem influenciar as decisões de adoção do OAD e o processo de publicação. Outros fatores como política e regulamentos, e comprometimento político também são citados.

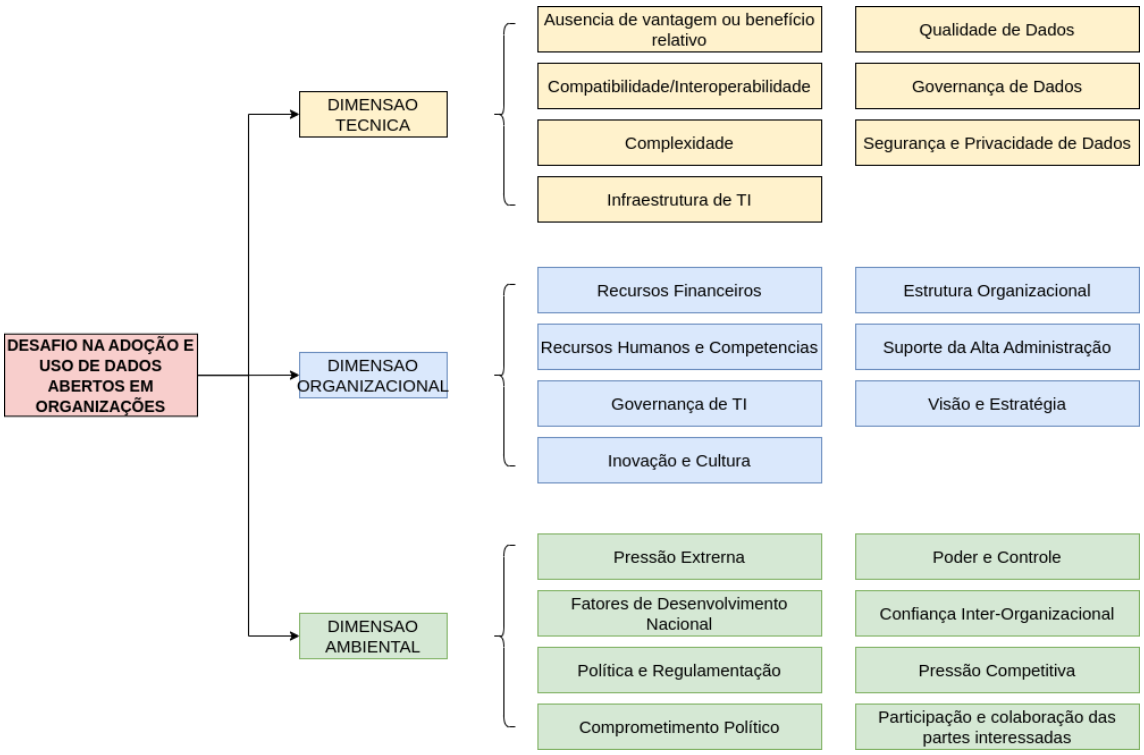


Figura 2.3: barreiras que afetam a adoção e utilização do ADO no contexto das organizações (Adaptado de (ÇALDAğ; GöKALP, 2023))

2.2 Leis de acesso à informação e proteção de dados no Brasil

A lei de acesso à informação (LAI) brasileira, L12527/2011, entrou em vigor em 18 de maio de 2012, e, para (ANGELI, 2016), 4 anos após entrar em vigor, prenunciava que uma grande mudança na "cultura da informação"havia se imposto às instituições e aos servidores públicos brasileiros, pois o sigilo, que antes era regra, seria agora é exceção, em substituição à tradicional confidencialidade das informações produzidas ou mantidas pela administração pública, como se privadas fossem, institui-se a abertura de informações a todo e qualquer cidadão, independentemente dos motivos determinantes da solicitação.

A LAI surge como uma importante ferramenta de controle social sobre as políticas públicas do país, pois o controle social emerge da necessidade de acompanhamento direto das ações de governo, que tem por objetivo coibir práticas de corrupção e contribuir para aproximar a sociedade do Estado, abrindo a oportunidade para os cidadãos fiscalizarem as ações dos governos, assim como os seus gastos, e exigirem uma boa gestão pública (CRUZ, 2022).

Além da LAI, outras leis brasileiras buscam aumentar a eficiência pública na era digital e a participação do cidadão através da transparência enquanto disponibilizam dados

governamentais, como, por exemplo a Lei 14129 de 2021 — conhecida como Lei do Governo Digital —, que reforça obrigações de transparência e o direito de se solicitar bases de dados de acesso público, e concede "permissão irrestrita de uso de bases de dados publicadas em formato aberto", ou seja, incentiva a utilização de dados abertos em seu artigo 3º, incisos XIV e XXV. E, ainda, a Lei complementar 131 de 2009, que prevê a disponibilização, em tempo real, de informações pormenorizadas sobre a execução orçamentária e financeira da União, dos estados, do Distrito Federal e dos municípios.

A política de dados abertos no Brasil define regras para promover a abertura de dados governamentais no âmbito dos órgãos e entidades federais e tem como pilar as disposições da LAI, sendo constituída por uma série de documentos normativos, que tratam de obrigações, planejamento e orientações, em especial o Decreto nº 8.777 de 2016 e a Resolução nº 3 de 2017, do Comitê Gestor da INDA (CGINDA). O órgão responsável pela gestão e monitoramento da política é a Controladoria-Geral da União (CGU), por meio da Infraestrutura Nacional de Dados Abertos (INDA) (GOV.BR, 2023).

Os 4 itens abaixo (de um total de 8 acessáveis em (GOV.BR, 2023)), expostos na política de dados abertos brasileira, vem de modo a retificar tendências mencionadas na Seção 2.1, que fala sobre LOD, dados abertos na ciência e os benefícios à sociedade:

- aprimorar a cultura de transparência pública;
- facilitar o intercâmbio de dados entre órgãos e entidades federais e as diferentes esferas da federação;
- fomentar o controle social e o desenvolvimento de novas tecnologias destinadas à construção de ambientes participativos e democráticos e à melhor oferta de serviços públicos para o cidadão;
- fomentar a pesquisa científica de base empírica sobre a gestão pública.

A Lei geral de proteção de dados (13.709/2018) (LGPD) tem como principal objetivo proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural, e tem como foco a criação de um cenário de segurança jurídica com a padronização de regulamentos e práticas para promover a proteção aos dados pessoais de todo cidadão que esteja no Brasil, de acordo com os parâmetros internacionais existentes (FEDERAL, 2023).

Para a LGDP, dado pessoal é a informação relacionada à pessoa natural identificada ou identificável. De acordo com o Supremo Tribunal Federal (STJ), exemplos de dados pessoais são: nome, RG (é um documento de identificação civil brasileiro), CPF (Cadastro de Pessoas Físicas), gênero, data e local de nascimento, telefone, endereço residencial, localização via GPS, foto, prontuário de saúde, cartão bancário, renda, histórico de pagamento, hábitos de consumo, preferências de lazer, endereço de IP, cookies, entre outros (STF, 2023). Além disso, a lei traz o conceito de dado pessoal sensível, que diz respeito a origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural.

O artigo segundo da LGPD diz que a disciplina da proteção de dados pessoais tem como fundamentos:

- o respeito à privacidade;

- a autodeterminação informativa;
- a liberdade de expressão, de informação, de comunicação e de opinião;
- a inviolabilidade da intimidade, da honra e da imagem;
- o desenvolvimento econômico, tecnológico e a inovação;
- a livre iniciativa, a livre concorrência e a defesa do consumidor;
- os direitos humanos, o livre desenvolvimento da personalidade, a dignidade e o exercício da cidadania pelas pessoas naturais.

Cabe ressaltar que a própria LAI restringe o acesso a informações pessoais de forma a garantir o respeito à intimidade, à vida privada, à honra e à imagem das pessoas, bem como às liberdades e garantias individuais (Art. 31). E a LGPD também restringe o acesso a dados pessoais, e, em paralelo, cria regras para o tratamento desses dados. Entre as regras da LGPD está a transparência nas relações entre quem fornece o dado e quem o coleta ou utiliza.

2.3 Proteção e privacidade de dados

De acordo com (CAMENISCH; FISCHER-HüBNER; RANNENBERG, 2011) o termo privacidade tem sido discutido há décadas por diferentes pessoas em diferentes ocasiões, mas com um significado (ligeiramente) diferente em mente. No entanto, para os autores, incontestável que a privacidade visa a proteger a autonomia das pessoas, em primeiro lugar. E aceitam a definição de Westin (PRIVACY AND FREEDOM, 1969) como amplamente aplicável: *"Privacidade é a reivindicação de indivíduos, grupos ou instituições de determinar por si próprios quando, como e em que medida as informações sobre eles são comunicadas a outros"*((PRIVACY AND FREEDOM, 1969). grifo dos autores).

As tecnologias da informação e comunicações (TIC) estão integradas de forma ubíqua nas nossas economias e sociedades, trazendo benefícios e desafios, e, à medida que avançam para uma implantação mais ampla, especialistas técnicos, analistas políticos e especialistas em ética levantam preocupações relativamente às consequências não intencionais e indesejadas da sua adoção generalizada (SOKOLOVSKA; KOCAREV, 2018).

Um exemplo do quão longe uma tecnologia pode ir em termos de privacidade de informações pessoais pode ser verificado no estudo de (KOSINSKI; STILLWELL; GRAEPEL, 2013), baseado no mecanismo de associação positiva — *like* — em conteúdos da rede social *Facebook*, e demonstra até que ponto registros digitais relativamente básicos do comportamento humano podem ser usados para estimar de forma automática e precisa uma ampla gama de atributos pessoais que as pessoas normalmente presumiriam serem privados, e, também, como a pevisão de informações pessoais para melhorar produtos, serviços e direcionamento pode levar a perigosas invasões de privacidade.

Um dos resultados do estudo mostra que através da análise de *likes* foi possível alcançar uma precisão na predição de variáveis dicotômicas, nas palavras dos autores, quase perfeita. A Figura 2.4, adaptada de (KOSINSKI; STILLWELL; GRAEPEL, 2013), mostra a acurácia da predição expressa em termos da área sob a curva característica de operação do receptor, ou curva de ROC — sendo equivalente à probabilidade de classificar corretamente dois usuários selecionados aleatoriamente, um de cada classe (por exemplo, homem e mulher). Cada barra lateral na Figura 2.4, mostra a porcentagem de precisão

alcançada pelo modelo, após a análise de likes. A maior precisão foi alcançada para origem étnica e gênero; os afro-americanos e os caucasianos americanos foram classificados corretamente em 95% dos casos e homens e mulheres foram classificados corretamente em 93% dos casos, sugerindo que os padrões de comportamento online expressos por curtidas diferem significativamente entre esses grupos, permitindo uma classificação quase perfeita.

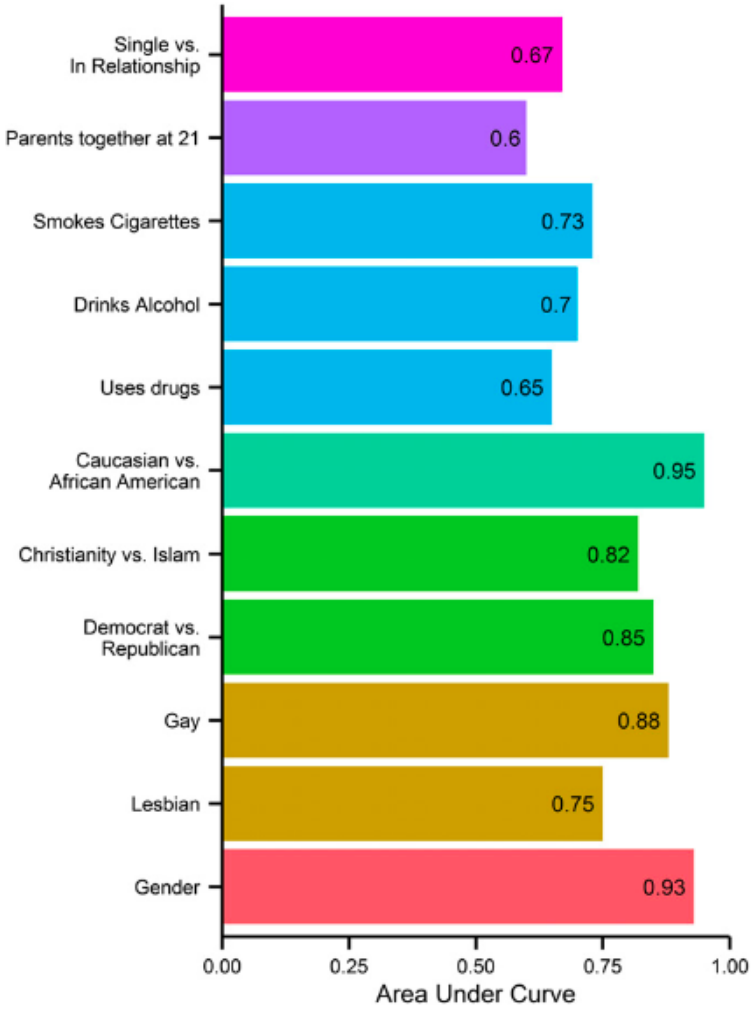


Figura 2.4: precisão de previsão da classificação para atributos dicotômicos expressos pela AUC.(Adaptado de (KOSINSKI; STILLWELL; GRAEPEL, 2013))

Uma das maiores preocupações em nossas sociedades é a privacidade, e esta foi identificada como um desafio político, regulamentar e legislativo fundamental do século XXI, sendo o âmbito da privacidade alargado para cobrir diferentes aspectos, incluindo o controle sobre a informação, a dignidade humana, a intimidade e as relações sociais. Além disto, a manipulação, coleta e utilização não autorizada de dados pessoais levantam importantes questões éticas e de privacidade (SOKOLOVSKA; KOCAREV, 2018).

2.4 Considerações finais

Os benefícios dos dados abertos são vastos e variados. E incluem o aumento da transparência governamental, a melhoria da prestação de serviços públicos, o estímulo

à inovação e ao empreendedorismo, a promoção da participação cívica e da governança democrática. Em contrapartida, a relação entre dados abertos e proteção de dados pessoais é complexa. De modo que a busca por um equilíbrio adequado entre transparência e privacidade requer uma abordagem cuidadosa e baseada em políticas bem definidas, como a observação das leis mencionadas previamente, tecnologias de segurança robustas e verificação contínua dos princípios éticos e legais envolvidos.

3 REVISAO DA LITERATURA E TRABALHOS RELACIONADOS

Neste capítulo, será destacada a metodologia de revisão da literatura utilizada e também serão sumarizados os trabalhos relacionados a: framework, plataforma ou metodologia de exposição de *Open Data*; e frameworks de validação de conformidade de dados em relação a leis de proteção de dados. O contexto é a disponibilização e gerenciamento do dado, de modo a viabilizar as análises com acurácia, otimização de recursos e dinamicidade de publicação e consumo dos dados abertos, além, conformidade com leis de proteção de dados.

Foi feita a subdivisão dos trabalhos relacionados em dois grupos: (1) escopo: plataformas ou meios de disponibilizar *Open Data* e (2) escopo: frameworks de conformidade com leis de proteção de dados — no Brasil e no mundo —, pois não foi possível encontrar, na literatura, abordagens que incorporassem ambos os assuntos num único trabalho de pesquisa, ou seja, framework que, além de disponibilizar dados abertos, tivesse módulo para verificação de conformidade com LGPD.

3.1 Método para revisao sistemática

A revisão sistemática da literatura foi desenvolvida baseada na abordagem de planejamento de (DYBA; DINGSOYR; HANSSEN, 2007) para condução da revisão. Esta abordagem diz que a revisão deve conter a identificação da necessidade da pesquisa e, também, o desenvolvimento do protocolo de revisão, devendo especificar questões de pesquisa, estratégia de busca, critérios de inclusão, exclusão e qualidade, extração de dados e métodos de síntese.

3.1.1 Identificação da necessidade e estratégia de busca

O objetivo do estudo é buscar por frameworks de dados abertos que tenham conformidade com leis de proteção de dados vigentes. Para conseguir isso, as questões de pesquisa foram elencadas na Tabela 3.1:

Acompanhando a abordagem de (DYBA; DINGSOYR; HANSSEN, 2007), a busca sistemática se deu pela identificação de palavras-chave e termos de busca que foram construídos a partir das questões de pesquisa, disponíveis na Tabela 3.1. Logo, as strings de busca utilizadas foram as elencadas na Tabela 3.2, onde é possível referenciar, através da coluna "ID"a questão de pesquisa atrelada.

Para cobrir a maior variedade de publicações relevantes, as bibliotecas eletrônicas consultadas foram as listadas na Tabela 3.3:

Tabela 3.1: questão de Pesquisa

ID	Question
RQ1	Quais plataformas ou framework de exposição de dados abertos existem?
RQ2	Quais plataformas ou framework de exposição de dados abertos existentes possuem meios de verificação de conformidade com leis de proteção de dados?
RQ3	Quais metodologias de exposição de dados abertos existentes possuem meios de verificação de conformidade com a LGPD?
RQ4	Quais metodologias de exposição de dados abertos existentes possuem meios de verificação de conformidade com a GDPR?
RQ5	Qual é o estado dos dados abertos no Brasil?

Tabela 3.2: string de busca em periódicos

ID	Strings de Busca
RQ2	"open data"AND "data protection law"
RQ2	"open data"AND "data protection framework"
RQ1 ǀ RQ5	"open data framework"OR "open data platform"
RQ3	"lgpd framework"
RQ4	"gdpr framework"

Tabela 3.3: bases de dados dos artigos

Biblioteca Eletrônica
Science Direct – Elsevier
Digital ACM Library
IEEE Explore Digital Library
Springer Link

3.1.2 Seleção e classificação dos estudos

Os critérios de inclusão e exclusão de trabalhos relacionados, podem ser observados a seguir:

- Inclusão
  - foco nos desafios e importância da proteção de dados;
  - conter metodologia para abertura de dados observando conformidade com LGPD;
  - conter relevância e benefícios a sociedade da abertura de dados observando a proteção de dados;
  - janela de busca dentro dos últimos 5 para metodologias de exposição de dados e desafios encontrados, dada a rápida evolução inerente ao meio da tecnologia da informação e seus procedimentos;
  - para a string de busca *"open data framework"OR "open data platform"*, de modo a manter o escopo de desenvolvimento do protótipo desta pesquisa, foi considerada a disciplina "ciência da computação"(*Computer Science*) quando disponível na biblioteca eletrônica.

- Exclusão
  - menciona os termos de busca, mas não mantém o foco nas perguntas da pesquisa, especialmente em relação a dados abertos;
  - frameworks que não tenham a finalidade de exposição/disponibilização de dados abertos, como, por exemplo, frameworks de análise de dados abertos;
  - que citem framework (em português 'estrutura'), porém, no escopo de estrutura conceitual. Que designem, por exemplo, "boas práticas", "guias"de implementação teóricos, ou, estruturas de "entrevistas com usuários", como visto nos estudos de (FORGÓ et al., 2021) e (RHAHLA; ALLEGUE; ABDELLATIF, 2021). Logo, não sendo no escopo de desenvolvimento de software;
  - foco em dados abertos de modo geral (seja na utilização de dados abertos, ou plataformas/frameworks de dados abertos);
  - disserta sobre a metodologia de exposição de dados abertos de outrem;
  - Disserta sobre proteção de dados, mas não disponibiliza framework — no contexto de desenvolvimento de software — para conformidade de alguma lei de proteção de dados (seja brasileira ou não);
  - o idioma de escrita do artigo não ser português ou inglês;
  - possuírem publicação anterior aos últimos 5 anos;
  - não possuírem acesso aberto ou conveniado com Universidade Federal do Rio Grande do Sul (UFRGS).

Os termos de busca serão aplicados aos resumos, palavras-chave do autor e títulos nas bases de dados eletrônicas identificadas. Serão considerados apenas tipos de conteúdo expostos a seguir:

- capítulos de livros (*chapter* ou *book chapter*);
- revistas acadêmicas (*journals*) — quando as bibliotecas eletrônicas tiverem esta opção — caso a base eletrônica não tenha a modalidade (*journals*), será utilizada a opção "trabalho de referência"(*reference work*);
- artigos de pesquisa (*research articles*) — caso a biblioteca não tenha artigos de pesquisa, serão considerados os artigos (*articles*);
- artigos de conferências (*conference paper*), quando não houver (*journals*) ou (*research articles*) disponíveis.

3.1.3 Resultados

De acordo com os critérios de inclusão/exclusão mencionados anteriormente, o resultado geral obtido por biblioteca e string pode ser visto na Tabela 3.4. A verificação da relação com o tema do trabalho se deu pela leitura do resumo do artigo e, quando necessário, pela metodologia utilizada.

O montante de artigos analisados totalizou 294, resultando nos trabalhos relacionados — seja por conformidade ou por framework de exposição — expostos na Tabela 3.5. Para cada string os resultados foram:

Tabela 3.4: artigos por biblioteca e string

Biblioteca Eletronica	String de Busca	Qtde.
Science Direct – Elsevier	*"open data"AND "data protection law"*	27
Science Direct – Elsevier	*"lgpd framework"*	0
Science Direct – Elsevier	*"gdpr framework"*	35
Science Direct – Elsevier	*"data protection framework"AND "open data"*	9
Science Direct – Elsevier	*"open data framework"OR "open data platform"*	8
Digital ACM Library	*"open data"AND "data protection law"*	28
Digital ACM Library	*"lgpd framework"*	0
Digital ACM Library	*"gdpr framework"*	4
Digital ACM Library	*"data protection framework"AND "open data"*	1
Digital ACM Library	*"open data framework"OR "open data platform"*	7
IEEE Explore Digital Library	*"open data"AND "data protection law"*	17
IEEE Explore Digital Library	*"lgpd framework"*	0
IEEE Explore Digital Library	*"gdpr framework"*	1
IEEE Explore Digital Library	*"data protection framework"AND "open data"*	0
IEEE Explore Digital Library	*"open data framework"OR "open data platform"*	20
Springer Link	*open data AND data protection law*	67
Springer Link	*"lgpd framework"*	0
Springer Link	*"gdpr framework"*	18
Springer Link	*"data protection framework"AND "open data"*	5
Springer Link	*"open data framework"OR "open data platform"*	47

– string *"open data"AND "data protection law"*; após a análise inicial, apenas 21 possíveis trabalhos tinham relação com a proteção de dados, porém, não no modelo do framework proposto neste trabalho, ou seja, não estavam em conformidade com leis de proteção de dados em dados abertos. Após a leitura completa dos artigos 21 artigos, apenas 3 artigos propunham um framework — no contexto de software — que fizesse a verificação de conformidade de dados já publicados, em relação a algumas leis de proteção de dados vigentes, porém nenhum relacionado a *dados abertos*. No entanto, é notável a relevância no contexto de proteção de dados em trabalhos como: a) (COMANDè; SCHNEIDER, 2021) exemplificando como a GDPR funciona como aliada (e não opositor) à abertura de dados; b) (von Grafenstein; JAKOBI; STEVENS, 2022) propondo uma metodologia de pesquisa empírica para especificação de propósito eficaz, usando métodos de design UX, para proteção de dados efetiva; e c) (PHILLIPS, 2021) mostrando os desafios de ficar em conformi-

dade com a GDPR no ambiente de educação continuada. Os 3 artigos mencionados inicialmente serão melhor explorados na Seção de trabalhos relacionados;

– string *"lgpd framework"*: nenhum resultado foi retornado;

– string *"gdpr framework"*: 57 artigos foram retornados, e após a análise inicial, apenas 2 de acordo com escopo do trabalho serão analisados na Seção de trabalhos relacionados;

– string *"data protection framework"and "open data"*: retornou 15 artigos, e após a análise inicial, apenas 2 de acordo com o escopo deste trabalho;

– a string *"open data framework"OR "open data platform"*: 82 artigos foram retornados, e após a análise inicial, apenas 8 de acordo com escopo do trabalho serão analisados na Seção de trabalhos relacionados.

3.2 Disponibilização de dados abertos

As plataformas de código aberto, também conhecidas como *Open Source*, facilitaram enormemente o trabalho das instituições envolvidas em iniciativas de Dados Abertos, tornando a configuração de portais de Dados Abertos uma tarefa quase trivial (NOGUERAS-ISO et al., 2021). O objetivo desta seção, é expor os trabalhos relacionados - no âmbito de plataformas de código aberto - para open data, uma vez que o acesso aos dados nestes portais, é realizado através destas plataformas.

3.2.1 SODAS

Smart Open Data as a Service (SODAS) é uma plataforma de dados abertos para compartilhamento e utilização de dados abertos (WON et al., 2021). A plataforma proposta foi desenvolvida com base em lacunas existentes no CKAN, através das seguintes estratégias core: expansão CKAN, suporte a *Vocabulário do Catálogo de Dados versão 2* (DCATv2) e *Mapeamento de Dados* (DataMap) extensível. Os problemas endereçados foram:

- limitação de gerenciamento de dados. Por causa de limitações funcionais do CKAN, plug-in de extensão adicional e trabalhos de instalação e gerenciamento são necessários;
- nenhum recurso em tempo real. Como o CKAN e plug-ins de extensão não suportam coleta e gerenciamento de dados em tempo real, há restrições nos domínios e formatos de dados que podem ser compartilhados;
- falta de metadados. O CKAN usa DCAT como um padrão de catálogo de dados. No entanto, o CKAN não explora a versão mais recente do DCAT e, portanto, limita os metadados que podem ser definidos;
- sem padrão de interconexão. A plataforma existente não possui guia de gerenciamento de metadados para publicação e distribuição de dados, reduzindo assim a interoperabilidade entre plataformas de dados abertos;
- baixa utilização. O CKAN, que se concentra principalmente no gerenciamento e recuperação de dados, tem aplicações limitadas.

Tabela 3.5: breve resumo dos trabalhos selecionados

No.	Autor e ano	Breve resumo
1	(FAZZINGA; GALASSI; TORRONI, 2022)	Propõe uma abordagem baseada em uma estrutura de argumentação computacional. A abordagem garante que os dados do usuário sejam gerenciados de acordo com a minimização de dados, limitação de finalidade e integridade.
2	(AHMAD; AUJLA, 2023)	Propõe uma abordagem para verificação de conformidade legal na Web Semântica que pode ser efetivamente utilizada para aplicações no ambiente Linked Open Data.
3	(FRANCESCONI; GOVERNATORI, 2023)	Propõe uma abordagem para verificação de conformidade legal na Web Semântica que pode ser efetivamente aplicada para aplicações no ambiente Linked Open Data.
4	(PIAO et al., 2019)	Propõe uma estrutura de privacidade diferencial para publicação de dados estatísticos governamentais baseados em fog computing.
5	(KIRSTEIN; BOHLEN, 2022)	Propõe uma arquitetura para superar as desvantagens dos dados abertos como: usabilidade, qualidade, barreiras legais, de privacidade, estratégicas e técnicas, utilizando os conceitos, especificações e tecnologias fornecidas pelos Espaços Internacionais de Dados.
6	(ESCOBAR et al., 2020)	Propõe uma abordagem para publicar dados estatísticos de repositórios públicos utilizando padrões de Web Semântica publicados pelo W3C, como RDF e SPARQL, a fim de facilitar a análise de modelos multidimensionais.
7	(AYDIN; AYDIN, 2020)	Desenvolvimento de modelo de aquisição de dados baseado em ontologia para plataformas de dados agrícolas abertos e implementação da ferramenta OWL2MVC.
8	(KIRSTEIN et al., 2020)	Propõe uma solução chamada 'Piveau', uma solução completa de gerenciamento de dados abertos, baseada em tecnologias da Web Semântica.
9	(ABIDI et al., 2019)	Uma abordagem de governança de segurança de serviços da Web baseada em microsserviços dedicados.
10	(SáNCHEZ-NIELSEN et al., 2021)	Framework de gerenciamento de dados governamentais abertos (na sigla em inglês OGD) sustentável para publicação e consumo de longo prazo.
11	(KIRSTEIN et al., 2021)	Plataforma de código aberto para coletar, processar e publicar dados abertos em tempo real com base em big data e ferramentas de processamento de dados.
12	(WON et al., 2021)	Plataforma de dados abertos para compartilhamento e utilização de dados abertos. A plataforma proposta foi desenvolvida com base em lacunas existentes no CKAN.

Então, de modo resumido, o SODAS resolve os problemas itemizados acima através da expansão do CKAN. Utiliza o DCATv2 para solucionar os problemas melhorando: o dado; o serviço de qualidade; definindo um sistema de gerenciamento de metadados; e, por fim, fazendo um mapeamento de dados extensível, melhorando assim a interoperabilidade entre plataformas. Na Figura 3.2, adpatada de (WON et al., 2021), está a visualização de alto nível da plataforma:

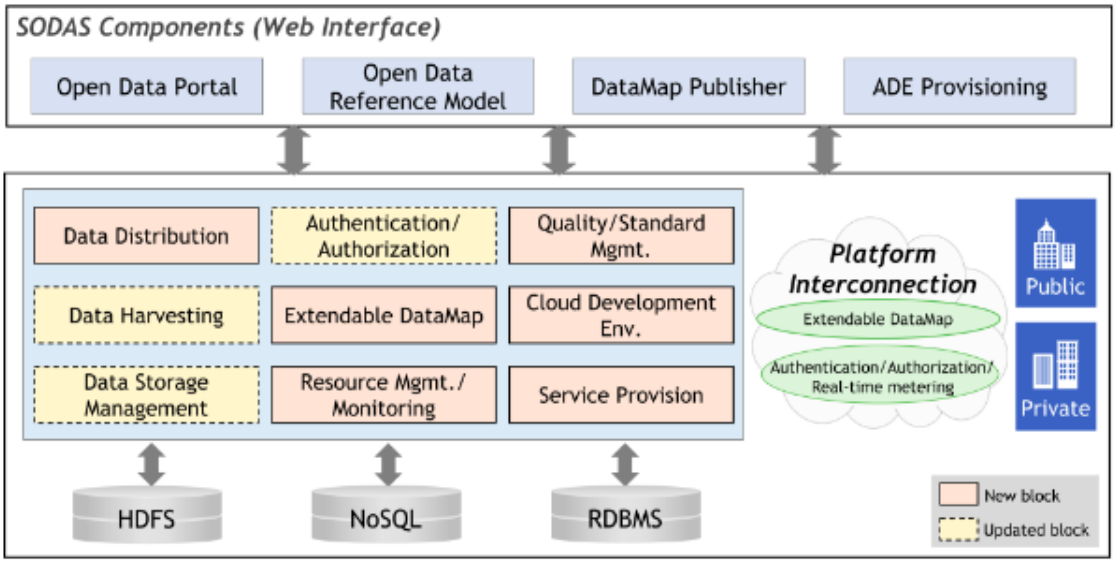


Figura 3.1: Estrutura geral do SODAS. (Adaptado de (WON et al., 2021))

3.2.2 SuDaMa

Do acrônimo em inglês *Sustainable Open Government Data Management Framework for Long-Term Publishing and Consumption*, o SuDaMa é um framework que busca resolver desafios relacionados a dados abertos como: pontualidade, acessibilidade e usabilidade, viabilizando ações como publicação eficiente de dados dinâmicos em plataformas de dados abertos, bem como melhorando a experiência de consumo de usuários finais (SáNCHEZ-NIELSEN et al., 2021).

O objetivo principal deste framework é funcionar de modo contínuo e automático e pode ser utilizado como uma solução genérica para gerenciar os desafios de publicação e consumo de dados abertos em diferentes domínios. A estratégia é introduzir um ecossistema evolutivo e escalável para garantir o acesso a dados dinâmicos quando são disponibilizados, e abordar a governança usando um agente autônomo para fornecer a capacidade de publicar/despublicar recursos de dados dinamicamente em plataformas de dados abertos.

As principais contribuições do framework são:

- apresentar uma estrutura de gerenciamento OGD como uma solução holística para publicação e consumo de longo prazo. O framework compreende as diferentes fases, desde a configuração até a pós-implantação;
- introduzir um ecossistema de API evolutivo e escalável como componente principal da arquitetura do sistema para garantir acesso em tempo real a dados dinâmicos, bem como gerar dinamicamente novos conjuntos de dados quando estiverem disponíveis;
- oferecer formas avançadas de consumir dados entre usuários finais e editores de dados, introduzindo um bot orientado por API OGD como uma interface de conversação;
- implementação da solução proposta e avaliação de dois componentes principais da arquitetura do sistema (ou seja, o ecossistema API e o bot orientado a API OGD) por um modelo de verificação, validação e teste em um cenário OGD do mundo real.

Na Figura 3.2, adpatada de (SáNCHEZ-NIELSEN et al., 2021), é exibida a arquitetura de alto nível implementada no SuDaMa, composta de quatro componentes: fonte de dados - com nome em ingles *data sources* - uma área intermediária chamada de *stage and storage*, ecossitema da API, *API ecosystem*, e a última etapa, consumo.

Os componentes principais, são: 1) o ecossistema da API, que garantirá a publicação a longo prazo — considerando os requisitos de dados oportunos e acessíveis — e 2) o componente de consumo, que deve garantir o consumo de longo prazo, considerando o requisito de dados utilizáveis.

3.2.3 Ronda

Ronda é uma plataforma de código aberto para coletar, processar e publicar dados abertos em tempo real com base em big data e ferramentas de processamento de dados comprovadas e estabelecidas no setor (KIRSTEIN et al., 2021). Esta plataforma permite que editores de dados abertos forneçam interfaces em tempo real para fontes de dados heterogêneas, promovendo casos de uso de dados abertos sofisticados e avançados.

As principais contribuições desse trabalho são:

- projetar uma arquitetura abrangente para recuperar, processar e fornecer dados abertos em tempo real com base em software de código aberto, padrões de mercado conhecidos e paradigmas de arquitetura de big data estabelecidos;
- desenvolver um protótipo funcional para coletar e disseminar dados em tempo real no contexto de um projeto de produção de cidade inteligente (*smart city*). A solução está disponível como código aberto e pode ser aplicada e estendida a diferentes domínios de dados abertos;
- a arquitetura e protótipo podem servir como um modelo para projetos similares de dados abertos em tempo real, demonstrando um caminho para a próxima geração de portais de dados abertos, que atuam como hubs de dados em tempo real, em vez de simples servidores de arquivos.

A arquitetura de alto nível do Ronda é exibida na Figura 3.3, adaptada de (KIRSTEIN et al., 2021), onde os principais componentes são:

- o chamado de *harvesting module*, módulo de coleta constituindo uma conexão com fontes de dados externas e principal ponto de entrada para o processamento de dados. Ele é responsável por recuperar dados de interfaces externas, transformando-os em uma representação interna e injetando-os no barramento de mensagens Kafka[1];
- o módulo de processamento (*processing module*) oferece recursos arbitrários e extensos para alterar os dados de origem em tempo real. Portanto, é nesse estágio que tarefas de agregação, modificação, normalização e limpeza são executadas;
- o módulo de armazenamento (*storage module*) é responsável por agrupar e armazenar periodicamente os dados em tempo real. Os conjuntos de dados resultantes são organizados cronologicamente e salvos em formatos de dados estruturados, como CSV e JSON;
- o módulo conector (*connector module*) representa a interface pública do sistema, dando acesso aos dados históricos e em tempo real. O acesso em tempo real é fornecido por meio do protocolo WebSocket;

[1] https://kafka.apache.org/

- o módulo agendador (*scheduler module*) funciona como um sistema de agendamento de tarefas recorrentes. É usado principalmente para conectar o sistema a portais de Dados Abertos.

3.2.4 Piveau

Para (KIRSTEIN et al., 2020), uma aplicação mais sofisticada de tecnologias da Web Semântica pode reduzir muitas barreiras na publicação e reutilização de dados abertos. Para tal, os autores propuseram uma plataforma de dados abertos chamada *Piveau*, uma solução de gerenciamento de dados abertos baseada em tecnologias da Web Semântica.

A motivação da criação se deu pela iniciativa de resolver os seguintes problemas encontrados na literatura: limitações nas Interfaces de Programação de Aplicativos (APIs); dificuldades de pesquisa e navegação; falta de informações sobre a qualidade dos dados; baixa capacidade de resposta e mau desempenho.

De acordo com os autores, a solução se diferencia do CKAN (amplamente utilizado) pois este é baseado em um esquema de dados JSON simples, armazenado em um banco de dados PostgreSQL. Segundo eles, isto impede a plena adoção dos princípios da Web Semântica. A expressividade de tal modelo de dados é limitada e não é adequada para uma integração direta de RDF.

A visão de alto nível da plataforma está na Figura 3.4, adpatada de (KIRSTEIN et al., 2020), e a etapa de ingestão pode ser sumarizada da seguinte forma:

- o processo de aquisição de dados dos provedores originais (no desenho identificado como *data providers*). O principal ponto de entrada para qualquer fluxo de trabalho e orquestração de dados é um escalonador. Cada fluxo de trabalho (*workflow*) é atrelado a um acionador, que pode ser executado a cada hora/dia/mês;

- após a execução, o escalonador passa a descrição para o primeiro serviço da fila, normalmente um importador (*importer*). O importador recupera os metadados do(s) portal(is) de origem, por exemplo, CKAN-API, a seguir, extrai os registros de metadados da API ou de um arquivo dump e, após, os envia para a próxima etapa de processamento;

- o transformador (*transformer*) gera o RDF a partir desses dados de origem, aplicando scripts de transformação escritos em JavaScript. A saída final é sempre RDF compatível com DCAT;

- O exportador (*exporter*) envia os dados RDF para o componente Hub.

As demais etapas seguem resumidamente: avaliação da qualidade desses dados, apresentação e gestão dos dados. Os autores ressaltam que o foco estava nas tecnologias e especificações da Web Semântica. Mais detalhes sobre a implementação podem ser consultados diretamente no trabalho do autor, para este trabalho de pesquisa o destaque dessa solução está na abordagem de exposição de dados abertos, buscando superar lacunas em plataforma de dados abertos existentes, como o CKAN.

3.2.5 OWL2MVC

A abordagem proposta por (AYDIN; AYDIN, 2020) é um modelo genérico de aquisição de dados baseado em ontologia para criar formulários no padrão de design model-view-controller (MVC), para publicar e utilizar nas plataformas de dados agrícolas abertos. O modelo proposto consiste em quatro partes diferentes: (1) upload da ontologia

para plataforma aberta de dados; (2) seleção, listagem e configuração de classes de ontologia; (3) criação de model-view-controller; e (4) processamento de formulários criados e manipulação de dados conforme Figura 3.5, adaptada de (AYDIN; AYDIN, 2020).

De modo resumido, as etapas da Figura 3.5 são:

- Na primeira etapa, a *"Carregando Ontologia para Plataforma de Dados Aberta"*, o criador da ontologia carrega a ontologia agrícola relevante na plataforma de dados abertos para extrair os elementos essenciais da ontologia, sendo classes, propriedades de objetos, propriedades de dados e indivíduos;

- na segunda etapa, *"selecionar, listar e definir classes de ontologias"*, quaisquer partes interessadas, como especialistas no domínio, pesquisadores e analistas, podem visualizar a estrutura da árvore da ontologia de forma expandida. A estrutura de visualização em árvore mostra quais classes são adequadas para uso como elemento de formulário e permite selecionar e listar classes dentro de uma tabela de resumo das classes de ontologia selecionadas;

- a terceira etapa *"Criando MVC"*, consiste em duas camadas, que são "Definição de tipos de controle adequados", que permite fazer consultas na ontologia de controles da web, e "Criação de MVC", que define a forma de aquisição de dados, após a seleção das classes e decisões sobre quais tipos de controle serão feitos;

- A última etapa, *"processamento de formulários criados e tratamento de dados"*, de modo resumido, mostra como lidar com formulários e como exportar e armazenar dados coletados por meio de formulários.

3.2.6 Publicação de LDO utilizando padrões de web semântica

Uma abordagem para publicar dados estatísticos de repositórios públicos utilizando padrões de Web Semântica publicados pelo W3C, como RDF e SPARQL, a fim de facilitar a análise de modelos multidimensionais, foi proposta por (ESCOBAR et al., 2020). Embora os dados abertos possam estar disponíveis online, estes dados são geralmente de má qualidade, desencorajando os usuários a fazerem contribuições ou reutilizações.

De acordo com o estudo, foi definida uma estrutura baseada no ciclo de vida da publicação de dados, incluindo etapas de avaliação de LOD e o uso de repositórios externos como base de conhecimento para enriquecimento de dados.

A Figura 3.6, adaptada de (ESCOBAR et al., 2020), detalha o fluxo da abordagem mencionada acima, onde as etapas são:

- especificação da fonte de dados, onde o principal objetivo é limpeza e normalização dos dados para, então, gerar como saída um arquivo único integrado;

- modelagem de dados RDF, que transforma os dados originais na forma de um modelo de dados multidimensional, incluindo componentes como dimensões, medidas e atributos;

- geração do dado, que inclui a transformação dos dados de origem em uma linguagem legível por máquina, ou seja, RDF, proporcionando assim interoperabilidade e links para outros conjuntos de dados;

- publicação do dado, sendo feita de forma direta, a fim de reduzir tarefas complexas de manutenção. A abordagem propõe a publicação do RDF como um arquivo que pode ser acessado por terceiros, incluindo metadados, como informações de licenciamento, e descrito por meio de Vocabulário de conjuntos de dados interligados (na sigla em inglês *VoID*);
- avaliação de LOD, uma lista de critérios de qualidade de dados para avaliar Gráficos de Conhecimento (KGs) no contexto LOD são aplicados;
- exploração de dados, permite a exploração dos dados através de painéis e, além disso, um endpoint público SPARQL poderia ser habilitado para facilitar o acesso e reutilização do conjunto de dados.

3.2.7 IDS como base para ecossistemas de dados abertos

Para (KIRSTEIN; BOHLEN, 2022), os dados abertos ainda enfrentam muitos problemas para desenvolver todo o seu potencial, incluindo barreiras de qualidade, usabilidade, privacidade, jurídicas, estratégicas e técnicas. O setor público continua a ser o seu principal fornecedor, enquanto as partes interessadas da indústria ainda estão relutantes em participar em ecossistemas de dados abertos. Para superar essas desvantagens, os autores propõe uma arquitetura utilizando os conceitos, especificações e tecnologias fornecidas pelos Espaços Internacionais de Dados.

A Figura 3.7, adpatada de (KIRSTEIN; BOHLEN, 2022), mostra a constituição do framework, que contém as seguintes etapas:

- a plataforma possui um conector de dados abertos chamado *conector IDS*, cada entidade de publicação de dados aplica uma instância do conector para anunciar a disponibilidade e conceder acesso aos recursos de dados.
- consumidores de dados solicitam esses dados do conector, que então responde servindo os dados reais dos sistemas internos de gerenciamento de dados;
- o corretor de dados aberto representa uma entidade central, que distribui informações sobre que tipo de dados estão disponíveis de cada participante e quais condições se aplicam à utilização dos dados;
- os consumidores de dados usam os metadados adquiridos do corretor para localizar e selecionar os dados desejados e solicitá-los diretamente da entidade de publicação.

Para os autores, a solução proposta *IDS* difere das demais pois fornece especificações mais rígidas, abrangendo não apenas os metadados, mas todo o processo de comunicação do fluxo do dado, permitindo uma comunicação muito mais harmônica e uma interoperabilidade melhorada.

3.3 Framework para conformidade com leis de proteção de dados

Nesta seção serão expostos os trabalhos que desenvolveram frameworks de software, para verificação de conformidade de dados, abertos ou não, com leis de proteção de dados, como a LGDP no Brasil ou a GDPR na Europa. Foram considerados também, trabalhos que contivessem técnicas que prezassem pela privacidade do indivíduo que interagia com a plataforma de dados, ou o software em questão.

3.3.1 S-GAMER

O *S-GAMER*, do acrônimo em ingles *Security Governance Approach Micro-sERvice*, é uma combinação de microsserviços baseada em um subconjunto de regras do GDPR. Referindo-se a um conjunto de políticas definidas presentes na ontologia de WS-Security (uma extensão do *SOAP* para aplicar segurança aos serviços da Web), o S-GAMER utiliza o microsserviço necessário que detecta vulnerabilidades existentes e verifica a correspondência entre os parâmetros de WS-Security e os requisitos de segurança do usuário (ABIDI et al., 2019).

O contexto desse trabalho é conformidade de regras da GDRP, utilizando parâmetros de *WS-Security* e microsserviços, em ambiente de nuvem. A solução tem o seguinte fluxo, de acordo com a pesquisa, com base em um conjunto de regras do GDPR, "regras de análise e decomposição"que visa decompor cada regra em sentenças, como no exemplo da Figura 3.8, adaptada de (ABIDI et al., 2019), baseado no *Stanford parser*, a saída do parser pode ser vista na Figura 3.9:

O passo (2) é a "extração de termos", que divide cada frase em termos e depois os analisa sintaticamente. Referindo-se a um glossário de *WS-Security*, o próximo passo verifica se os termos extraídos estão relacionados à *WS-Security* ou não. Caso os termos existam no glossário, eles são selecionados e considerados como palavras-chave para, então, ser a base de construção do schema RDF (na sigla em inglês *Resource Description Framework*, que, de acordo com a W3C — principal organização internacional de padrões para a World Wide Web — é uma estrutura para representar informações na Web), que pode ser visto na Figura 3.10.

Os autores complementam que, no que diz respeito às palavras-chave selecionadas, apenas três constituintes (sujeito, verbo e objeto) são considerados na etapa de "construção do esquema RDF"referente ao padrão de segurança apresentado e baseado no esquema RDF construído. Finalmente, a ontologia de microsserviços é criada referindo-se ao esquema construído em RDF. Uma visão geral do framework é apresentada na Figura 3.11, adaptada de (ABIDI et al., 2019).

3.3.2 Sistema de diálogo que preserva a privacidade baseado em argumentação

Os sistemas de diálogo são uma classe de soluções baseadas em inteligência artificial (IA) cada vez mais populares para apoiar a comunicação oportuna e interativa com usuários em muitos domínios. Devido à aparente possibilidade de os utilizadores divulgarem os seus dados sensíveis ao interagirem com tais sistemas, garantir que os sistemas cumpram as leis, regulamentos e princípios éticos relevantes deve ser a principal preocupação (FAZZINGA; GALASSI; TORRONI, 2022).

O trabalho de (FAZZINGA; GALASSI; TORRONI, 2022) propõe uma arquitetura de sistema de diálogo inspirada nos princípios e valores da IA confiável e, segundos os autores, aborda explicitamente os seguintes pontos: (1) a interação do usuário através de linguagem natural, não apenas para fornecer informações ao usuário, mas também para responder às dúvidas do usuário sobre os motivos que levaram à saída do sistema (explicabilidade); (2) o sistema seleciona respostas com base em um módulo de raciocínio transparente, construído sobre uma estrutura de argumentação computacional com uma semântica rigorosa e verificável (transparência, auditabilidade); (3) o tratamento dos dados dos usuários é efetuado de acordo com os princípios de minimização de dados, limitação de finalidade e limitação de armazenamento. Para tal, a interface de linguagem natural e o módulo de raciocínio são dissociados de forma a garantir que nenhum dado

pessoal seja transmitido de um módulo para outro (privacidade e governação de dados).

O modelo de alto nível da arquitetura pode ser visto na Figura 3.12 adaptada de (FAZZINGA; GALASSI; TORRONI, 2022). A arquitetura modular compreende: (1) uma Base de Conhecimento (na sigla em inglês *KB*) feita por especialistas, contendo todos os possíveis casos relevantes, respostas e relações entre eles; (2) um módulo de linguagem que processa a entrada do usuário, incluindo informações confidenciais, e a mapeia para os casos correspondentes da KB; e (3) um módulo de argumentação para raciocinar sobre tais casos de KB e calcular respostas.

Os autores afirmam ainda que é importante destacar que a troca de dados pessoais e sensíveis ocorre apenas entre o usuário e o Módulo de Linguagem, e que, portanto, o módulo Argumentação tem acesso apenas a uma representação geral e ampla que seja estritamente necessária para fornecer a resposta. É destacado, ainda, que qualquer informação considerada irrelevante pelo módulo Linguagem nunca chega ao módulo seguinte. Essas duas propriedades refletem os princípios de minimização de dados e limitação de finalidade, respectivamente, cumprindo, então, a conformidade com regras de proteção de dados.

3.3.3 Framework decidível

O trabalho de (FRANCESCONI; GOVERNATORI, 2023) propõe uma abordagem para verificação de conformidade legal na Web Semântica que pode ser aplicada em LOD. A abordagem baseia-se na modelagem de normas deônticas em termos de classes de ontologias e restrições de propriedades de ontologias.

Segundo os autores, para compartilhar informações na Web de forma compreensível tanto para humanos quanto para máquinas, os princípios da Web Semântica e LOD recomendam a utilização de padrões *Ontology Web Language* OWL/RDF(S), capazes de fornecer uma descrição semântica de um cenário de informação de interesse. RDF é utilizado para descrever instâncias de tal cenário em termos de classes e propriedades, OWL (incluindo RDF Schema) é a linguagem para representar modelos de conhecimento (ontologias) capazes de dar significado a tais classes e propriedades.

O ponto importante do trabalho relacionado é a abordagem para verificação de conformidade legal baseada na distinção entre Disposições e Normas para modelagem de conhecimento e representação de regras, já que a lógica para fazer modelagem de conhecimentos e regras, para então fazer checagem de conformidade, é de interesse da pesquisa.

A representação da modelagem de normas para verificação de conformidade legal pode ser acompanhada no exemplo, com a suposta regra R1, que diz *O fornecedor deverá comunicar ao consumidor todos os termos e condições contratuais* . Os autores exemplificam que, no caso de R1, o cenário ao qual R1 se aplica pode ser modelado em termos de uma ontologia incluindo uma classe *Supplier*, possuindo uma propriedade booleana *hasCommunicatedConditions*. Em termos de OWL, o cenário relativo a R1 pode ser expresso como na Figura 3.13 adaptada de (FRANCESCONI; GOVERNATORI, 2023), onde *myo*: é um namespace fictício para a ontologia *"MyOntology"*. A Figura 3.13 mostra a norma R1 representada como restrição à propriedade do *Supplier hasCommunicatedConditions* e exemplos de indivíduos não conformes (s1) e conformes (s2) (observe que a relação de subclasse entre *FornecedorR1Compliant* e *Supplier* é inferida).

Para os autores, a consulta, SPARQL, na equação 3.1, adaptada de (FRANCESCONI; GOVERNATORI, 2023), é capaz de selecionar os indivíduos que possuem reclamação com R1 (no nosso caso s2). Os autores afirman, aindda, que o raciocínio jurídico em termos de verificação do cumprimento das normas é, portanto, realizado numa estrutura

LOD, utilizando raciocinadores disponíveis num perfil de complexidade computacional decidível.

$$[!ht]SELECT?xWHERE\{?xrdf : typemyo : DriverR2Compliant\} \quad (3.1)$$

3.3.4 Abordagem de privacidade diferencial baseada em Computação Fog para publicação de dados com preservação de privacidade

Endereçando problemas na proteção da privacidade dos cidadãos, após órgaos governamentais publicarem seus dados em ambientes conhecidos como *nuvem*, (PIAO et al., 2019) propõem uma estrutura de privacidade diferencial para a publicação de dados estatísticos governamentais baseada na computação fog. Para o autor, há um grande número de riscos operacionais nas atuais plataformas governamentais em nuvem, pois, quando uma plataforma em nuvem é atacada, a maioria dos modelos existentes de proteção de privacidade para publicação de dados não consegue resistir aos ataques se o invasor tiver conhecimento prévio.

Os autores explicam a estrutura, presente na Figura 3.14 adaptada de (PIAO et al., 2019), que, em suas palavras, consiste em uma arquitetura de publicação de dados e um processo de preservação de privacidade. A arquitetura de publicação de dados serve como base para apoiar o processo de preservação da privacidade e, portanto, o algoritmo de preservação da privacidade em execução no processo. Para levar em conta os riscos de privacidade inerentes à nuvem do *governo eletrônico*, os autores utilizam um modelo híbrido cloud-fog para apoiar a arquitetura de publicação de dados.

Para os autores, as principais vantagens em usar o modelo híbrido de computação cloud-fog para agências governamentais realizarem publicação de dados públicos são: (1) o algoritmo de preservação da privacidade pode ser executado em equipamentos e recursos de computação existentes no governo, reduzindo os gastos com grandes quantidades de equipamentos de alto desempenho; (2) terminais multipartidários ou dispositivos de borda de rede podem servir como armazenamento de dados, reduzindo a necessidade de alta largura de banda e alto desempenho de nuvens governamentais, melhorando a eficiência do processamento de dados e reduzindo o custo de construção de nuvens governamentais; (3) As funções de proteção e de privacidade da computação fog quem melhoram a segurança da divulgação de dados governamentais e aumentam a credibilidade.

3.3.5 Abordagem baseada em blockchain para verificar conformidade com o GDPR em ambientes multinuvem

Para (AHMAD; AUJLA, 2023), as abordagens baseadas em blockchain ganharam popularidade nos últimos anos para enfrentar o desafio de verificar a conformidade com o GDPR em ambientes multinuvem. Isso porque, ao implantar contratos inteligentes no blockchain, os autores afirmam que é possível criar registros transparentes e imutáveis de processos de dados na esperança de automatizar a verificação de conformidade com o GDPR. No entanto, os autores finalizam dizendo que os trabalhos existentes estão limitados a fornecer uma verificação de conformidade centrada no usuário, logo, sua solução é uma verificação de conformidade com a lei de proteção de dados europeia, GDPR, utilizando blockchains, porém centrada no usuário.

Na explicação dos autores, na estrutura centrada no usuário e baseada em blockchain, todas as operações de dados relevantes ao GDPR ocorrem no blockchain por meio de contratos inteligentes bem definidos.

No trabalho dos autores todos os contratos são explicados, mas, para o foco da pesquisa, apenas o contrato de GDPR é escopo, e então será brevemente explanado. O contrato chamado de *contrato inteligente de verificação GDPR* é responsável pelo registro das operações de dados que ocorrem. Os autores explicam que, ao registrar essas operações de dados, elas passam por um filtro de perguntas relacionadas ao GDPR para detectar violações, e seguem então o fluxo: (1) as violações de dados são então registadas numa "tabela de violações"; (2) o contrato de verificação também é usado para detectar dados incorretos mantidos por outras entidades; (3) métodos nos contratos do cliente, e entidades são chamados para calcular valores de *hash* para os dados armazenados; (4) esses valores de hash são recuperados pelo contrato de verificação e comparados para verificar a exatidão dos dados, com valores de hash incompatíveis resultando em uma violação do GDPR que é registrada na tabela de violações; (5) a tabela de operações de dados é exibida na página de Verificação e a tabela de violações do GDPR é enviada à página do Cliente para informar os clientes.

Um exemplo prático da operação pode ser visto na Figura 3.15 adaptada de (AHMAD; AUJLA, 2023) , onde, na sequência da esquerda para direita, estão os dados sendo gerados na página da web *PrettyBigThing*; e então sendo transferidos para o contrato de mesmo nome, que na sua saída gera um evento com detalhes da operação; sendo então recebido pela aplicação web, através da ferramenta de captura de eventos; enviando, por fim, as informações para o contrato de verificação que fará a checagem de conformidade GDPR.

3.4 Considerações finais

Este capítulo expôs uma ampla gama de tópicos nos quais plataforma e frameworks de dados abertos vem sendo utilizados. Incluindo definições, benefícios, desafios, padrões de interoperabilidade, impactos sociais e econômicos, entre outros aspectos. A literatura estudada mostrou como, atualmente, são utilizadas ferramentas de exposição de dados e frameworks de verificação de conformidade com leis de proteção de dados. A variedade de tópicos, de certo modo, corrobora com o tamanho dos desafios encontrados, pois mostra as diferentes iniciativas necessárias para que os dados sejam expostos com qualidade e verificando a proteção a privacidade de usuários. Quanto as subquestões de pesquisa que guiaram a busca de trabalhos relacionados, foi verificado que sistemas como SODAS e SuDaMa, por exemplo, são plataformas de exposição de dados abertos, que cobrem lacunas de plataformas amplamente utilizadas — como CKAN —, no entanto não possuem meio de verificação de conformidade com a LGPD. Outras soluções encontradas, fazem a verificação de conformidade com proteção de dados pessoais (como S-GAMER), porém não são para dados abertos. Por fim, frameworks de exposiçao de dados abertos e verificação de conformidade com proteção de dados aberto no Brasil, não foram encontrados durante a pequisa, utilizando a metodologia de revisão da literatura exposta nesse capítulo. Sendo assim a a tabela 3.6 demonstra o diferencial do framework proposto neste trabalho em comparação com as soluções existentes.

Tabela 3.6: comparativo entre soluções vs questões de pesquisa

No.	Autor e ano	RQ1	RQ2	RQ3	RQ4
1	(FAZZINGA; GALASSI; TORRONI, 2022)		X		
2	(AHMAD; AUJLA, 2023)		X		X
3	(FRANCESCONI; GOVERNATORI, 2023)		X		
4	(PIAO et al., 2019)		X		
5	(KIRSTEIN; BOHLEN, 2022)	X			
6	(ESCOBAR et al., 2020)	X			
7	(AYDIN; AYDIN, 2020)	X			
8	(KIRSTEIN et al., 2020)	X			
9	(ABIDI et al., 2019)		X		X
10	(SáNCHEZ-NIELSEN et al., 2021)	X			
11	(KIRSTEIN et al., 2021)	X			
12	(WON et al., 2021)	X			
13	CompOD (CARMO et al., 2024)	X	X	X	

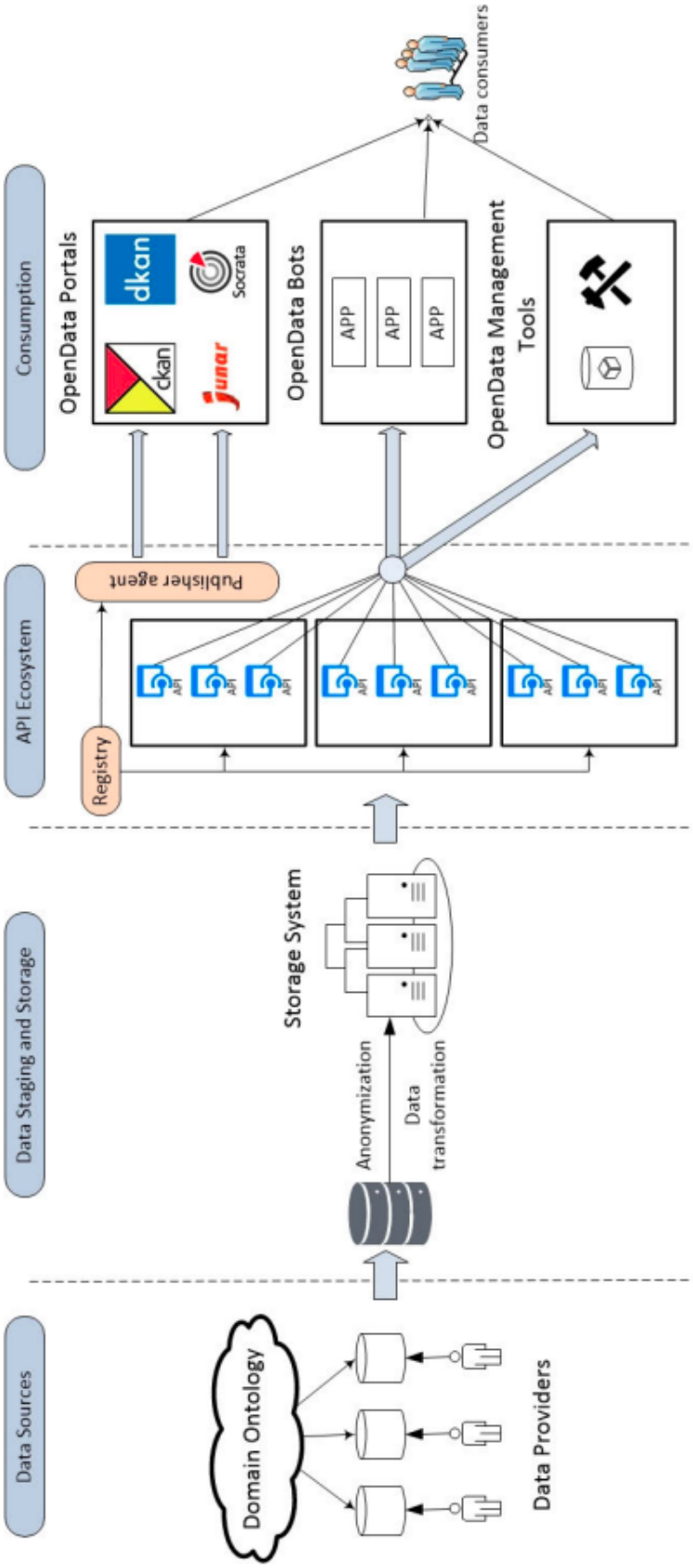


Figura 3.2: arquitetura geral do sistema SuDaMa. (Adaptado de (SáNCHEZ-NIELSEN et al., 2021))

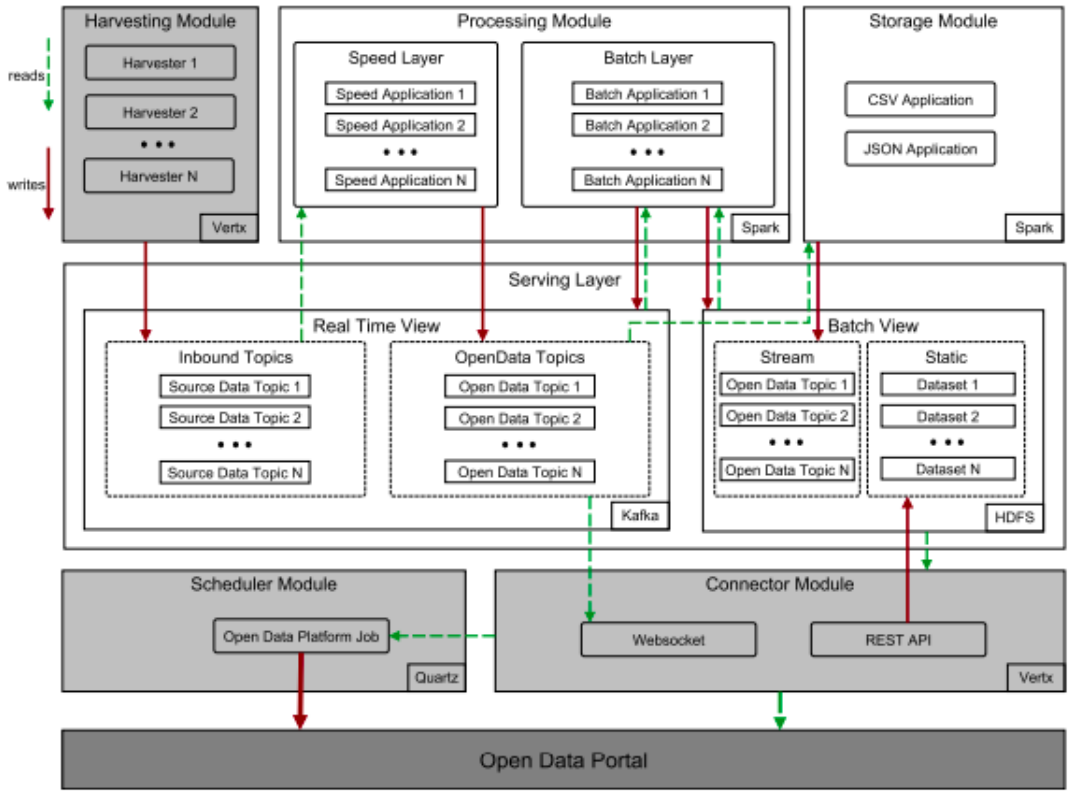


Figura 3.3: estrutura geral da arquitetura da plataforma de dados Ronda. (Adaptado de (KIRSTEIN et al., 2021))

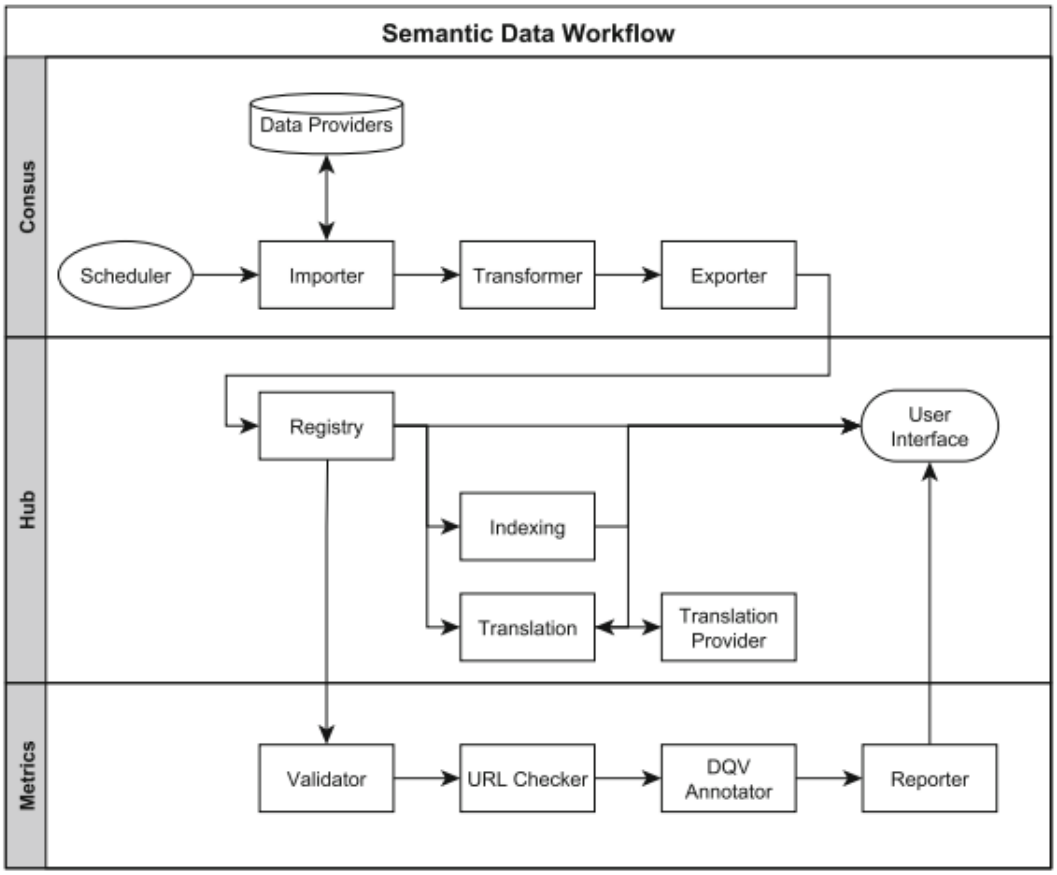


Figura 3.4: visão geral Piveau. (Adaptado de (KIRSTEIN et al., 2020))

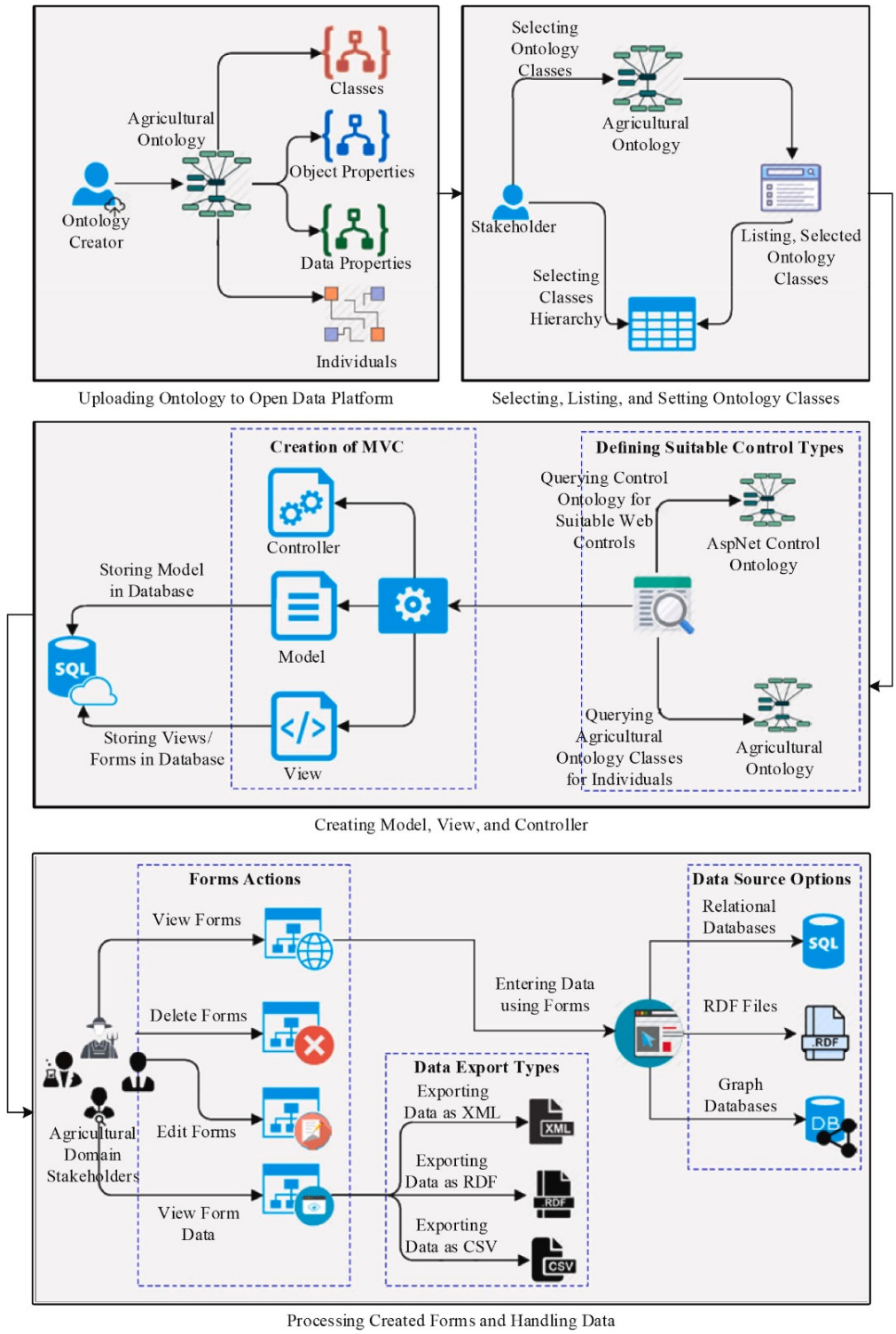


Figura 3.5: modelo de aquisição de dados baseado em ontologia. (Adaptado de (AYDIN; AYDIN, 2020))

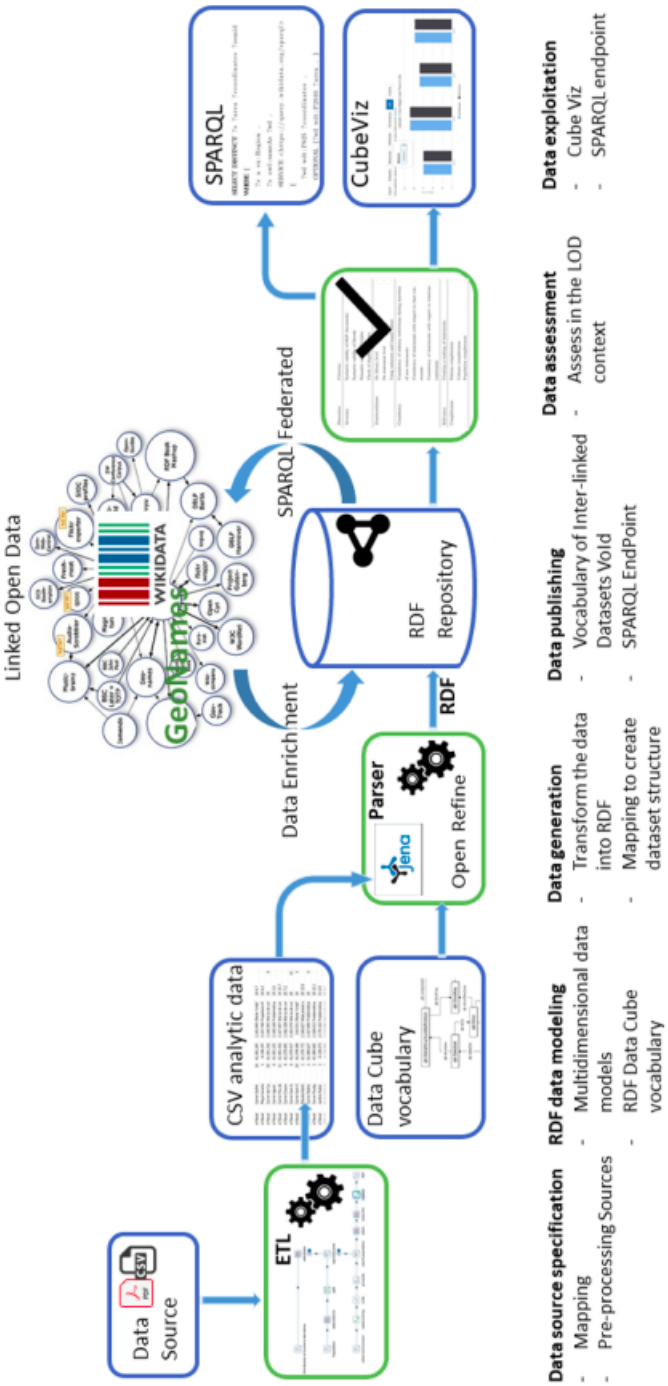


Figura 3.6: A estrutura para publicação de Linked Open Data. (Adaptado de (ESCOBAR et al., 2020))

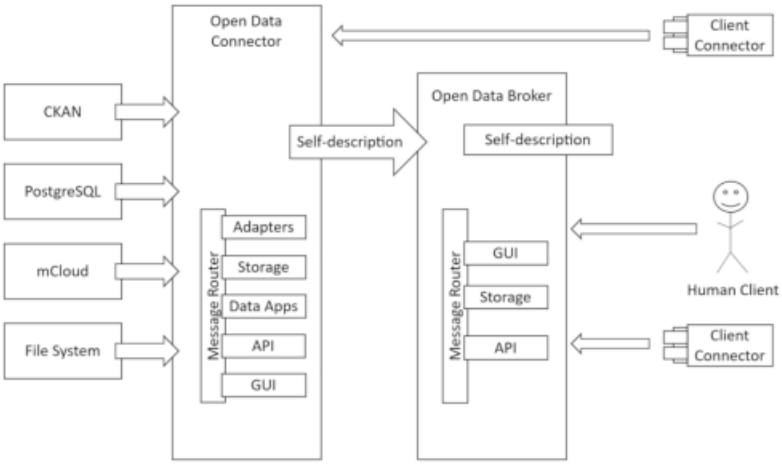


Figura 3.7: visão geral dos componentes e fluxo de (meta)dados no ecossistema de dados abertos do IDS. (Adaptado de (KIRSTEIN; BOHLEN, 2022))

(7)Those developments require a strong and more coherent data protection framework in the Union, backed by strong enforcement, given the importance of creating the trust that will allow the digital economy to develop across the internal market. Natural persons should have control of their own personal data. Legal and practical certainty for natural persons, economic operators and public authorities should be enhanced.

Figura 3.8: exemplo de Regra GDPR. (Adaptado de (ABIDI et al., 2019))

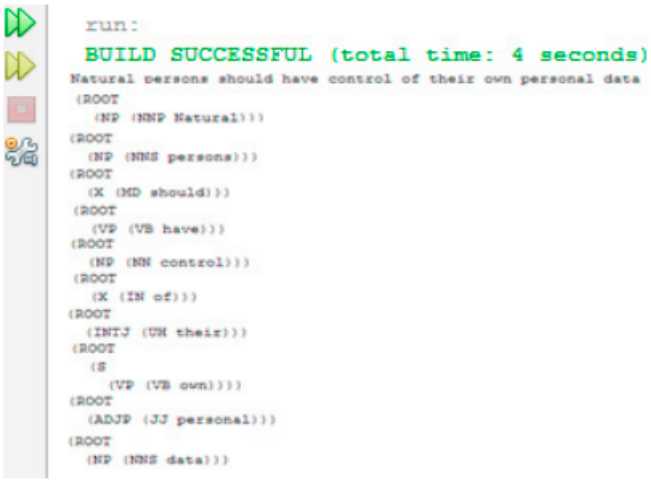


Figura 3.9: saída do Stanford Parser. (Adaptado de (ABIDI et al., 2019))

```
run:
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
  <rdf:Description rdf:nodeID="A0">
    <rdfs:label rdf:parseType="Literal">Data_protection</rdfs:label>
    <rdfs:label xml:lang="fr">enforcement</rdfs:label>
    <rdfs:label xml:lang="en">Control</rdfs:label>
  </rdf:Description>
</rdf:RDF>
```

Figura 3.10: amostra schema RDF. (Adaptado de (ABIDI et al., 2019))

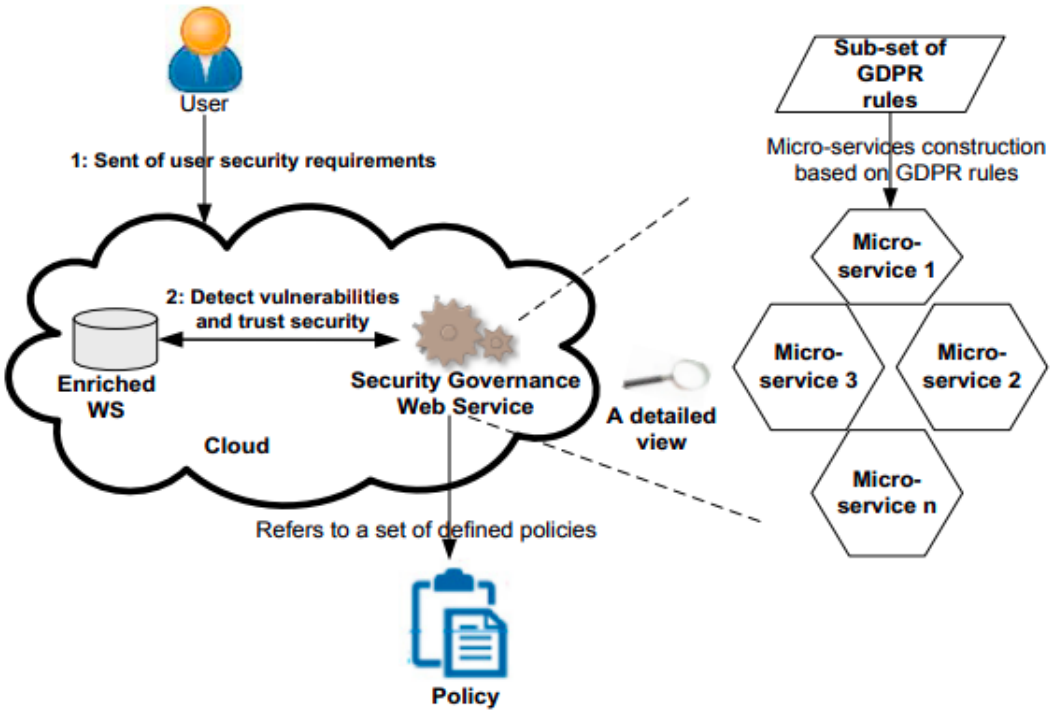


Figura 3.11: visão geral S-Gamer. (Adaptado de (ABIDI et al., 2019))

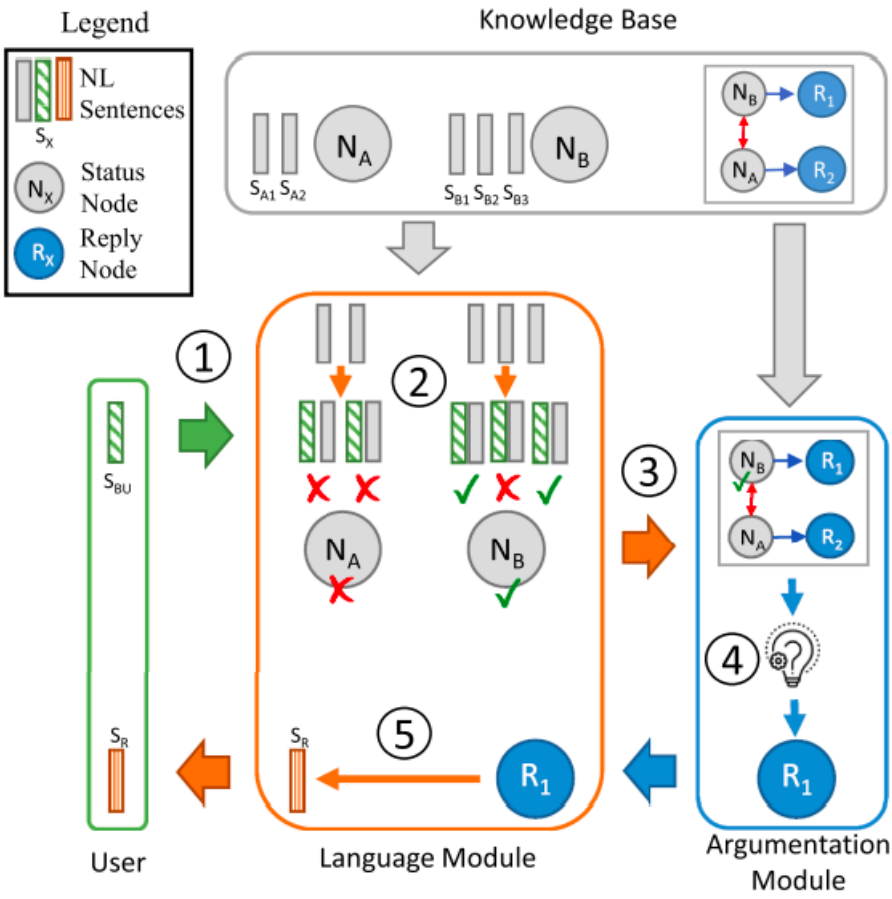


Figura 3.12: visão geral sistema de argumentação. (Adaptado de (FAZZINGA; GALASSI; TORRONI, 2022))

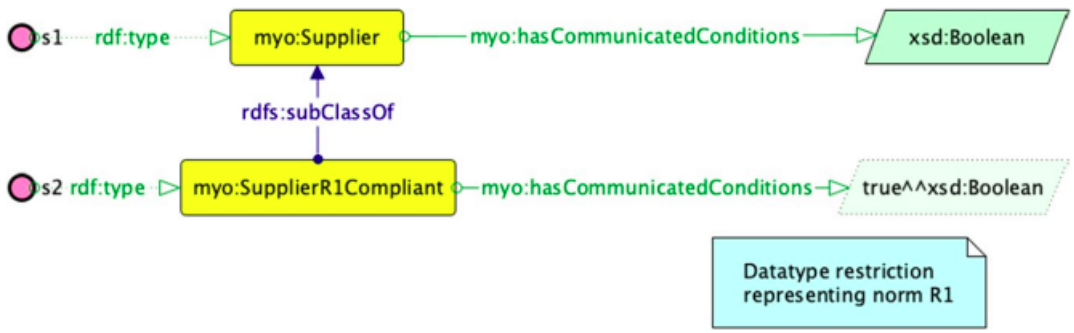


Figura 3.13: representação de modelagem de normas. (Adaptado de (FRANCESCONI; GOVERNATORI, 2023))

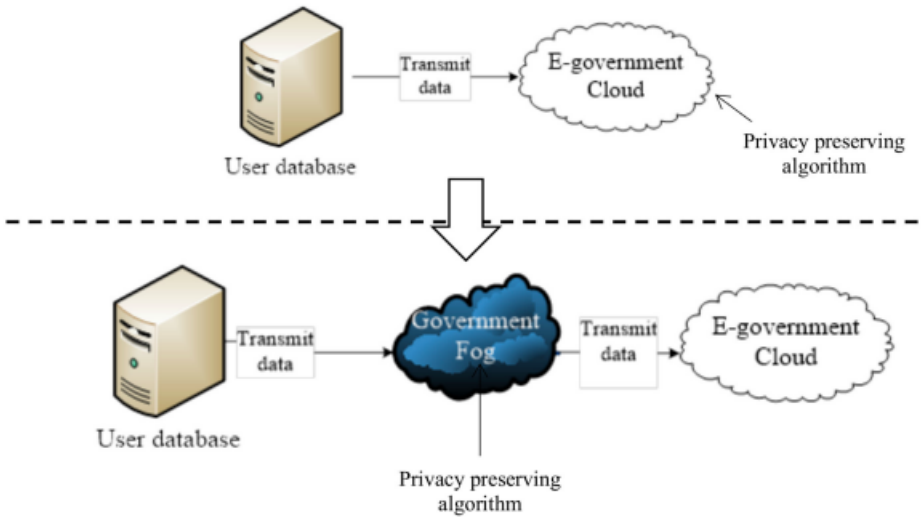


Figura 3.14: modelo de nuvem tradicional vs. modelo híbrido cloud-fog. (Adaptado de (PIAO et al., 2019))

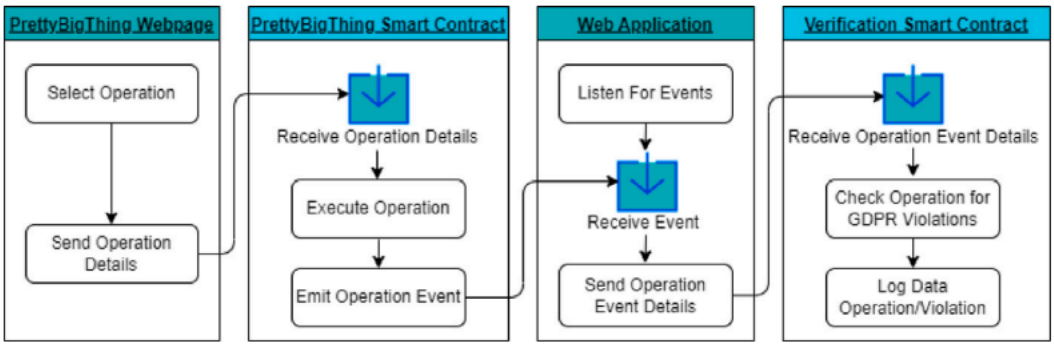


Figura 3.15: visão geral dos componentes e fluxo de (meta)dados no ecossistema de dados abertos do IDS. (Adaptado de (AHMAD; AUJLA, 2023))

4 MODELO PARA DESENVOLVIMENTO DO FRAMEWORK

Para atender aos objetivos deste trabalho, o modelo foi divididos em dois módulos:

- Um módulo para análise quantitativa e qualitativa dos dados abertos no Brasil, que tem como objetivo demonstrar o panorama atual destes;
- um módulo de conformidade com a LGPD, que, como o nome sugere, busca demonstrar se os dados abertos disponíveis estão em conformidade com a lei brasileira de proteção de dados.

Para facilitar a fluidez da leitura nas próximas seções e capítulos, os módulos do framework serão mencionados por suas siglas conforme a Tabela 4.1

Tabela 4.1: Módulos desenvolvidos na pesquisa

Sigla	Descrição
M1	Protótipo de estudo sistemático da quantidade e qualidade da exposição dos dados abertos.
M2	Protótipo de conformidade com a LGPD.

É importante ressaltar que os trabalhos relacionados ao Módulo M1, poderiam ser utilizados por terem a mesma finalidade deste trabalho (inclusive alguns com melhorias em lacunas do CKAN), no entanto, não foi possível reproduzir os estudos feitos e reaproveitá-los, pois o código-fonte não foi encontrado. Logo, apesar de entender que o CKAN possui suas limitações — como exposto nos trabalhos relacionados — para a metodologia GODI [1], na qual a análise de quantidade e qualidade foi baseada, a utilização da *Action API* do CKAN atendeu aos requisitos necessários.

Quanto ao Módulo M2, de conformidade com a LGDP, não foi encontrado na literatura um trabalho relacionado que atendesse ao escopo desta pesquisa, *conformidade LGPD em portais de dados abertos*.

4.1 Módulo M1

O projeto *Open Definition* afirma que o termo "aberto"no contexto de "dados abertos"e "conteúdo aberto"significa que os dados podem ser livremente acessados, usados, modificados e compartilhados por qualquer pessoa, com qualquer pessoa, estando sujeitos, no máximo, a exigências que preservem a procedência e a abertura (FOUNDATION, 2023).

[1] http://index.okfn.org/methodology/index.html

A *Open Knowledge Foundation* (OKFN) criou o *Open Data Index* (ODI), uma iniciativa pioneira na promoção da transparência, ajudando a avaliar políticas, identificar gargalos e orientar os municípios a melhorarem as suas políticas de dados abertos (INDEX, 2018). O Índice avalia não apenas os governos federais, mas também os municipais, atuando para garantir a escalabilidade necessária. De acordo com a Fundação Getúlio Vargas (FVG, 2017)[2], o ODI foi trazido ao Brasil por meio de uma parceria entre a Diretoria de Análise de Políticas Públicas (DAPP), da FGV, e a *Open Knowledge Brasil* (OKBR)[3].

O OKBR também define as principais condições para que os dados sejam considerados abertos. Em síntese: (1) disponibilidade e acesso, ou seja, devem estar totalmente disponíveis e apenas ao custo de cópia; (2) também num formato conveniente e mutável; (3) deve ser possível sua reutilização e redistribuição, ou seja, além de ser possível reutilizar, deve ser possível combiná-los com outro conjunto de dados; (4) e participação universal, ou seja, todos devem poder utilizá-los, sem qualquer discriminação contra pessoas, grupos ou campos de ação relativos (como apenas para fins sem fins lucrativos ou educacionais).

De modo a validar os itens mencionados no parágrafo anterior, o modelo de análise de dados desta pesquisa possui as seguintes etapas, conforme a Figura 4.1: (1) os portais de dados abertos são verificados manualmente (acessando browser e fazendo a pesquisa) pelo usuário para validação do modo de exposição dos dados. O uso de frameworks de exposição padronizados e amplamente utilizados pela comunidade é um requisito para análise (mesmo que o portal de dados possua um padrão de exposição de dados, se não atender a um padrão de exposição superior e amplamente utilizado — como o CKAN/DKAN — não será possível aplicar estudos automatizados, juntamente a outros portais); (2) o mapeamento de portais de dados abertos deve ser feito e o material necessário preparado para inicialização do estudo sistemático; (3) em seguida, deve-se executar a validação desejada, que terá sua lógica contida em um ambiente de desenvolvimento interativo reproduzível, inicializando uma função para buscar recursos nos portais; (4) isso deverá acontecer através de um framework de exposição de dados abertos padronizado, como CKAN/DKAN, por exemplo. Após a inicialização da validação, na próxima etapa; (5) o framework de validação executará as métricas a serem analisadas, e, por fim retornará o resultado ao usuário.

Uma abstração da validação de métrica desenvolvido está na Figura 4.2: Inciando pela verificação manual do requisito de uso do CKAN no portal; seguido pelo módulo M1, executando a chamada da API CKAN desejada (detalhada na Seção 5.1 de *Metodologia*); caso a chamada inicial retorne sucesso, a validação de *bulks* (massa de dados) do recurso é também validada, o retorno da chamada é processado, e, por fim, seu status final disponibilizado ao usuário. Este fluxo validaria, por exemplo, se os recursos disponibilizados num conjunto de dados têm suas rotas retornando status de sucesso ou falha. Ou seja, se os recursos estão acessíveis.

4.2 Módulo M2

De acordo com (LOGAREZZI, 2016), o livre acesso à informação pública é fundamental para o funcionamento das democracias. A falta de informação dificulta a avaliação das políticas públicas, o controle social e a participação qualificada da população. Portanto, os autores afirmam, ser indispensável que os cidadãos conheçam o modo de

[2]FGV é uma instituição brasileira conhecida mundialmente como referência em ensino e pesquisa
[3]https://ok.org.br/

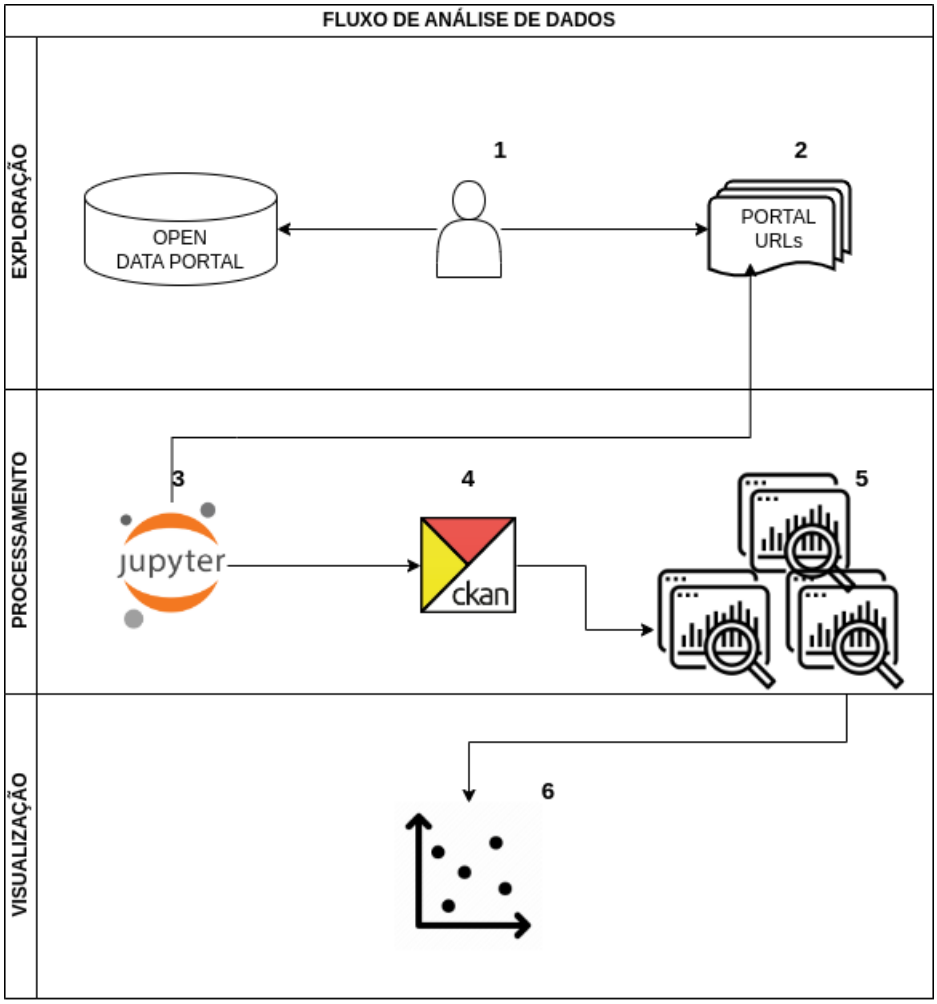


Figura 4.1: modelo de desenvolvimento M1. Fonte: autor.

funcionamento dos órgãos estatais e as ações dos governos para poderem exigir que seus direitos sejam cumpridos.

No entanto, assim como o direito à informação, a proteção de dados pessoais também é um fator fundamental para a garantia da democracia, especialmente em uma sociedade cada vez mais orientada a dados e progressivamente tecnológica (BIONI; SILVA; MARTINS, 2022).

Com a finalidade de fundamentar o exposto no parágrafo anterior, o objetivo principal deste módulo é verificar, em portais de dados abertos, se existem dados pessoais expostos, ou se, conforme a LAI, o dado aberto disponibilizado sempre está respeitando os limites de exposição, conforme a LGPD, com relação a informações pessoais identificáveis (na silga em inglês PII) ,como CPF, nome completo e endereço, por exemplo.

O fluxo semântico de verificação de conformidade com a LGPD, está na Figura 4.3, onde será iniciado por:

- (1) parâmetros disponibilizados pelo usuário, como URL do portal e *termo de busca* (por exemplo, CPF);

- seguido pelo módulo (2) de extração, com base nos parâmetros informados pelo usuário, que utilizará uma API padronizada e amplamente utilizada na comunidade

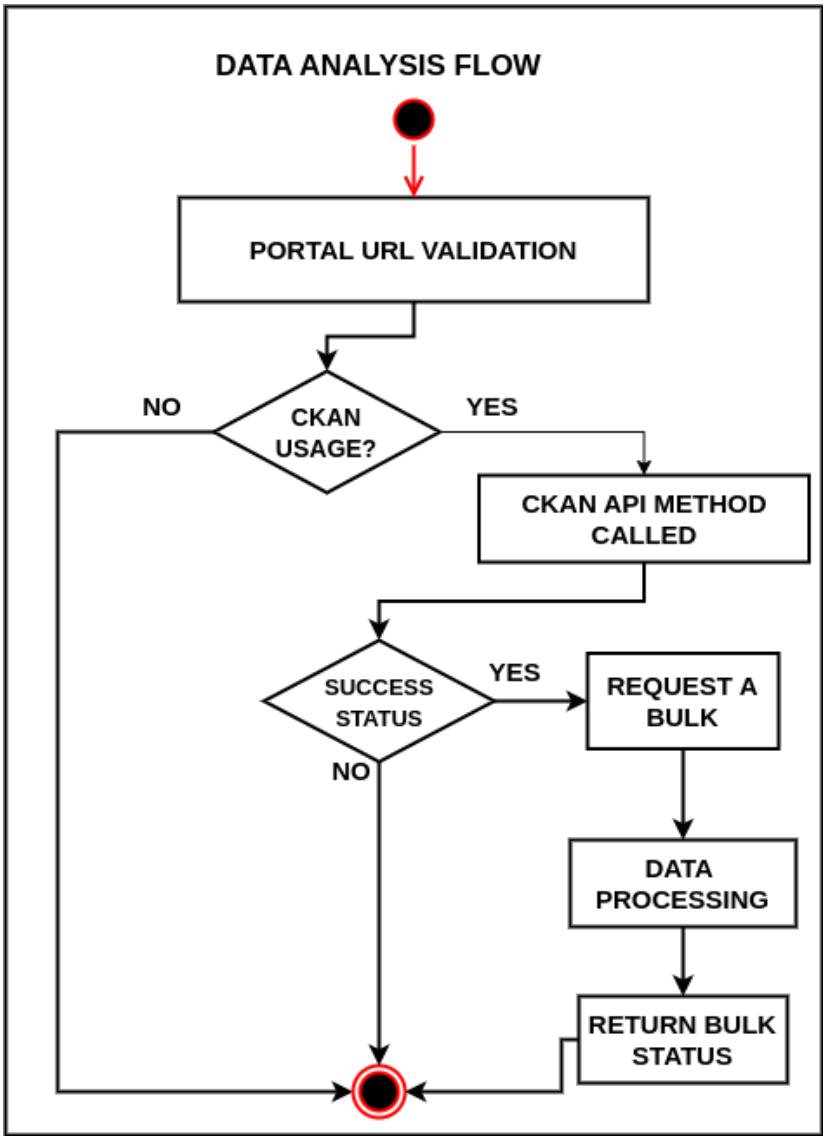


Figura 4.2: Fluxo de Validação Análise Qualitativa. Fonte: Autor

para extração dos dados e, então, fará a busca correspondente nas bases de dados dos portais de dados abertos;

- Ao receber o recurso, o (3) o modulo de extração salva os dados temporariamente numa tabela bidimensional de linhas e colunas, ajustando o retorno conforme os parâmetros passados pelo usuário;
- por fim, (4) salva os dados numa stage efêmera;
- a seguir, na etapa (5), o módulo de processamento acessará o recurso encontrado na stage, e fará uma requisição de busca na rota;
- o conteúdo da rota recebido (6) passará por uma etapa de limpeza, sendo removidos os dados mal formatados, nulos e outras anomalias nos dados;
- na etapa (7), o dado será processado e sua saída (conhecida tecnicamente como output) será verificada pelo *LGPD conformidade checker*;

- na etapa (8) o dado recebido será analisado pelo algoritmo, e o *termo de busca* informado pelo usuário será buscado pelo algoritmo, de modo que, caso seja encontrado dado sensível compatível com o *termo de busca*, será retornado para o usuário na etapa (9);
- se o conjunto de dados, conhecido como *dataset*, estava ou não em conformidade, essa informação estará disponível em um frontend, na etapa (10), para o usuário final.

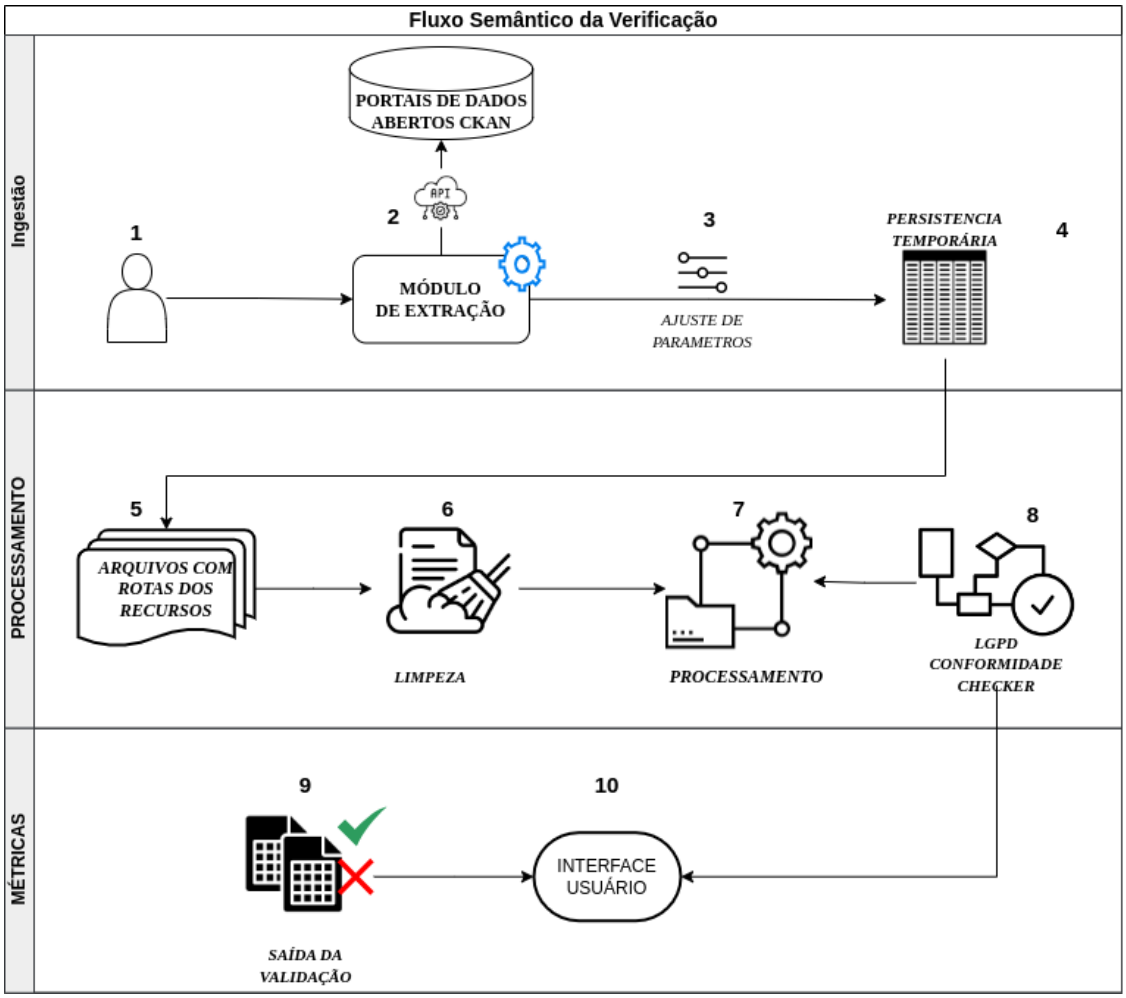


Figura 4.3: modelo de desenvolvimento, M2, de conformidade LGDP - CompOD. Fonte: autor.

No modelo de desenvolvimento, o fluxo de decisão necessário até a aplicação do validador de conformidade foi desenhado conforme a Figura 4.4. Após o recebimento dos parâmetros, deve ocorrer a verificação se estes atendem os requisitos previamente definidos, de modo a manter a qualidade e evitar falhas no framework, por exemplo, se o portal de dados abertos está exposto com API amplamente acessíveis, e se o termo de busca é válido; depois, a consulta do recurso é iniciada, e logo em seguida deve ser verificado se está acessível; em caso de sucesso, o recurso deve conter um formato previamente definido, e ser comumente utilizado em exposições de dados abertos (isto deve ocorrer para que a possibilidade de sucesso na validação final seja maior); estando no formato correto, os dados devem ser baixados para o ambiente de validação, deve ser aplicada limpeza

nos dados e, então, os dados serão processados; se o parâmetro de verificação desejado existir no conteúdo do dado baixado, o algoritmo que verifica conformidade será aplicado. Nesta etapa, a seguinte verificação será feita: supostamente o usuário solicitou verificação de conformidade para CPF, então o algoritmo previamente testado buscará pelo parametro CPF como atributo do dado importado e buscará pelo valor retornado. O algoritmo deve ser capaz de reconhecer se a informação é um dado sensível ou não, e também se está exposto ou protegido. Por fim, este resultado será retornado ao usuário que requereu a verificação inicialmente.

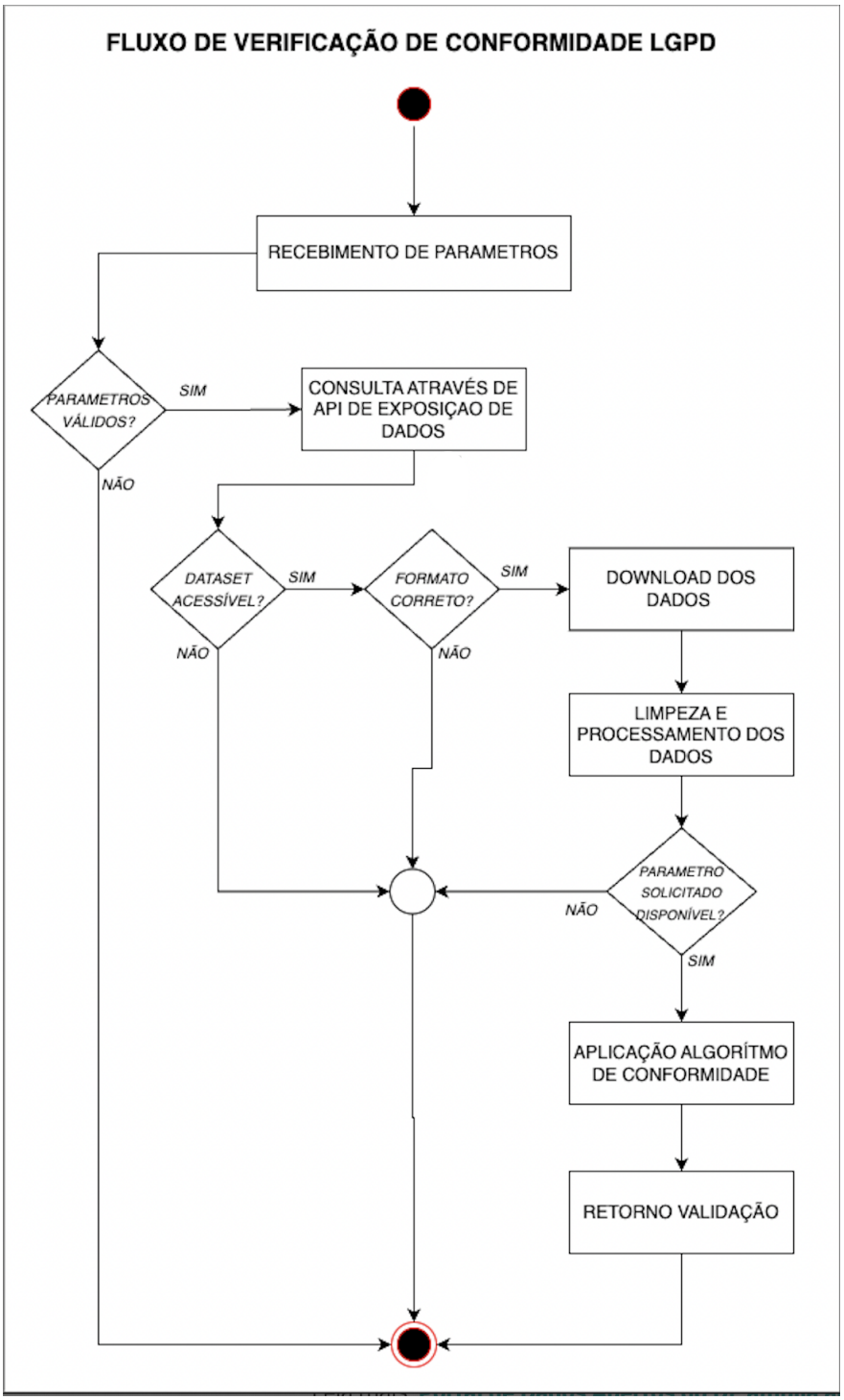


Figura 4.4: fluxo de decisão, M2 - CompOD. Fonte: autor.

5 AVALIAÇÃO

Neste capítulo será apresentada a metodologia de desenvolvimento com base no modelo exposto previamente. Será detalhado o escopo dos experimentos aplicados e seus resultados. O objetivo é fornecer o fluxo através do qual os resultados foram alcançados, utilizando dados empíricos e tarefas bem definidas.

5.1 Metodologia

A metodologia é o método de desenvolvimento aplicado aos módulos M1 e M2, de maneira a construir o framework proposto por este trabalho, ou seja, o *CompOD*. Sendo assim, a metodologia buscou organizar os procedimentos feitos de forma sistemática para desenvolver o framework. Assim, serão apresentadas as tecnologias utilizadas no desenvolvimento do código, como ambiente de desevolvimento e bibliotecas; será destacado o processo para aquisição dos dados abertos, explicando a tecnologia escolhida e seus métodos; por fim, será esclarecido o escopo dos dados utilizados, e suas terminologias.

5.1.1 Ambiente de desenvolvimento e reprodutibilidade

As tecnologias utilizadas foram: linguagem de programação *Python* para desenvolvimento dos módulos M1 e M2; para o ambiente de desenvolvimento do módulo M1, a distribuição *Anaconda* em *Notebooks Jupyter* (os *notebooks* são espaços de desenvolvimento interativo, que facilitam a análise de conjuntos de dados). Através da biblioteca *Matplotlib* a plotagem de dados feita; para o módulo M2 o ambiente de desenvolvimento utilizado foi o *Visual Studio Code*, com o treinamento do modelo de aprendizado máquina sendo viabilizado pela biblioteca *Scikit-Learn*; a etapa de disponibilização dos dados e métricas foram desenvolvidos com *Flask* para hospedagem da aplicação; em ambos os módulos, o processamento dos dados foi executado utilizando *Pandas*, que é uma biblioteca para análise e processamenro de dados; e o versionamento de código foi feito utilizando o sistema de controle de versão *Github*.

Os códigos desenvolvidos nos módulos M1 e M2 estão disponíveis no Github, conforme a Tabela 5.1. Desenvolvidos de modo complementar, porém com finalidades/tecnologias diferentes, cada módulo está em um repositório próprio.

Tabela 5.1: repositório Git dos módulos M1 e M2

Módulo	Repositório Git
M1	https://github.com/shluh/ufrgs-open-data-analysis
M2	https://github.com/shluh/ufrgs-open-data-lgpd-compliance-probe

5.1.1.1 Tecnologias

- Anaconda é uma distribuição das linguagens de programação Python e R para computação científica, que visa simplificar o gerenciamento e implantação de pacotes. A distribuição inclui pacotes de ciência de dados adequados para Windows, Linux e macOS (ANACONDA , PYTHON DISTRIBUTION).

- O Projeto Jupyter é um projeto de código aberto sem fins lucrativos, que evoluiu para apoiar a ciência de dados interativa e a computação científica em todas as linguagens de programação. O Jupyter é software 100% de código aberto, gratuito para uso de todos e lançado sob os termos liberais da licença BSD modificada (PROJECT JUPYTER, 2023).

- Scikit-Learn (anteriormente scikits.learn e também conhecido como sklearn) é uma biblioteca de software livre de aprendizado de máquina para a linguagem de programação Python. Ele apresenta vários algoritmos de classificação, regressão e clustering, incluindo máquinas de vetores de suporte, florestas aleatórias, aumento de gradiente, k-means e DBSCAN, e é projetado para interoperar com as bibliotecas numéricas e científicas Python NumPy e SciPy (SCIKIT-LEARN, 2023).

- Visual Studio Code, também conhecido como VS Code, é um editor de código-fonte desenvolvido pela Microsoft para Windows, Linux e macOS. Os recursos incluem suporte para depuração, destaque de sintaxe, preenchimento inteligente de código, snippets, refatoração de código e Git incorporado (VISUAL STUDIO CODE, 2023).

- Flask é um micro framework web escrito em Python. É classificado como micro framework porque não requer ferramentas ou bibliotecas específicas. Não possui camada de abstração de banco de dados, validação de formulário ou qualquer outro componente onde bibliotecas pré-existentes de terceiros forneçam funções comuns(FLASK , WEB FRAMEWORK).

- Matplotlib é uma biblioteca de plotagem para a linguagem de programação Python e sua extensão matemática numérica NumPy. Ele fornece uma API orientada a objetos para incorporar gráficos em aplicativos usando kits de ferramentas GUI de uso geral, como Tkinter, wxPython, Qt ou GTK (MATPLOTLIB, 2023).

- Pandas é uma biblioteca de software escrita para a linguagem de programação Python para manipulação e análise de dados. Em particular, oferece estruturas de dados e operações para manipulação de tabelas numéricas e séries temporais. É um software livre lançado sob a licença BSD de três cláusulas (PANDAS , SOFTWARE).

- Python é uma linguagem de programação de alto nível e de uso geral. Sua filosofia de design enfatiza a legibilidade do código com o uso de recuo significativo. Ele suporta vários paradigmas de programação, incluindo programação estruturada, orientada a objetos e funcional (PYTHON , PROGRAMMING LANGUAGE).

- Git é um sistema de controle de versão distribuído e gratuito, de código aberto projetado para lidar desde projetos pequenos a muito grandes, com velocidade e eficiência (GIT, 2023).

5.1.2 Aquisição de Dados Abertos

A ferramenta de verificação e consulta de dados abertos utilizada foi o CKAN[1][2], pois países como Estados Unidos, Canadá, Austrália, e muitos órgãos governamentais brasileiros, o usam, facilitando um estudo sistemático em diferentes continentes, caso seja necessário, através dos módulos M1 e M2.

CKAN é uma ferramenta para criar sites de dados abertos (pense em um sistema de gerenciamento de conteúdo como o WordPress[3] – mas para dados, em vez de páginas e postagens de blog). Mais do que isso, ele ajuda a gerenciar e publicar coleções de dados, sendo utilizado por governos nacionais e locais, instituições de investigação e outras organizações que recolhem muitos dados (CKANGUIDE, 2023).

A extração dos dados para os módulos M1 e M2 foi feita através da *Action API* do CKAN. Portanto, foram analisados apenas conjuntos de dados que expõem dados por meio desta API – ou derivações como DKAN[4]. Para a estrutura técnica foram utilizadas as *funções de API compatíveis com GET* (*GET-able API functions*), com recursos por conjunto de dados, grupo e recursos. Para consultar todas as chamadas da API utilizadas no processamento de dados, consulte a Tabela 5.2.

Tabela 5.2: lista de datasets JSON-formatted

Função API	Descrição
/action/package_list	retorna a lista de datasets disponíveis por plataforma
/action/group_list	retorna os grupos/temas que estão contidos nos datasets
/action/tag_list	retorna o detalhamento das tags por conjunto de dados
/action/organization_list	retorna as organizações que gerenciam os datasets contidos no portal de dados
/action/package_show?	retorna a representação do conjunto de dados especificado no parâmetro 'id'
/action/package_search?	para procurar conjuntos de dados (pacotes) que correspondam à consulta de pesquisa

A Figura 5.1 mostra o exemplo de construção de uma rota que funcionará como alvo de uma análise. Três variáveis devem ser definidas inicialmente: a URL do portal de validação; a ação da API desejada, conforme Tabela 5.2; o ID da organização, que contém os dados a serem analisados. A saída ao final, será conforme a saída destacada em azul.

5.1.3 Conjunto de dados e recursos

No guia do usuário[5] do CKAN, existe uma distinção conceitual entre conjunto de dados (em inglês *datasets*) e recursos. Entendê-la é importante, pois, cada módulo M1 e M2 tem o foco num objeto específico. O módulo M1 analisa os datasets, seu gerenciadores (organizações), grupos/temas, tags e metadados (como por exemplo, formato dos recursos disponíveis no dataset). Já o módulo M2, analisa os recursos e a conformidade com

[1] https://github.com/ckan/ckan
[2] https://ckan.org/
[3] https://wordpress.com/
[4] https://dkan.readthedocs.io/en/latest/
[5] https://docs.ckan.org/en/2.9/user-guide.html

```
portal_url = 'https://legado.dados.gov.br/'
api_action = f'api/3/action/package_search'

org_id = 'instituto-brasileiro-de-geografia-e-estatistica-ibge'
add_organization = f'fq=organization:{org_id}'


url_ibge_get_resources_qtde = f'{portal_url}{api_action}?{add_organization}'

print(url_ibge_get_resources_qtde)
```

```
https://legado.dados.gov.br/api/3/action/package_search?fq=organization:instituto-brasileiro-de-geografia-e-estatistica-ibge
```

Figura 5.1: modelo de rota CKAN. Fonte: autor.

a LGPD. No exemplo da Figura 5.2, usando o portal de dados aberto do estado de São Paulo[6], o Módulo M1 analisa os pontos 1-4, que são organizações, grupos e metadados (tags, formato e licença do dados) e datasets. E o M2 analisa os pontos 4-5, datasets e, principalmente, recursos.

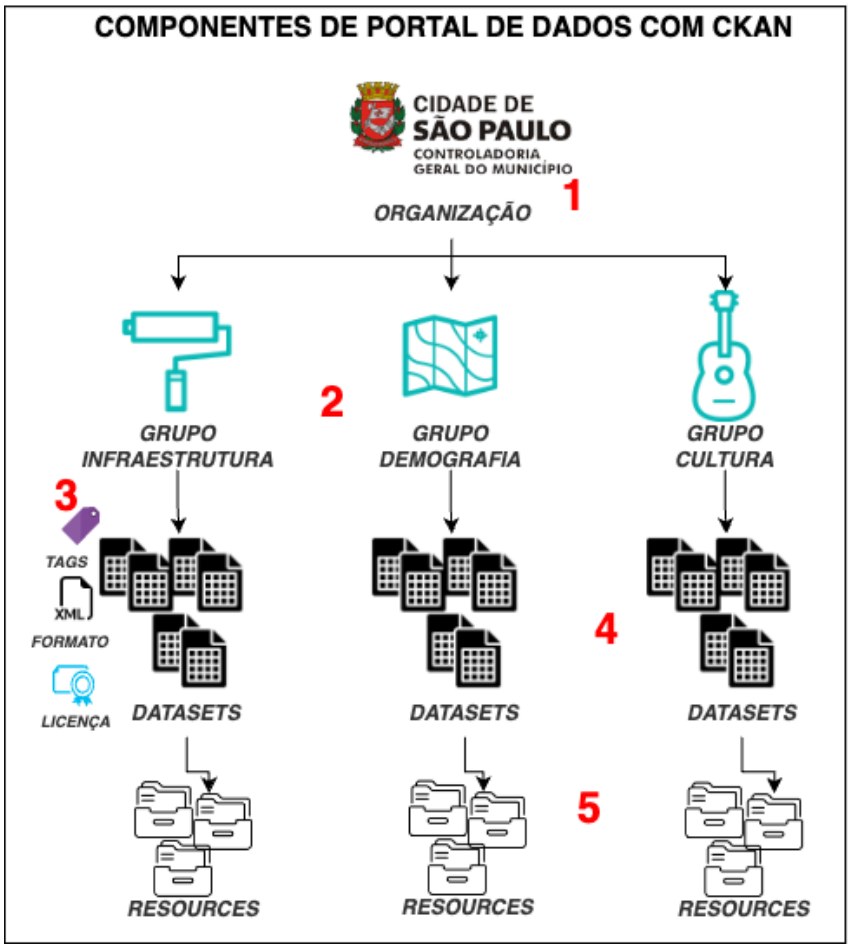


Figura 5.2: componentes de portais de dados abertos com CKAN. Fonte: autor.

Segundo CKAN, os dados são publicados em unidades denominadas "conjuntos de

[6]http://dados.prefeitura.sp.gov.br/

dados". Um conjunto de dados é uma parcela de dados que podem ser, por exemplo, as estatísticas de criminalidade de uma região, os números de gastos de um departamento governamental ou leituras de temperatura de várias estações meteorológicas. Quando os usuários pesquisam dados, os resultados da pesquisa que veem são conjuntos de dados individuais.

Para o CKAN, um conjunto de dados contém duas partes:

- informações ou “metadados” sobre os dados, por exemplo, o título e a editora, a data, os formatos em que estão disponíveis, sob que licença foram lançados;
- uma série de "recursos"(que podem eventualmente, nesta pesquisa, ser chamados *resources*, do nome em inglês) que contêm os próprios dados. O CKAN não se importa com o formato dos dados. Um recurso pode ser uma planilha CSV ou Excel, arquivo XML, documento PDF, arquivo de imagem, dados vinculados em formato RDF, um link, estando o próprio recurso em outro lugar na web. Um conjunto de dados pode conter qualquer número de recursos. Por exemplo, recursos diferentes podem conter dados de anos diferentes ou podem conter os mesmos dados em formatos diferentes.

Um exemplo da disponibilização de datasets e recursos pode ser visto na Figura 5.3, adaptada do portal de dados abertos da Presidência da República do Brasil[7]. Nela é possível verificar que o dataset *Gestores e fiscais de contratos da Presidência da República* (***Nro 1***), pertence ao grupo *Casa Civil da Presidência da República* (***Nro 2***), e que possui inúmeros recursos (***Nro 3***), inclusive em diferentes formatos, como *PDF* e *CSV*.

5.1.4 Framework *CompOD* - Metodologia aplicada ao Módulo M1

A metodologia de pesquisa utilizada é a combinação da abordagem quantitativa e qualitativa em dados de portais de dados abertos brasileiros. Os dados serão categorizados, sumarizados, contabilizados e, posteriormente, será feita a análise através do módulo M1, identificando a qualidade do dado encontrado e das dimensões predefinidas; ao final serão feitas observações percebidas durante a execução da análise.

Os portais de transparência foram considerados e analisados individualmente — quando não encontrado o portal de dados abertos —, mas desconsiderados quando não utilizaram o CKAN para exposição dos dados. Segundo o portal brasileiro de dados abertos, a diferença entre o portal de transparência e o portal de dados abertos é que os portais de transparência têm o objetivo de aumentar o controle das despesas e receitas do governo, enquanto que os portais de dados abertos, por outro lado, têm uma abordagem diferente, o objetivo é ser referência única para busca e acesso a dados públicos brasileiros sobre todo e qualquer assunto ou categoria. Na ausência do portal de dados abertos do estado e do portal de transparência do estado, foi feita uma busca pelo portal de dados abertos das capitais de cada estado.

O trecho de código da Figura 5.4 está relacionado à análise quantitativa (que na Figura 4.1 do modelo, representa o ponto 5), e mostra um exemplo da construção da *URL* alvo (4 primeiras linhas), a recuperação das informações através do contrato *Json* definido (2 linhas seguintes) e o retorno dos formatos de recursos encontrados na busca (demais linhas). Essa é uma das etapas utilizadas na análise quantitativa por portal de dados abertos. Em etapas posteriores, estes dados serão acumulados por tipo, e, ao final da análise, será tirada uma média do formato mais utilizado em portais de dados abertos no Brasil.

[7]https://dadosabertos.presidencia.gov.br/

Figura 5.3: dataset do portal de dados abertos da Presidencia da República. Fonte: dados abertos PR

```python
portal_url = 'https://legado.dados.gov.br/'
api_action = f'api/3/action/package_search'

search_target = 'facet.field=[%22res_format%22]&facet.limit=10&rows=0'
url_target = f'{portal_url}{api_action}?{add_organization}'


#Formatos mais frequentes
formatos = pd.read_json(url_target).get("result")
dados = formatos['facets']['res_format']

fGroup = dict()
def addPlatformfGroup(platform, groupQtde):
    for i in fGroup:
        if i == platform:
            return
    fGroup[platform] = groupQtde


for item in sorted(dados, key = dados.get, reverse=True):
    addPlatformfGroup(item, dados[item])

print(fGroup)
```

Python

```
{'CSV': 9584, 'HTML': 6774, 'JSON': 4300, 'PDF': 3738, 'wsdl': 3084, 'ZIP': 2470, 'KML': 2185, 'GeoJSON': 2158, 'ArcGIS
```

Figura 5.4: trecho de código análise quantitativa, módulo M1. Fonte: autor.

A Figura 5.5 está relacionada à análise qualitativa, e mostra um exemplo de retorno do *jupyter notebook* contendo a lógica de verificação de atualidade dos recursos. De cima para baixo, o método de validação retornará: (1) o conjunto total de dados encontrados;

(2) destes, quantos tinham metadados de *"atualidade"* do recurso; (3) a porcentagem dos que contém metadados; (4) quantos conjuntos de dados estão atualizados; (5) quantos desatualizados; (6) e, por fim, a porcentagem de atualizados dos conjuntos de dados encontrados por organização.

```python
    # National Statistics+National Map
    # IBGE

    ibge_org_id = 'instituto-brasileiro-de-geografia-e-estatistica-ibge'
    return_freshness_info(br_portal_url, br_api_action, ibge_org_id)
```

Python

```
Total of datasets:  424
Total of datasets with freshness info available:  53
Percentage of dataset with freshness information: 12.5 %
Up to date Datasets:  2
Not Up to date Datasets:  51
Percentage of updated information (of the total that informed freshness): 3.77 %
```

Figura 5.5: trecho de código análise qualitativa, módulo M1. Fonte: autor.

5.1.5 Framework *CompOD* - Metodologia aplicada ao Módulo M2

O principal alvo do módulo M2 são os **recursos** (conforme explicado na Seção 5.1.3) disponíveis em portais de dados abertos. A Figura 5.6 demonstra um trecho da classe de extração (equivalente à etapa de *ingestão* demonstrada no modelo da Figura 4.3).

```
13
14  def build_url(portal_url, rows_per_page, start_at):
15      portal_url = f'{portal_url}/api/3/action/'
16      ckan_getable_function = f'package_search?fq=res_format:CSV&rows={rows_per_page}&start={start_at}'
17      get_url = f'{portal_url}{ckan_getable_function}'
18
19      return get_url
20
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS zsh

```
https://dados.ba.gov.br/api/3/action/package_search?fq=res_format:CSV&rows=20&start=0
```

Figura 5.6: Modelo de URL alvo, M2. Fonte: Autor

Nas linhas 15 e 16 são recebidos os parâmetros da função python *"build url"*; os parâmetros são: a <*URL do portal*>, a quantidade de linhas por página <*rows per page*>, e o parâmetro <*start at*>, que indica a página para carregamento das linhas. Os valores são concatenados (linha 17), e geram a URL em destaque como retorno da função, que será o alvo da requisição GET. Essa busca, se feita manualmente, equivaleria à Figura 5.7, feita no portal de dados abertos de São Paulo[8].

Na próxima etapa, a classe de extração irá iterar em todos os *datasets* encontrados (no caso do exemplo da Figura 5.7, seriam 124 execuções). Ao acessar o dataset, a classe de extração acessará individualmente cada recurso, conforme os destacados na Figura 5.8 para o dataset de exemplo, *Remuneração dos servidores aposentados da prefeitura de*

[8]http://dados.prefeitura.sp.gov.br/pt_PT/dataset?res_format=CSV

Figura 5.7: busca manual por termos em portais de dados abertos. Fonte: Portal de Dados Abertos de São Paulo

São Paulo. Por fim, as rotas de todos os recursos do tipo CSV serão armazenados num arquivo de texto *.txt* temporário, conforme a Figura 5.9, que servirá de insumo para a etapa seguinte, a de *processamento*.

Na fase de processamento, o *output* da fase anterior é utilizado como base para o download de recursos, acessados através de requisições *HTTP GET*. O recurso é recebido num dataframe pandas, e os seguintes passos são executados:

- deleção de registros nulos;
- deleção de registros duplicados;
- filtro apenas pelo atributo *<termo de busca>* informado pelo usuário, conforme a Figura 5.6 (caso exista no recurso baixado);
- formatação dos valores e, caso seja numérico, serão removidos caracteres não numéricos;
- remoção de registros mal formados;
- remoção de linhas vazias.

Figura 5.8: recursos disponíveis por dataset. Fonte: Portal de Dados Abertos de São Paulo

```
[
'https://dadosabertos.ibama.gov.br/dados/SINAFLOR/poaAmazoniaLegal.csv'
, 'URL: http://siscom.ibama.gov.br/geoserver/publica/ows?service=WFS&version=1.0.0&request=GetFeature&typeName=publica:veg
, 'http://siscom.ibama.gov.br/geoserver/publica/ows?service=WFS&version=1.0.0&request=GetFeature&typeName=publica:veg_mata
, 'http://siscom.ibama.gov.br/geoserver/publica/ows?service=WFS&version=1.0.0&request=GetFeature&typeName=publica:veg_pamp
, 'http://siscom.ibama.gov.br/geoserver/publica/ows?service=WFS&version=1.0.0&request=GetFeature&typeName=publica:veg_pant
, 'http://siscom.ibama.gov.br/geoserver/publica/ows?service=WFS&version=1.0.0&request=GetFeature&typeName=publica:veg_caat
, 'https://dadosabertos.ibama.gov.br/dados/AATIPP/autorizacao_empresa/DF/2016.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SINAFLOR/poaOutrosBiomas.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_suspensao/suspensao.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_apreensao/termo_apreensao.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_apreensao/bem_apreendido.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_apreensao/especime_apreendido.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_apreensao/anexo.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/auto_infracao/auto_infracao/auto_infracao.csv'
, 'httpS://dadosabertos.ibama.gov.br/dados/SIFISC/auto_infracao/enquadramento/enquadramento.csv'
, 'httpS://dadosabertos.ibama.gov.br/dados/SIFISC/auto_infracao/especime/especime.csv'
, 'httpS://dadosabertos.ibama.gov.br/dados/SIFISC/auto_infracao/coordenada/coordenada.csv'
, 'httpS://dadosabertos.ibama.gov.br/dados/SIFISC/auto_infracao/bioma/bioma.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/auto_infracao/anexo/anexo.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/termo_embargo/termo_embargo.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/itens/itens.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/coordenadas/coordenadas.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/anexo/anexo.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/decisao/decisao.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/enquadramento/enquadramento.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/enquadramento_complementar/enquadramento_complementar.csv'
, 'https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/termo_embargo_historico/termo_embargo_historico.csv'
]
```

Figura 5.9: output final da etapa de ingestão, M2. Fonte: autor.

Ao final da limpeza, o dataset processa o recurso recebido e salva num arquivo de saída no formato CSV, conforme a Figura 5.10 (os CPFs foram ocultados, pois são reais e foram encontrados em portais de dados abertos de órgãos governamentais do Brasil).

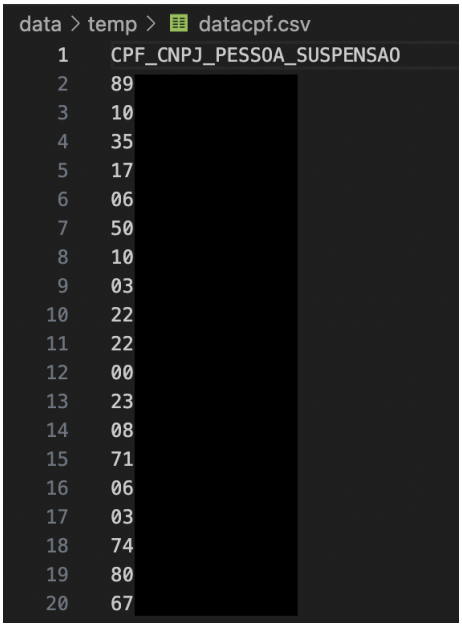


Figura 5.10: Output 01 da Etapa de Processamento, M2. Fonte: Autor

Para a verificação da existência de dados pessoais nos conjuntos de dados, foi utilizado processamento de linguagem natural (na sigla em inglês NLP), pois, de acordo com (SILVA et al., 2020), dentro do domínio aprendizado de máquina (na sigla em inglês ML), o processamento de NLP e o Reconhecimento de Entidades Nomeadas (na sigla em inglês NER) permitem um monitoramento e detecção transparentes de PII, revelando, assim, possíveis violações de privacidade.

Segundo (WU et al., 2008) o algorítimo de classificação de texto, conhecido como máquina de vetores de suporte (na sigla em inglês SVM) é considerado "obrigatório", pois oferece um dos métodos mais robustos e precisos entre todos os algoritmos conhecidos. Os autores também afirmam que métodos eficientes para treinamento de SVM também estão sendo desenvolvidos em ritmo acelerado.

Então, para o treinamento do modelo SVM, as seguintes etapas foram executadas:

- ingestão do dataset de treino, no modelo da Figura de exemplo 5.11, com dados sintéticos de CPF;
- ajuste de variáveis para o treinamento do modelo. Na Figura 5.11, a amostra (comumente referenciado pelo nome em inglês *sample*) e o objetivo (comumente referenciada pelo nome em inglês *target*) são definidos. Além disso, os dados são divididos em 70% para treino e 30% para testes;
- as variáveis categóricas são encodadas para numéricas, para que a máquina leia, ficando, após o encode, como na Figura 5.12, e não mais como na Figura de exemplo 5.11;
- o modelo SVM é treinado com os dados configurados nas etapas anteriores, e ficará disponível na fase final, de *disponibilização e métricas*;

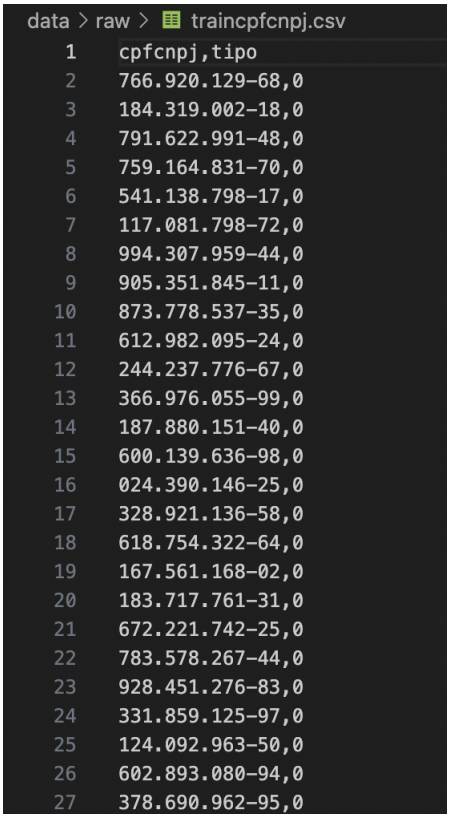
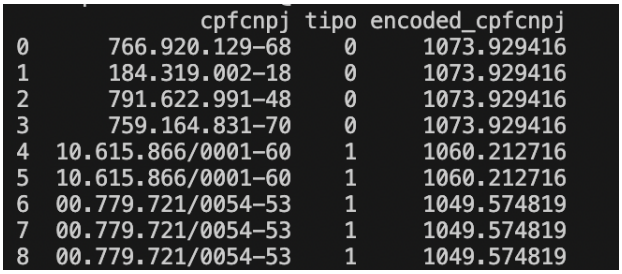


Figura 5.11: dataset sintético com dados de treino para o aprendizado de máquina, M2. Fonte: autor.



	cpfcnpj	tipo	encoded_cpfcnpj
0	766.920.129-68	0	1073.929416
1	184.319.002-18	0	1073.929416
2	791.622.991-48	0	1073.929416
3	759.164.831-70	0	1073.929416
4	10.615.866/0001-60	1	1060.212716
5	10.615.866/0001-60	1	1060.212716
6	00.779.721/0054-53	1	1049.574819
7	00.779.721/0054-53	1	1049.574819
8	00.779.721/0054-53	1	1049.574819

Figura 5.12: variáveis encodadas para o treino do aprendizado de máquina, M2. Fonte: autor.

Na fase de disponibilização, o resultado da busca por dados pessoais em datasets será retornado ao usuário após a inserção dos seguintes valores: endereço (URL) do portal de dados abertos e termo que se deseja buscar. Atualmente, os termos possíveis de serem buscados são: CPF; nome e sobrenome; coordenadas (GPS); e o protocolo de internet, conhecido como IP, conforme a Figura 5.13.

A tela de busca após escaneamento e classificação do texto encontrado nos recursos acessados dentro do portal de dados abertos analisado pode ser vista na Figura 5.14. Neste caso, a busca encontrou CPF. Com esse resultado, o retorno conterá: (1) a organização responsável pelo recurso encontrado; (2) o título da licença encontrado; (3) a classificação como aberto; (4) o id do dataset (no CKAN identificado como pacote (ou em inglês *package*); (5) a URL do recurso onde o CPF foi encontrado; e (6) a URL do dataset.

Conforme a Figura 5.14, verificam-se ao menos 6 recursos que possivelmente contêm CPFs expostos — é utilizada a palavra *possivelmente* porque o modelo pode predizer

Figura 5.13: opções de termo de busca, M2 - CompOD. Fonte: autor.

erroneamente. Para verificação e comprovação, foi feita uma busca manual no primeiro recurso retornado na lista em destaque, o de *resource route*, ou seja, URL do recurso, conforme a Tabela 5.3. O retorno pode ser verificado na Figura 5.15, onde o modelo fez a classificação corretamente. O CPF não apenas está exposto, como está atrelado ao nome completo (dado pessoal e sensível) da pessoa (ambas as informações foram cobertas com retângulo preto, pois são dados reais que estão expostos na internet).

Tabela 5.3: rota de recurso retornado pelo modelo

https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/decisao/decisao.csv

5.2 Limitações da metodologia

O modelo utilizado na classificação de palavras é o SVM, que é um modelo de aprendizado de máquina amplamente utilizado para esta finalidade. No entanto, como é conhecido, algoritmos de predição não possuem, em sua maioria, 100% de acurácia. Logo, uma limitação evidente do estudo são as predições imprecisas e falsos positivos que o modelo pode gerar. Como no exemplo (real) da Figura 5.14, onde, no retorno de rotas com CPFs possivelmente expostos, na terceira linha do retorno, a rota de recurso 5.4 na verdade não apresenta CPFs expostos, e sim protegidos, como pode ser visto na Figura 5.16.

Tabela 5.4: rota de recurso com predição imprecisa retornado pelo modelo

https://dadosabertos.ibama.gov.br/dados/AATIPP/autorizacao_empresa/DF/2016.csv

Atualmente, o framework de verificação de conformidade com a LGDP, valida uma funcionalidade por vez. Ou seja, é feita apenas uma validação por vez: ou CPF, ou nome+sobrenome, ou IP, ou GPS. No entanto, como visto na Figura 5.15, é possível verificar que, além de retornar o CPF, o recurso continha também o nome completo das pessoas, assim como no exemplo real, da Figura 5.17, onde está exposto, além do nome completo da pessoa, o e-mail pessoal, combinação capaz de identificar uma pessoa, uma vez que e-mails são únicos.

5.3 Escopo do experimento - análise aplicada aos módulos M1 e M2 do framework *CompOD*

As dimensões analisadas para os experimentos do módulo M1 foram baseadas no OFKN que, como já mencionado, trata-se de um esforço anual para medir o estado dos dados governamentais abertos em todo o mundo. O módulo M1 fará a análise de quantidade e qualidade dos dados, conforme as dimensões:

- estatísticas nacionais;
- orçamento do governo;
- gastos públicos;
- legislação;
- resultados eleitorais;
- mapa nacional;
- emissões poluentes;
- cadastro de empresa;
- conjuntos de dados de localização;
- licitações de compras governamentais (passadas e atuais);
- qualidade da água;
- previsão do tempo;
- propriedade da terra;
- desempenho em saúde.

Os órgãos que representam as dimensões, ou seja, de onde os dados serão extraídos, estão na Tabela 5.5. No módulo M1 não haverá classificação e contagem de pontos como produzido no OKFN, mas o resultado da análise será mostrado para todos os conjuntos de dados disponíveis em portais de dados abertos.

Tabela 5.5: agências governamentais por dimensão

Dimensão	Agências governamentais
Estatísticas Nacionais	IBGE/IPEA
Mapa Nacional	IBGE
Legislação	Senado Federal
Resultados Eleitorais	TSE
Orçamento e Gastos do Governo	CGU/TCU/BNDES/ME
Emissões de Poluentes	MMA/IBAMA
Cadastro de Empresa	DREI
Conjuntos de dados de localização	
Licitações de compras governamentais (passadas e presentes)	ME/MGISP/SIASG
Qualidade da Água	ANA
Previsão do tempo	INPE/INMET
Propriedade da Terra	INCRA/SNCR/CAFIR
Desempenho em saúde	Ministério da Saúde

Além disso, os portais oficiais de dados abertos brasileiros serão alvo de análise dos módulos M1 e M2; Sendo que o módulo M1 fará a análise de quantidade e qualidade, e o M2 fará a verificação de conformidade com a LGPD. Os dados analisados estão no portal de dados abertos do Brasil e Distrito Federal (ver Tabela 5.6). Na data da redação da dissertação, o portal brasileiro estava passando por uma reformulação e, para compatibilidade com o CKAN, foi utilizada a URL legada do portal *legado.dados.gov.br* nos experimentos do módulo M1, e não a principal, *dados.gov.br*, pois esta não continha, à época, exposição através do CKAN. E também foram analisados os portais de cada um dos 26 estados do Brasil (ver Tabela 5.7).

Tabela 5.6: portais de dados abertos do Brasil e Distrito Federal

Distrito	Acronimo	URL do Portal de Dados
Distrito Federal	DF	http://www.dados.df.gov.br/
Brazil	BR	*https://legado.dados.gov.br/

5.4 Resultados - módulo M1

A análise dos dados pelo módulo M1 ocorreu no primeiro semestre de 2023, e abrangeu as questões apresentadas na Tabela 5.8, que, assim como suas descrições, estão de acordo com a metodologia do Indice Global de Dados Abertos (GODI)[9].

Através do módulo M1, o objetivo é disponibilizar o panorama da saúde dos dados abertos no Brasil, como, por exemplo, quais estados e organizações brasileiros abrem seus dados; Destes, quais utilizam padronização no momento da disponbilização dos dados, e, também, qual a qualidade desses dados disponibilizados. Além disso, foram aplicadas

[9]http://index.okfn.org/methodology/

Tabela 5.7: portal de dados abertos por estado

Estado	*Portal de Dados Abertos	Usa CKAN
Acre	Sim	Não
Alagoas	Sim	Sim
Amapá	Sim	Não
Amazonas	Sim	Não
Bahia	Sim	Sim
Ceará	Sim	Não
Espírito Santo	Sim	Sim
Goiás	Sim	Sim
Maranhão	Sim	Não
Mato Grosso	Sim	Não
Mato Grosso do Sul	Sim	Sim
Minas Gerais	Sim	Sim
Pará	Sim	Não
Paraíba	Sim	Não
Paraná	Sim	Não
Pernambuco	Sim	Sim
Piauí	Sim	Não
Rio de Janeiro	Sim	Não
Rio Grande do Norte	Sim	Não
Rio Grande do Sul	Sim	Sim
Rondônia	Sim	Não
Roraima	Sim	Não
Santa Catarina	Sim	Sim
São Paulo	Sim	Sim
Sergipe	Sim	Não
Tocantins	Sim	Não

* *ou portal de transparência*

verificações como: se o portais possuem seus recursos em formatos legíveis por máquina — para permitir análises sistemáticas; qual a quantidade de grupos/temas existentes no portal; e qual a qualidade dos metadados.

Os pontos analisados foram (1) quantidades de grupos/tema por plataforma (Seção 5.4.1.1); (2) número de conjuntos de dados por plataforma (Seção 5.4.1.2); (3) formatos de dados disponíveis mais comumente usados, por plataforma (Seção 5.4.1.3); (4) qualidade dos dados por dimensão (Seção 5.4.2).

5.4.1 Analise quantitativa

5.4.1.1 Grupos/temas por plataforma

Por meio de grupos, é possível encontrar conjuntos de dados classificados por temas, por exemplo, o grupo 'Governo e Política' do portal brasileiro de dados abertos traz dados relativos aos censos legislativos, dados sobre a estrutura organizacional do poder executivo federal etc. A diversidade de grupos, desde que bem estruturada, potencializa o uso correto dos dados, além de ajudar a mapear quais grupos estão com cadastro pendente e estimular os responsáveis por eles a cadastrá-los no portal.

Tabela 5.8: perguntas da metodologia de qualidade do GODI

Pergunta	Contexto
Os dados existem?	Podem estar em qualquer formato (papel ou digital, offline ou online etc.)
Os dados estão em formato digital?	Esta questão aborda se os dados estão em formato digital (armazenados em computadores ou armazenamento digital) ou se estão apenas, por exemplo, em formato de papel
Estão disponíveis publicamente?	Esta questão aborda se os dados são "públicos". Isto não exige que estejam disponíveis gratuitamente, mas exige que alguém fora do governo possa acessá-los de alguma forma. Se for necessário um pedido de liberdade de informação ou similar para acessar os dados, este não é considerado público
Os dados estão disponíveis gratuitamente?	Esta questão aborda se os dados estão disponíveis gratuitamente ou se há cobrança
Os dados estão disponíveis online?	Esta questão aborda se os dados estão disponíveis online a partir de uma fonte oficial
Os dados são legíveis por máquina?	Os dados são legíveis por máquina se estiverem em um formato facilmente estruturado por um computador. Os dados podem ser digitais, mas não legíveis por máquina. Por exemplo, considere um documento PDF contendo tabelas de dados. Estes são digitais, mas não são legíveis por máquina porque um computador teria dificuldade para acessar as informações tabulares
Os dados estão disponíveis em massa?	Os dados estarão disponíveis em massa se todo o conjunto de dados puder ser baixado ou acessado facilmente. Por outro lado, é considerado não em massa se os cidadãos estiverem limitados a obter apenas partes do conjunto de dados (por exemplo, se estiverem restritos a consultar um formulário web e recuperar alguns resultados de cada vez a partir de uma base de dados muito grande)
O dados possui licenças *abertas*?	Esta questão aborda se o conjunto de dados é aberto conforme a *Open Definition (http://opendefinition.org)*. É preciso indicar os termos de uso ou licença que permitem a qualquer pessoa usar, reutilizar ou redistribuir livremente os dados (sujeitos, no máximo, a requisitos de atribuição ou compartilhamento). É vital que uma licença esteja disponível (se não houver licença, os dados não serão licenciados abertamente). As licenças abertas que atendem aos requisitos da Definição Aberta estão listadas em *http://opendefinition.org/licenses/*
Os dados são fornecidos em tempo hábil e estão atualizados?	Esta questão aborda se os dados estão atualizados, ou se estão muito atrasados

De acordo com a Figura 5.18, das plataformas analisadas, a que possui mais grupos é a do estado de São Paulo (39), seguida por, Espírito Santo (26), Rio Grande do Sul (25), Santa Catarina (25), Bahia (21), Mato Grosso do Sul (12), Distrito Federal (11), Alagoas (10), Minas Gerais (4), Pernambuco (3) e Goiás (0). Excluindo o maior e o menor valor

(menor valor maior que zero), as plataformas possuem uma média aproximada de 17,7 grupos cadastrados, sendo que o maior grupo é quase seis vezes maior que o menor grupo.

5.4.1.2 Número de conjuntos de dados por plataforma

Conjuntos de dados são os dados disponíveis por cada plataforma/grupo, por exemplo, conjuntos de dados do governo federal brasileiro são a *Declaração Anual de Uso de Recursos Hídricos - DAURH; o Número de Unidades Básicas de Saúde em construção* e as *Contratações Públicas do Governo Federal*.

Tão importante quanto uma grande variedade de grupos (ou temas) em um portal de dados, é a quantidade de conjuntos de dados (ou, em inglês, *dataset*) por grupo, pois fornece informações sobre o tema escolhido, possibilitando a utilização dos dados do grupo para diversos fins, que vão desde fiscalização e transparência, até utilização para implementação de novas aplicações em prol da sociedade.

De acordo com a Figura 5.19, das plataformas analisadas, as que possuem o maior número de conjuntos de dados são São Paulo (484), Alagoas (332), Santa Catarina (302), Rio Grande do Sul (302), Distrito Federal (161), Espírito Santo (95), Mato Grosso do Sul (44), Goiás (33), Pernambuco (27), Minas Gerais (26) e Bahia (11). Excluindo o maior e o menor valor (menor valor maior que zero), em média, as plataformas possuem 1.322 conjuntos de dados. E o maior conjunto de dados é quase 12 vezes maior que o menor conjunto de dados. Também é possível verificar que, embora o Espírito Santo possua maior variedade de grupos (Figura 5.18), o estado de Alagoas possui maior disponibilidade de conjuntos de dados.

5.4.1.3 Formatos de dados disponíveis mais comumente usados

De acordo com o GODI, uma das métricas utilizadas na avaliação de conjuntos de dados é se estes são legíveis por máquina. Os formatos considerados legíveis são: formato de arquivo Microsoft Excel, um formato de arquivo de planilha abreviado como XLS; valores separados por vírgula, formato abreviado como CSV, constituído de um arquivo de texto que permite que os dados sejam salvos em estrutura de tabela; JavaScript Object Notation, mais comumente conhecido pela sigla JSON, que é um formato aberto de intercâmbio de dados legível por humanos e por máquinas; a linguagem de marcação extensível XML, que é um conjunto de códigos, ou tags, que descrevem o texto em um documento digital.

Na análise quantitativa foi verificado, cruzando os dados de todos os portais abertos analisados, qual a quantidade total de conjunto de dados

Cruzando os dados de todos os portais de dados abertos analisados, foi possível verificar a quantidade total de conjuntos de dados legíveis por máquina. O resultado dos 3 formatos mais utilizados é mostrado na Figura 5.20, onde 70,45% dos dados estão em formato CSV, 29,08% dos dados estão no formato JSON e 0,46% dos dados estão no formato XLS.

5.4.2 Analise qualitativa

Cada subseção nas nesta seção apresentará as questões do OKFN abordadas pela análise, especificadas na Tabela 5.8, para as dimensões que contêm conjuntos de dados disponíveis suficientes. Os órgãos governamentais que contêm os dados das dimensões analisadas podem ser vistos na Tabela 5.5 e terão sua análise realizada nas seções seguintes.

Pela característica do estudo — análise apenas em portais de dados abertos — as per-

guntas: *'Os dados estão em formato digital?'*, *'Disponíveis publicamente?'*, *'Os dados estão estão disponíveis gratuitamente?'*, *'Os dados estão disponíveis online?'*, da Tabela 5.8, têm em comum a mesma resposta afirmativa, que é: todas as dimensões onde os dados existem foram encontradas em portais de dados abertos: seja no portal de dados abertos brasileiro ou no portal de dados abertos específico da organização governamental. Portanto, a resposta comum para as dimensões: Estatísticas Nacionais, Mapa Nacional, Resultados Eleitorais, Orçamento e Despesas do Governo, Registo de Empresas, Conjuntos de dados de localização, Licitações de compras governamentais, Qualidade da Água, Previsão do Tempo e Propriedade da Terra, é: os dados estão disponíveis digitalmente, também estão disponíveis publicamente (*no novo portal brasileiro de dados abertos é necessário se cadastrar e depois fazer login*) e, por fim, estão disponíveis online e gratuitamente.

5.4.2.1 Os dados existem?

Para a análise *'Os dados existem?'*, todas as dimensões analisadas — exceto *'Conjuntos de dados de localização'* e *'Registro de Empresa'* — possuem seu próprio portal de dados aberto ou expõem seus dados no Portal brasileiro de dados abertos. Ou seja, em 83% das dimensões analisadas, os dados existem.

Para as dimensões *Cadastro de Empresas*, o órgão público é o DREI (instituto vinculado ao Ministério da Economia), que não possui portal próprio de dados abertos, nem possui dados disponíveis no portal brasileiro de dados abertos. Porém, o DREI possui site próprio, com planos de abertura de dados vinculados à agenda do Ministério da Economia, o que ainda não foi implementado. Para a dimensão *Location datasets*, não foram encontrados dados abertos em nenhum portal ou site de órgão público.

5.4.2.2 Os dados podem ser lidos por máquina?

Para validar esta categoria, o formato dos dados encontrados foi verificado e contabilizado, desde que estivessem de acordo com a definição do Índice Global de Dados Abertos para formatos legíveis por máquina, ou seja, que sejam dos tipos <'XLS', 'CSV', 'JSON', 'XML'>.

Algumas organizações governamentais, como o *Senado Federal*, tinham poucos (menos de 100) conjuntos de dados disponíveis no Portal brasileiro de dados abertos ou em seu próprio portal. Cenários que podem causar isso: a organização não possui dados expostos através da API CKAN, ou estão disponíveis em outro modelo não padrão, inviabilizando a análise sistemática. Isto fará com que o resultado não seja relevante para a dimensão em análise. Quando isso acontecer, será sinalizado no texto. O resultado de dados legíveis por máquina para cada dimensão são:

- Estatísticas Nacionais e Mapa Nacional: no total, foram avaliados 430 conjuntos de dados, sendo 424 do IBGE e 6 do IPEA. Destes, 373 eram CSV, o que equivale a 86,74% dos formatos disponíveis para esta dimensão, seguido de 52 recursos disponíveis do formato JSON, e 49 do formato XML.

- Legislação: à disposição do Senado Federal, havia apenas dois conjuntos de dados, e um total de 3 recursos, sendo 2 legíveis por máquina: 1 CSV e 1 XLS. Para que a leitura não seja tendenciosa, não serão fornecidos percentuais dos casos em que o volume de dados é insignificante, já que, quando a quantidade de recursos é baixa, o percentual daquele legíveis por máquina é alto.

- Resultados Eleitorais: no total, foram avaliados 144 conjuntos de dados do TSE, e e 169 recursos foram utilizados com eles. Dos recursos disponíveis, 95 eram CSV, o equivalente a 64,37% dos recursos disponíveis para esta dimensão. Dos formatos legíveis por máquina, apenas o CSV estava disponível.

- Orçamento e Gastos do Governo: nesta dimensão, os órgãos governamentais analisados foram o Banco Nacional de Desenvolvimento Econômico e Social (BNDES), a Controladoria-Geral da União (CGU) e o Tribunal de Contas da União (TCU). Para as três instituições, havia 92 conjuntos de dados disponíveis, com 69,56% de CSV.

  Das três instituições, o BNDES se destaca positivamente, com 93 recursos disponíveis em seus 46 conjuntos de dados. Dos 93 recursos disponíveis 47 são legíveis por máquina. Isso significa que, do total de recursos disponibilizados pelo BNDES, 50,54% podem ser utilizados em estudos automatizados com auxílio de software.

- Emissões de Poluentes: As instituições governamentais analisadas foram o Ministério do Meio Ambiente e Mudanças Climáticas (MMA) e o Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA). No total, estavam disponíveis 98 conjuntos de dados e, estes continham 314 recursos, dos quais 212 são legíveis por máquina. No total, para esta dimensão, 67,51% dos recursos disponíveis podem ser utilizados de forma sistemática.

- Qualidade da Água: Com 308 conjuntos de dados disponíveis em seu portal de dados abertos, a Agência Nacional de Águas e Saneamento Básico (ANA) conta com 1.611 recursos, dos quais 15,46% são legíveis por máquina, e CSV é o tipo de arquivo predominante.

- Previsão do tempo: para esta dimensão, o Instituto Nacional de Meteorologia (INMET), do Ministério da Agricultura e Pecuária, foi analisado fornecendo os seguintes formatos de dados legíveis por máquina, com sua respectiva quantidade: 'CSV',27; 'XML',14; 'JSON',12. Isso representa 55,79% de recursos reutilizáveis, com um total de 95, espalhados por 56 conjuntos de dados.

- Propriedade da Terra: o Instituto Nacional de Colonização e Reforma Agrária (INCRA) representa esta seção. No entanto, apenas 1 conjunto de dados estava disponível, e não era legível por máquina.

5.4.2.3 Disponível em massa?

Para validar que os dados estavam disponíveis em massa, foram coletados todos os recursos disponíveis para a dimensão em análise e feita uma solicitação do tipo GET ao endpoint de cada recurso. Sempre que a solicitação retornasse o status HTTP 200, o recurso era contabilizado como acessível, e quando retornasse um valor diferente do status HTTP 200, ou quando a consulta ultrapassasse o tempo limite de 15 segundos para retorno da resposta, o recurso era contabilizado como não disponível.

- Estatísticas Nacionais e Mapa Nacional: IBGE e IPEA juntos possuem 595 recursos, dos quais 504 retornaram status de sucesso, e 91 atingiram o limite máximo de espera de resposta, 15 segundos, ou retornaram status diferente de 200. Dos recursos disponíveis, 84,70% dos volumes são acessíveis.

- Legislação: o Senado Federal possui 16 recursos, e todos retornaram ao status de sucesso.
- Orçamento e Gastos do Governo: CGU, TCU e BNDES juntos possuem 334 recursos, destes 333 são acessíveis e, apenas 1 com volume indisponível. Então, 96,80% dos recursos acessíveis estão retornando status de sucesso para consumo.
- Emissões de Poluentes: o MMA e o IBAMA juntos possuem 335 recursos, destes 128 acessíveis, e 217 (todos do IBAMA) com recurso indisponível ou falho. Portanto, apenas 38,20% dos recursos acessíveis estão retornando status de sucesso para consumo.
- Qualidade da Água: a ANA possui 249 recursos, e os volumes acessíveis são 240. Destes, 9 são volumes indisponíveis ou com falha. Portanto, 96,38% dos recursos acessíveis estão retornando status de sucesso para consumo.
- Previsão do tempo: o INMET dispõe de 325 recursos, dos quais 293 retornaram sucesso. Ou seja, 90,15% dos recursos acessíveis estão retornando status de sucesso para consumo.
- Propriedade da Terra: o INCRA não possui recursos disponíveis.

5.4.2.4 Licenciado abertamente?

Para validar esta categoria, foi consultada, nos metadados, a lista de licenças disponíveis para a organização em análise. Nem todos os órgãos/institutos públicos analisados forneceram esta informação, portanto, o número de licenças disponíveis pode ser menor que a quantidade de conjuntos de dados disponíveis, mas nunca maior, uma vez que a licença é atribuída a um conjunto de dados (e não aos múltiplos recursos que um conjunto de dados pode conter).

- Estatísticas Nacionais e Mapa Nacional (IBGE/IPEA): para esta dimensão havia 372 conjuntos de dados com licença do tipo *'Licença não especificada'*, 36 conjuntos de dados licenciados *'Outros (Domínio Público)'*, seguidos por *'Creative Commons Attribution'*, *'Open Data Commons Open Database License (ODbL)'*, *'Other (Open)'*, cada um com 6 conjuntos de dados e, finalmente, 4 conjuntos de dados com *'Outra licença (Atribuição)'*. Do total de licenças reportadas para esta dimensão, 86,51% possuem licença 'não especificada'.
- Legislação (Senado Federal): para o único conjunto de dados disponível, a licença foi *'Creative Commons Attribution and Share Alike'*.
- Resultados Eleitorais (TSE): dos 144 conjuntos de dados disponíveis, 140, ou 97,22%, utilizam a licença *'Creative Commons Attribution'*, a única disponível para a dimensão.
- Orçamento e Gastos Governamentais (CGU/TCU/BNDES): as licenças disponíveis por conjunto de dados foram 46 sob a licença *'Open Data Commons Open Database License (ODbL)'*, 18 sob a licença *'Licença Não Especificada'*, 13 sob a licença emph'Outro (Aberto)', 9 sob a licença *'Creative Commons Attribution'*, 5 sob a licença *'Outro (Domínio Público)'*. Mais uma vez, a instituição BNDES se destaca como responsável por disponibilizar todos os seus conjuntos de dados (46) sob a licença 'Open Data Commons License (ODbL)'.

- Emissões Poluentes (MMA/IBAMA): as licenças são: 36 sob a licença *'Creative Commons Attribution'*, 35 sob a licença *'Outro (Domínio Público)'*, 20 sob a licença *'Outro (Aberto)'*, 7 sob a licença *'Open Data Commons Open Database License (ODbL)'*. No total, são 98 conjuntos de dados, todos licenciados abertamente.

- Qualidade da Água (ANA): para esta dimensão a consulta aos metadados não retornou licenças disponíveis.

- Previsão do tempo (INMET): para esta dimensão, todos os 56 conjuntos de dados disponíveis possuem as seguintes licenças: 48 sob a licença *'Creative Commons Attribution'*, 6 sob a licença *'Open Data Commons Open Database License (ODbL)'*, 2 sob a licença *'Outro (Aberto)'*.

- Propriedade de Terra (INCRA): para o único conjunto de dados disponível a licença é *'Creative Commons Attribution'*.

5.4.2.5 Os dados são fornecidos em tempo hábil e estão atualizados?

Para esta categoria, buscou-se as palavras-chave "frequência de atualização" nos metadados do conjunto de dados, "frequência de atualização (meses)", "periodicidade" e "frequência de publicação". Como os atributos de peridiocidade e frequência de atualização dos dados não são obrigatórios, alguns órgãos governamentais não os colocaram, ou se o colocaram, não foi de forma padronizada, de modo que, caso haja alguma periodicidade informada que não atenda às palavras-chave mencionadas no início do parágrafo, essa peridiocidade não foi contabilizada.

Entendendo a importância e relevância das informações, foram realizadas buscas manuais por maior diversidade de palavras-chave que pudessem de alguma forma conter a periodicidade. Porém, conforme mencionado, era um atributo não obrigatório e, portanto, não padronizado na forma como era disponibilizado.

Devido ao volume de dados, uma pesquisa manual não é eficaz e suficiente. Quando informada a periodicidade, verificou-se, então, se o valor estava na seguinte lista: ['semestral', 'diário', 'semanal', 'quinzenal', 'mensal', 'bimestral' , 'trimestral', 'anual', 'bienal']. Se sim, o valor do atributo ['metadados modificados'] foi reduzido do valor encontrado na lista. Por exemplo, o conjunto de dados X continha uma periodicidade de atualização 'mensal', portanto, foi verificado se a última atualização do conjunto de dados havia sido feita no último mês. Se o valor retornado para a periodicidade fosse diferente daquele da lista acima, como <frequência de atualização: 'on demand'>, o cálculo era impossível e o conjunto de dados não podia ser contabilizado. A seguir, serão itemizados as análises para as dimensões, de acordo a métrica de atualização dos conjuntos de dados:

- Estatísticas Nacionais e Mapa Nacional (IBGE/IPEA): o IPEA não retornou — em seus metadados — dados de atualização dos conjuntos de dados. Quanto ao IBGE, dos seus 424 conjuntos de dados, 53 possuíam informações de atualização, dos quais apenas 2 estavam atualizados de acordo com os cálculos feitos entre a janela de atualização informada e a última modificação do conjunto de dados. Isso representa 3,77% dos conjuntos de dados atualizados, do total que continha dados de atualização.

- Legislação (Senado Federal): para o único conjunto de dados disponível, não havia informações atualizadas nos metadados.

- Resultados Eleitorais (TSE): dos seus 144 conjuntos de dados, 20 continham informações de atualização, e nenhum – segundo a janela de atualização informada – estava atualizado.

- Orçamento e Gastos do Governo (CGU/TCU/BNDES): para o TCU, não havia informações atualizadas em seus conjuntos de dados. Para a CGU, dos seus 43 conjuntos de dados, 12 possuíam informações de atualização, ou seja, 27,91%. Destes, nenhum dos 12 estava atualizado. Quanto ao BNDES, 58,7% dos seus conjuntos de dados continham informações atualizadas. Ou seja, de um total de 46, 27 possuíam metadados com informações de atualização e estavam, de fato, atualizados.

- Emissões de Poluentes (MMA/IBAMA): para o MMA, 6,06% dos conjuntos continham informações de atualização — 4 de um total de 66 conjuntos — contudo, apenas 1 destes estava atualizado, representando 25% do total. Esse cenário deve ser analisado com atenção porque, apesar das percentagens de atualizações serem relativamente consideráveis, a quantidade total de conjuntos de dados disponíveis é inferior a 100, o que é uma quantidade baixa. Para o IBAMA, 56,25% dos conjuntos continham informações de atualização — 18 de uma total de 32 conjuntos — e todos estavam atualizados.

- Qualidade da Água (ANA): apesar de contar com 300 conjuntos de dados — uma quantidade alta comparada a outros órgãos governamentais —, a ANA está com todos os conjuntos de dados desatualizados, o que chamou a atenção durante este estudo. Devido à qualidade dos resultados do ANA em outras análises, foi verificado manualmente que de fato este campo não existe nos metadados dos conjuntos de dados, portanto, neste caso, não foi possível verificar se os dados estão atualizados ou não.

- Previsão do tempo (INMET): o INMET possui 56 conjuntos de dados, dos quais 13, ou 23,21%, contam com informações de atualização. Apenas 1 conjunto de dados estava desatualizado, representando 7,69%.

- Propriedade da Terra (INCRA): o INCRA possui apenas um conjunto de dados disponível e nenhuma informação atualizada.

5.4.3 Percepções da análise

A análise quantitativa validou 1.817 conjuntos de dados, de 196 grupos, contidos em 19 portais de dados abertos – subdivididos em 12 portais de dados abertos dos estados; 2 portais de dados abertos no Brasil e no Distrito Federal; 5 portais de órgãos governamentais com portais próprios de dados abertos (TSE/BNDES/MMA/IBAMA/ANA).

Embora o Brasil tenha 26 estados, apenas 11, menos da metade, possuem portal de dados abertos com exposição padrão utilizando CKAN – amplamente utilizado em países que expõem seus dados. Porém, a ocorrncia de alguns estados com portal de dados abertos, ainda que sem exposição de dados com padronização, como o Rio de Janeiro, demonstra disposição e compreensão da importância desse assunto.

De todo modo, o conhecimento do papel dos dados abertos evidencia uma das principais dificuldades relatadas no estudo: a falta de promoção e divulgação da importância de se ter padrões na exposição dos dados, para que seja possível reutilizar dados, ou aplicar estudos sistemáticos, por exemplo, através de múltiplos portais de dados em diferentes continentes.

A análise das dimensões foi importante para um aprofundamento nos dados, de acordo com questões importantes para a população, como emissões de poluentes e estatísticas nacionais. Para esta categoria, foram realizadas buscas manuais para verificar a existência de portais de dados abertos de órgãos do governo federal, responsáveis por disponibilizar esses serviços aos cidadãos brasileiros. Por exemplo, a dimensão Estatística Nacional foi representada pelos órgãos federais IBGE e IPEA e a dimensão Emissão de Poluentes foi representada pelas instituições MMA e IBAMA.

A partir das análises, foi possível perceber que, apesar de possuírem portais de transparência, alguns órgãos federais importantes, como Senado Federal, não possuem portais próprios de dados abertos, constituindo portais para fins diferentes. Além de os dados analisados precisarem ser acessados através do portal brasileiro de dados abertos, os conjuntos de dados do Senado Federal disponíveis (inferiores a 100) representam uma quantidade baixa quando comparada a outras organizações, como o IBGE, que contém mais de 400 conjuntos.

Em relação à categoria 'os dados são legíveis por máquina', das dimensões que possuíam volume relevante de dados (excluindo Legislação e Propriedade da Terra), todas tinham pelo menos 50% de seus dados legíveis por máquinas, ou seja, passíveis de reutilização. A dimensão Qualidade da Água foi a exceção, com 15%. Já em relação à categoria 'Disponível em massa?', das dimensões que possuíam dados (excluindo Propriedade da Terra), todas tiveram pelo menos 80% dos seus recursos acessados com sucesso. Apesar do elevado volume, 80% de recursos disponíveis em massa, o valor precisa ser olhado com atenção, pois ter 80% dos dados disponíveis não significa um volume relevante de recursos, mas que as possibilidades de acesso com sucesso eram grandes para os recursos disponíveis.

As duas últimas categorias avaliadas foram as que apresentaram os piores indicadores para todas as dimensões, representando as dificuldades de leitura em materiais relacionados, a saber: licenciamento pouco claro, pois, em sua maioria, as dimensões analisadas possuem licenças do tipo *'License não especificado'*, *'Licença não especificada'* e *'Outro (Aberto)'*, quando especificado; e recursos desatualizados, pois menos de 10% dos recursos mostraram-se atualizados.

5.5 Resultados - módulo M2

A análise dos dados pelo módulo M2 ocorreu no segundo semestre de 2023. Logo, apesar da massa de dados ser, em parte, a mesma do módulo M1, alguns recursos/portais podem não estar mais dispoíveis para análise, ou, por outro lado, ter sido incluídos/adicionados pelos estados e estar disponíveis para validação.

O objetivo principal do experimento foi verificar os estados brasileiros que expõem seus dados, via CKAN, levando em consideração a lei geral de proteção de dados brasileira (LGPD)[10]. Para isso, foram escaneados os *recursos* disponíveis nos alvos mapeados, e foi verificada a existência de propriedades que contivessem CPF (de acordo com a LGPD, considerado dado pessoal, e que, portanto, deve ser protegido). Considerando que os experimentos do módulo M1 demonstraram que CSV é o formato com mais dados disponíveis, a busca deste módulo filtrou apenas recursos nesse formato.

As tabelas de resultados nas próximas seções seguirão o modelo conforme:

- Tabela 5.9 de rotas e status, contendo os parâmetros:

[10]https://www.planalto.gov.br/ccivil_03/_ato20152018/2018/lei/l13709.htm

  - *Alvo*: o objeto da análise — por exemplo, *Ministério da Saúde*;
  - *URL do portal de dados*: endereço de acesso do portal;
  - *Status de retorno*: o status retornado no momento do experimento, se sucesso ou falha.

- Tabela 5.10 de totais de recursos analisados, com os parâmetros:
  - *Alvo*: o objeto da análise — por exemplo, *Ministério da Saúde*;
  - *Qtde de datasets*: o número de datasets encontrados no portal;
  - *Qtde de recursos*: a quantidade de recursos nos datasets do portal.

- Tabela 5.11 com datasets em discordância com a LGPD, com os parâmetros:
  - *Alvo*: *X* opcional para cada seção;
  - *Datasets*: endereço de acesso do dataset. Para acessar o dataset, basta substituir o *{PORTAL_URL}* pelo valor correspondente para o *Alvo*;
  - *Qtde de Recursos*: a quantidade de recursos no referido dataset, identificados como contendo ao menos um registro com CPF desprotegido;
  - *Total de Linhas*: o total de linhas em cada recurso do dataset. É importante ressaltar que cada recurso pode conter de dezenas a milhares de linhas, e que cada linha pode conter dados pessoais expostos. Porém, a contagem de linhas do recurso tem o objetivo principal de mostrar quantos dados pessoais poderiam estar expostos naquele recurso. Por exemplo, um recurso (CSV) que tem CPF/CNPJs na mesmma coluna, pode conter 100 linhas, sendo apenas 01 com CPF desprotegido, e os demais com CNPJs. Deste modo, a contagem de linhas terá o valor de 100, mas apenas 1 CPF. Assim como existem cenários onde todas as linhas analisadas do recurso eram CPFs desprotegidos.

Tabela 5.9: modelo - Rotas dos alvos analisados e seus status de acesso

Alvo	URL do Portal de Dados	Status de Retorno
XXXX	https://XXXXX	Falha ou Sucesso

Tabela 5.10: modelo - Totais de recursos analisados

X	Alvo	Qtde de Datasets	Qtde de Recursos
X	XXXX	XXX	XXX

Tabela 5.11: modelo - Datasets em discordância com LGPD

Alvo	Datasets	Qtde de Recursos	Total de Linhas
XXX	{PORTAL_URL}/dataset/XXXX	XXXX	XXXX

Dos 26 estados brasileiros, 13 possuem exposição com CKAN, sendo eles: Rondônia, Acre, Ceará, Pernambuco, Alagoas, Bahia, Minas Gerais, Espírito Santo, São Paulo, Santa Catarina, Rio Grande do Sul, Mato Grosso do Sul e Goiás. Um estado, Maranhão,

expõe seus dados com DKAN (derivado do CKAN). O Distrito Federal também expõe seus dados com CKAN. Ou seja, pouco mais da metade dos estados brasileiros, 53%, expoe seus dados de modo padronizado com CKAN ou derivados, conforme a Figura 5.21.

A análise foi mapeada então para 15 alvos: 14 estados mais Distrito Federal. No entanto, no momento do experimento apenas 12 alvos obtiveram êxito no acesso. Logo, 2 alvos estiveram indisponíveis (veja as evidências na Figura 5.22): o portal de dados abertos de Fortaleza, inacessível via chamada HTTP e manualmente, foi verificado estar com quantidade de conjunto de dados zerados; e o de Minas Gerais com erro do lado do servidor. O estado do Maranhão também não foi analisado, pois o protótipo do módulo M2 é compátivel no momento apenas com CKAN.

Tabela 5.12: rotas dos alvos analisados e aeus status de acesso

Alvo/Sigla	URL do Portal de Dados	Status de Retorno
Acre (AC)	https://dados.ac.gov.br	200 Sucesso
Rondônia (RO)	https://dados.ro.gov.br	200 Sucesso
Ceará (CE)	https://dados.fortaleza.ce.gov.br	JSONDecodeError
Alagoas (AL)	https://dados.al.gov.br/catalogo	200 Sucesso
Bahia (BA)	https://dados.ba.gov.br	200 Sucesso
Pernambuco (PE)	https://dados.pe.gov.br	200 Sucesso
Mato Grosso do Sul (MS)	http://www.dados.ms.gov.br	200 Sucesso
Goiás (GO)	https://dadosabertos.go.gov.br	200 Sucesso
Distrito Federal(DF)	http://www.dados.df.gov.br	200 Sucesso
Minas Gerais (MG)	https://dados.mg.gov.br/	502 bad gateway
Espirito Santo (ES)	https://dados.es.gov.br	200 Sucesso
São Paulo (SP)	http://dados.prefeitura.sp.gov.br	200 Sucesso
Santa Catarina (SC)	https://dados.sc.gov.br	200 Sucesso
Rio Grande do Sul (RS)	https://dados.rs.gov.br	200 Sucesso

Dos portais disponíveis, o experimento avaliou 11.154 recursos, contidos em 812 datasets, gerenciados por 217 organizações e distribuidos em 144 temas, conforme a Tabela 5.13. A maior quantidade de datasets foi observada no Rio Grande do Sul, em São Paulo e no Distrito Federal, o que indica maior variedade de dados nesses alvos. No entanto, a maior quantidade de recursos (porção de dados consumível) está nos estados de Pernambuco, Rio Grande do Sul e São Paulo, conforme Figura 5.23. O estado de Pernambuco (PE) foi removido da visualização pois a quantidade de recursos difere drasticamente dos demais — possui 5293 recursos — e a visualização ficaria comprometida, todavia, os dados foram analisados normalmente.

5.5.1 Análise por regiões do Brasil

Durante a análise, nenhum dataset com atributo CPF foi encontrado nos estados da região *Norte* disponíveis para análise (Acre e Rondônia). Logo, nenhum recurso foi escaneado, e consequentemente nenhuma visualização de dados desta região será apresentada.

Dos estados da região *Nordeste* disponíveis para análise — Pernambuco, Alagoas e Bahia —, 2 possuíam atributos de CPF em seus datasets, sendo 7 datasets do estado da Bahia, e 8 do estado de Pernambuco. Destes, o modelo classificou 4 datasets da Bahia, e ao menos 6 datasets de Pernambuco com possíveis dados de CPF expostos (veja Tabela 5.14):

Tabela 5.13: Totais de recursos analisados

Região	Alvo	Qtde de Datasets	Qtde de Recursos
Norte	Acre	20	20
Norte	Rondonia	5	9
Nordeste	Alagoas	66	184
Nordeste	Bahia	12	91
Nordeste	Pernambuco	27	5293
Centro-Oeste	Mato Grosso do Sul	28	776
Centro-Oeste	Goiás	25	238
Centro-Oeste	Distrito Federal	113	919
Sudeste	Espirito Santo	99	1024
Sudeste	São Paulo	124	1181
Sul	Santa Catarina	30	82
Sul	Rio Grande do Sul	263	1337
		812	11154

Tabela 5.14: datasets em discordância com a LGPD - Região Nordeste

Alvo	Datasets	Qtde de Recursos	Total de Linhas
BA	/dataset/diarias	7	9616
BA	/dataset/covid-19	2	2242
BA	/dataset/obras	6	1926
BA	/dataset/contratos	4	1284
PE	/dataset/bb965e9a-c108-496a-a0f4-f27ef803f531	43	5008
PE	/dataset/b94815fd-0327-4849-b3d2-6b9e88a8abcc	15	49
PE	/dataset/0e1282e5-0189-487f-90a6-dddec519b5b0	2	346
PE	/dataset/0fad561c-e79b-40c4-babf-b0b8189273ec	3	6230
PE	/dataset/06b932d9-139b-45f5-94b3-2a3abe87ec73	1	229
PE	/dataset/26b5138b-a6a2-43ee-8c89-3f4a1f0d7bca	20	4391
		103	29587

Na Figura 5.24, está representada a região Nordeste, demarcada, em preto; os estados que possuem CPFs expostos em vermelho, o estado Alagoas em laranja, pois, apesar de ter sido feita a análise, não existiam CPFs nos datatsets, então não foi possível verificar como as organizações daqueles estados tratam dados pessoais no momento de sua exposição.

Na *Centro-Oeste* do Brasil, entre os estados de Mato Grosso do Sul, Goiás e o Distrito Federal, disponíveis para a análise, Goiás possuía 6 datasets com atributos de CPF. Destes, o modelo classificou ao menos 5 datasets com dados de CPF expostos, veja Tabela 5.15:

Na Figura 5.25, está representada a região *Centro-Oeste*, demarcada em preto; os estados que possuem CPFs expostos, em vermelho; e os que possuem CPF verificado como protegido, em verde:

Na região *Sudeste* do Brasil, ambos os estados disponíveis para a análise — São Paulo e Espírito Santo — possuíam atributos de CPF em seus datasets, sendo 5 datastes do estado de Espírito Santo com CPF, e 2 do estado de São Paulo. Destes, o modelo classificou ao menos 1 dos 2 datasets do Espírito Santo com CPF expostos, e também, 2 do estado de São Paulo, veja Tabela 5.16:

Na Figura 5.26, está representada a região *Sudeste*, demarcada em preto, e os estados

Tabela 5.15: Datasets em discordância com LGPD - Regiao Centro-Oeste

Alvo	Datasets	Qtde de Recursos	Total de Linhas
GO	/dataset/fornecedores	9	2806
GO	/dataset/empenhos	11	6834
GO	/dataset/liquidacoes	11	4357
GO	/dataset/beneficiarios-dos-pagamentos	11	6460
GO	/dataset/contratos	7	25
GO	/dataset/pagamentos	10	5342
		59	25824

Tabela 5.16: datasets em discordância com a LGPD - Região Sudeste

Alvo	Datasets	Qtde de Recursos	Total de Linhas
SP	/dataset/997502fd-f084-47dd-97a8-910f1ea8bef8	2	1028
SP	/dataset/bd6b595b-b1ab-465b-821f-6bbdb099991b	21	5624
ES	/dataset/83ff8fa2-8d2b-4100-aa7d-92cb9d8f0d3b	1	266
ES	/dataset/99e16b13-0e6f-4504-8544-00de842ab1fd	15	11340
ES	/dataset/c8ae0a93-096a-4f74-add8-441ff620c68b	1	5
ES	/dataset/cfe5c4c8-6214-48a0-a873-a4b1be978155	18	270
ES	/dataset/ea970c7f-c524-45b0-a346-74ef5b1af218	14	5667
		72	24200

que possuem CPFs expostos, em vermelho:

Na região *Sul* do Brasil, dos estados disponíveis para análise — Santa Catarina e Rio Grande do Sul —, Santa Catarina possuía atributos de CPF em seus datasets, sendo 2 datasets retornados. Destes, o modelo classificou 1 datasets com dados de CPF expostos (veja Tabela 5.17).

Tabela 5.17: datasets em discordância com a LGPD - Região Sul

Alvo	Datasets	Qtde de Recursos	Total de Linhas
Santa Catarina	{PORTAL_URL}/dataset/transferencias	2	39
		2	39

Na Figura 5.27, está representada a região *Sul* demarcada em preto, os estados que possuem CPFs expostos em vermelho, o estado do Rio Grande do Sul em laranja, pois, apesar de ter sido feita a análise, não exitiam CPFs nos datatsets, então não foi possível verificar se os dados estariam em conformidade com LGPD ou não.

5.5.2 Percepções da Análise

De acordo com o verificado no módulo M2, a Figura 5.28 demonstra, em verde, os estados que estão em conformidade com a LGDP, relacionado ao dado pessoal CPF; em laranja, os estados onde não foi encontrado o atributo de CPF, e, assim não foi possível verificar como as organizações cadastram dados pessoais; e, em vermelho, os estados que o modelo classificou como tendo dados de CPF expostos nos conjuntos de dados. Os demais estados representados, ou não possuíam portais de dados abertos com CKAN, ou os portais existentes não estavam disponíveis no momento do experimento, como menci-

onado anteriormente, na Figura 5.22.

Resultado da busca

organization	license_title	is_open	package_id	resource_route	dataset_url
ibama	Outra (Aberta)	True	4b292c2e-56bb-4639-9ea2-a9b02fea9558	https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_embargo/decisao/decisao.csv	http://dadosabertos.ibama.gov.br/dataset/fiscalizacao-termo-de-embargo
ibama	Outra (Domínio Público)	True	89568872-a865-4e19-8a7c-6f11965a3195	https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_apreensao/termo_apreensao.csv	http://dadosabertos.ibama.gov.br/dataset/fiscalizacao-termo-de-apreensao
ibama	Outra (Aberta)	True	049740d8-a3c5-425f-9161-431a42a87109	https://dadosabertos.ibama.gov.br/dados/AATIPP/autorizacao_empresa/DF/2016.csv	http://dadosabertos.ibama.gov.br/dataset/sinaflor-poa-outros-biomas
ibama	Outra (Aberta)	True	09850847-5ba2-435a-942c-a5dcd7e970c9	https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_apreensao/termo_apreensao.csv	http://dadosabertos.ibama.gov.br/dataset/fiscalizacao-auto-de-infracao
ibama	Outra (Aberta)	True	47c97994-cd26-4d16-a2d8-4fbc27b1d9a0	https://dadosabertos.ibama.gov.br/dados/SIFISC/termo_suspensao/suspensao.csv	http://dadosabertos.ibama.gov.br/dataset/fiscalizacao-termo-de-suspensao
ibama	Other (Open)	True	2dc12ec3-de3d-4bf4-964b-374d9bacad6f	https://dadosabertos.ibama.gov.br/dados/AATIPP/autorizacao_empresa/DF/2016.csv	http://dadosabertos.ibama.gov.br/dataset/autorizacao-ambiental-para-o-transporte-interestadual-de-produtos-perigosos-aatipp

Figura 5.14: resultado da consulta por dados pessoais em datasets do Ibama, M2 - CompOD. Fonte: autor.

Figura 5.15: resultado da classificação do modelo, em recursos do Ibama. Fonte: autor.

Figura 5.16: dados pessoais protegidos em recursos, M2 - CompOD. Fonte: autor.

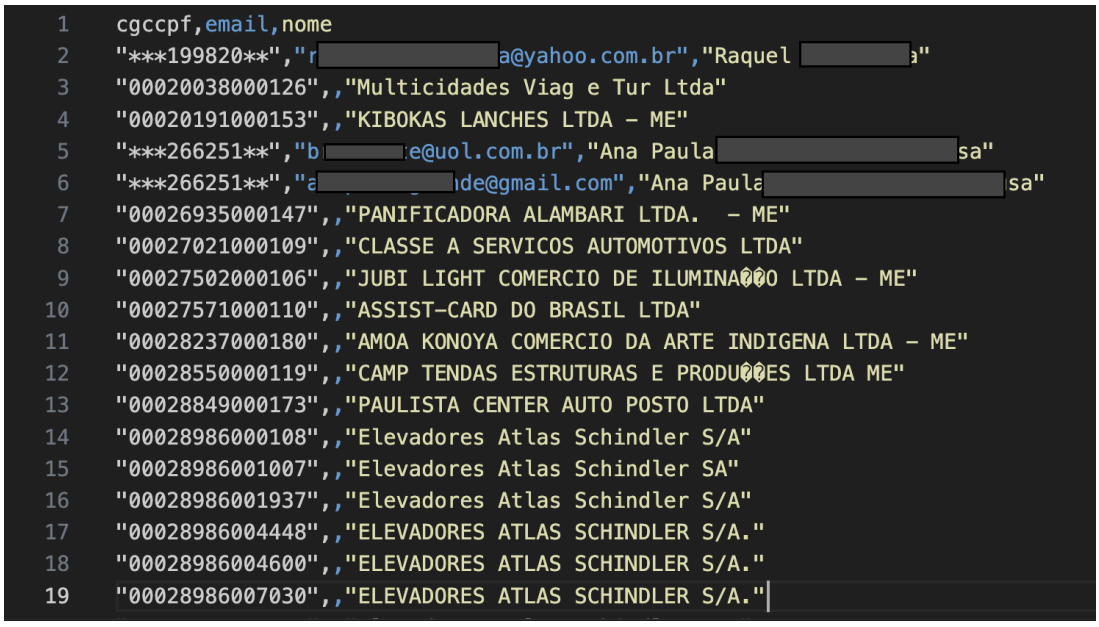


Figura 5.17: combinação de múltiplos dados pessoais no recurso, M2 - CompOD. Fonte: autor.

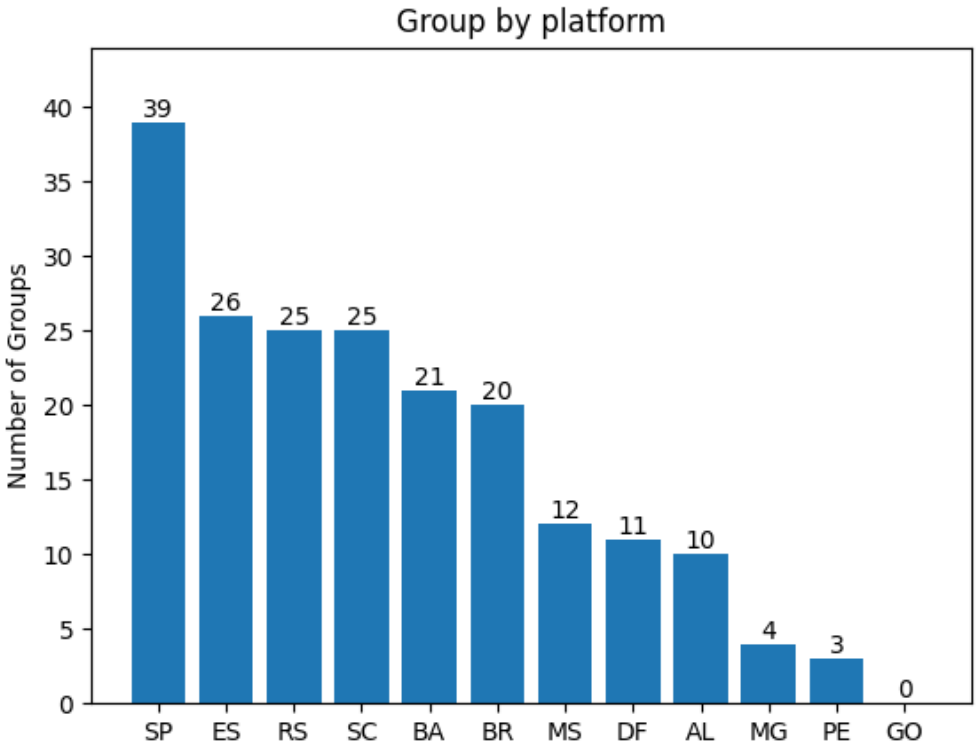


Figura 5.18: quantidade de grupos por estado e distrito federal. Fonte: autor.

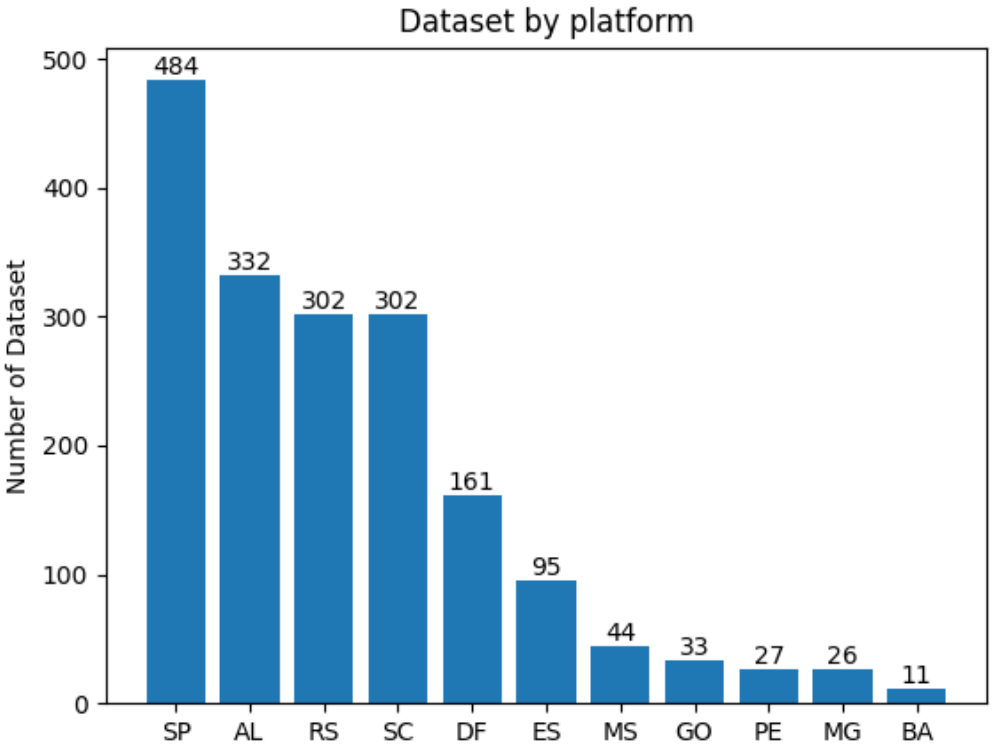


Figura 5.19: quantidade de conjuntos de dados por estado e Distrito Federal. Fonte: autor.

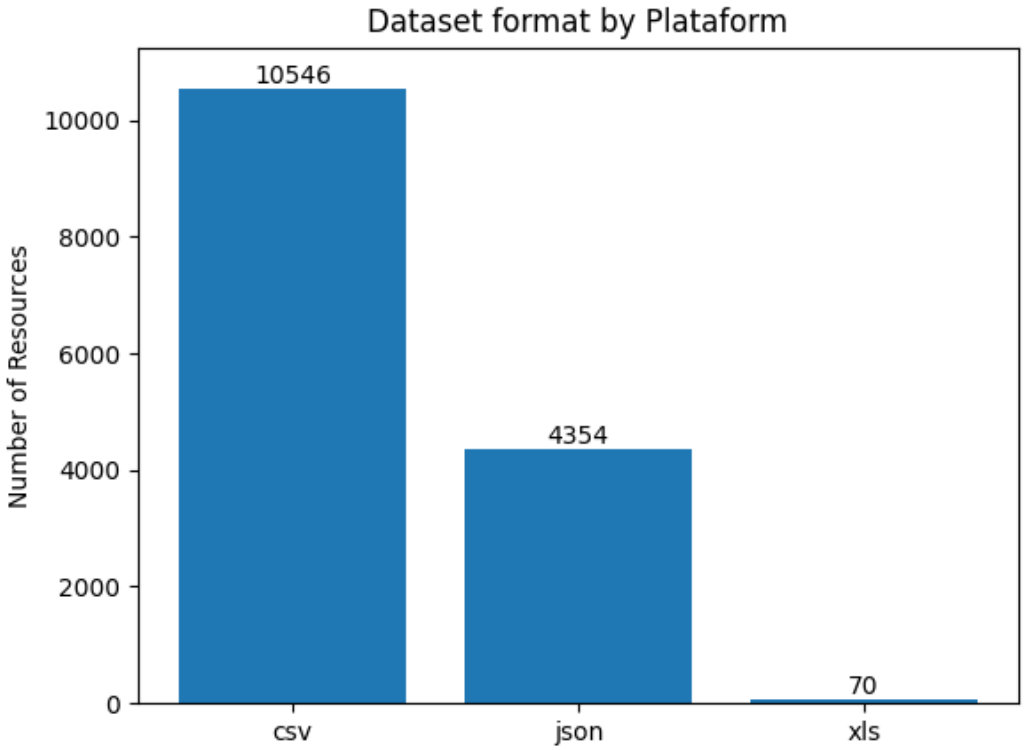


Figura 5.20: formatos de dados disponíveis mais comumente usados. Fonte: autor.

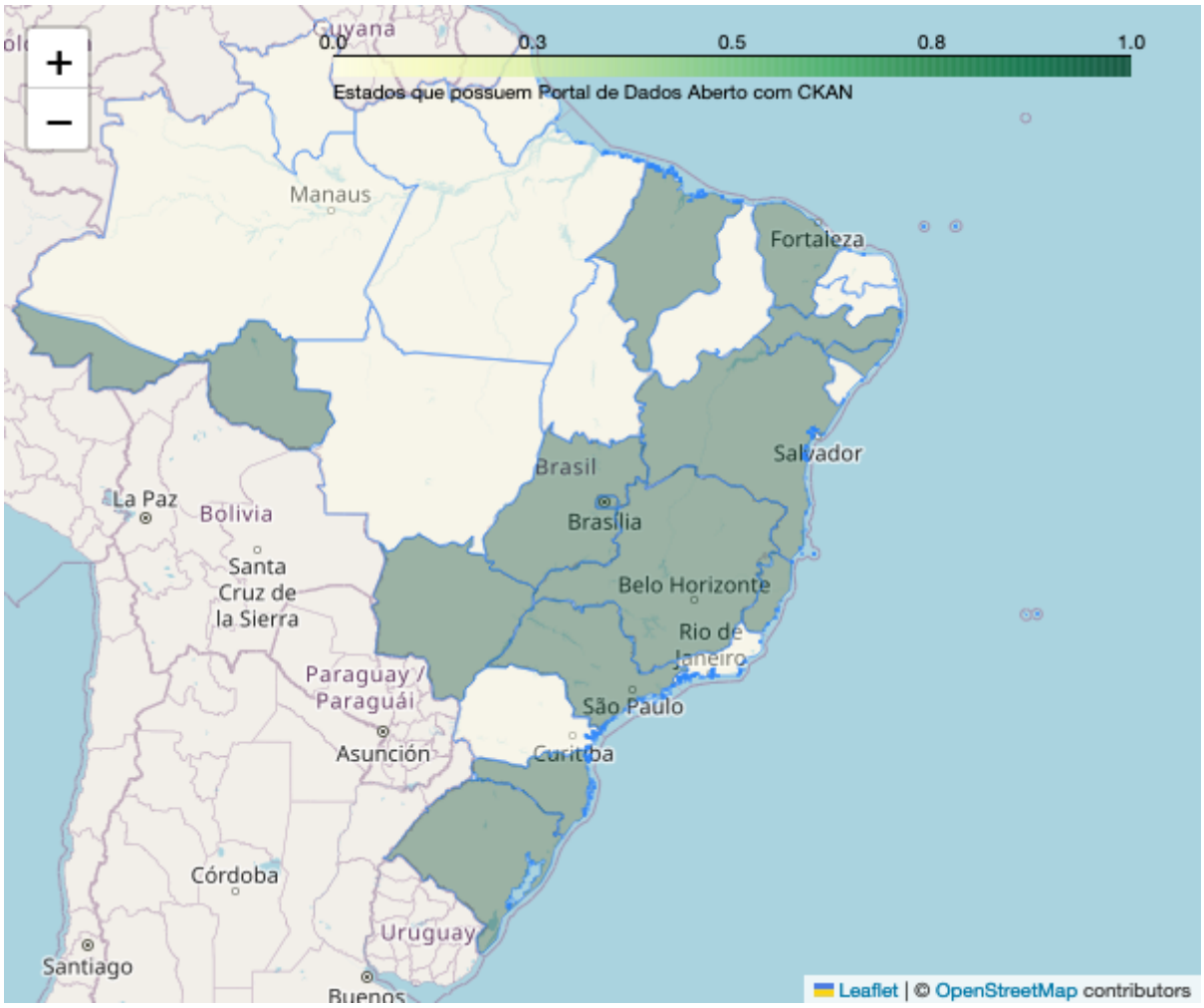


Figura 5.21: estados que expoem seus dados abertos com CKAN. Fonte: autor.

Figura 5.22: indisponibilidade de portais de Dddos abertos. Fonte: autor.

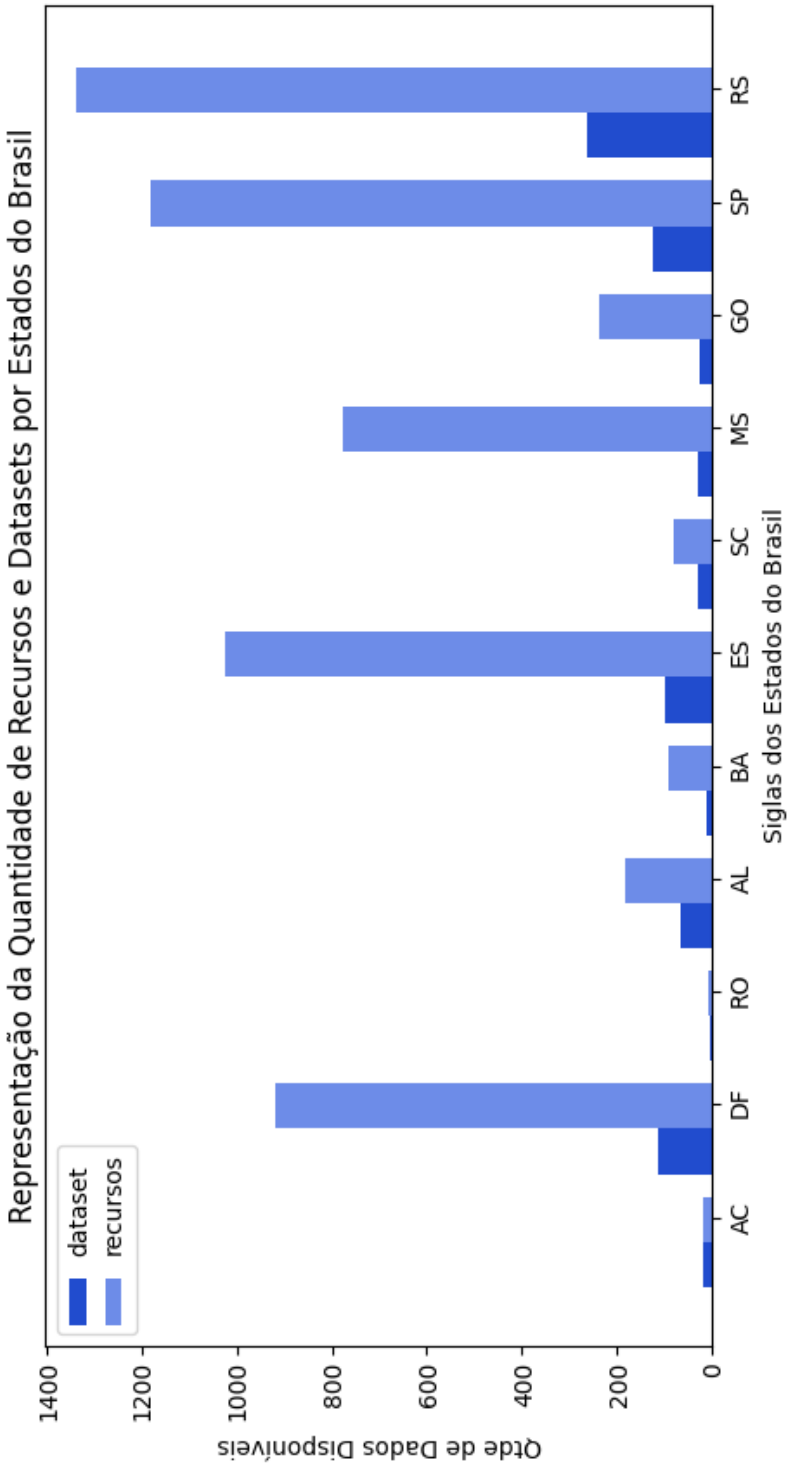


Figura 5.23: datasets e recursos por estado. Fonte: autor.

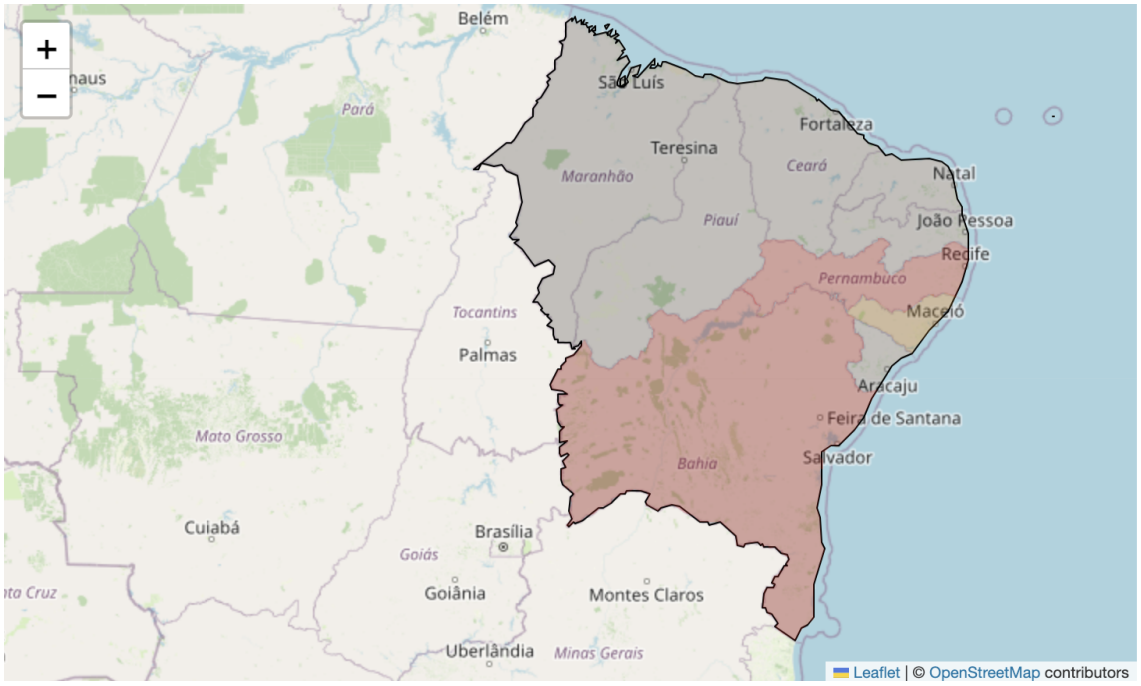


Figura 5.24: estados que possuem dados pessoais expostos, Região Nordeste do Brasil. Fonte: autor.

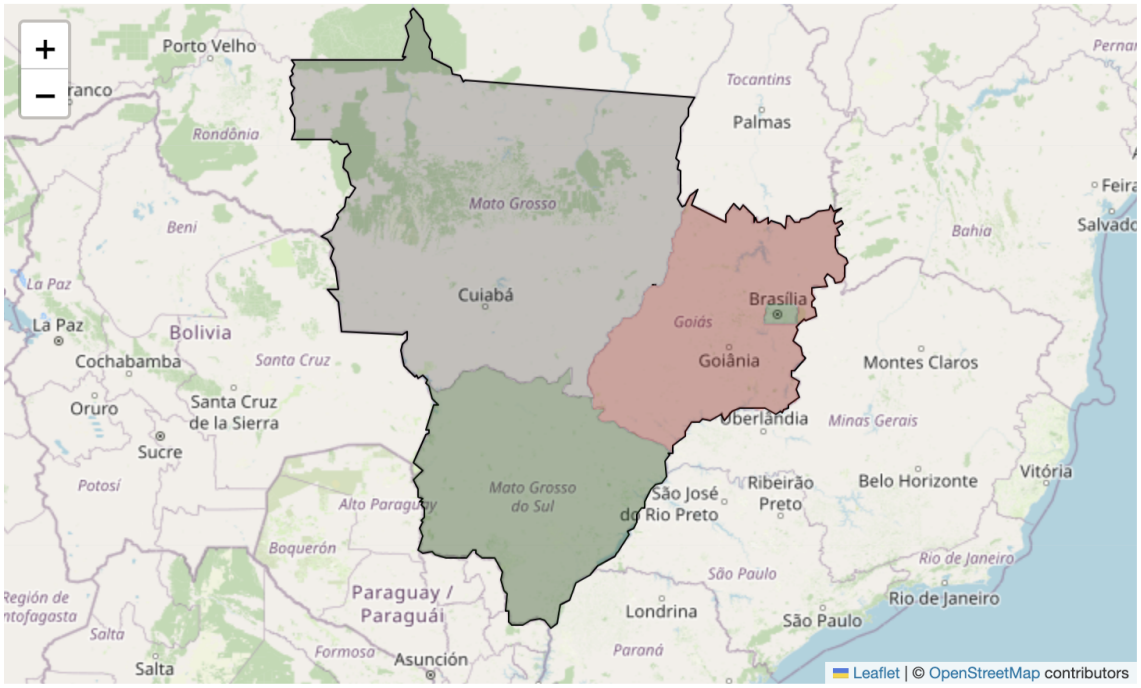


Figura 5.25: estados que possuem dados pessoais expostos, Região Centro-Oeste do Brasil. Fonte: autor.

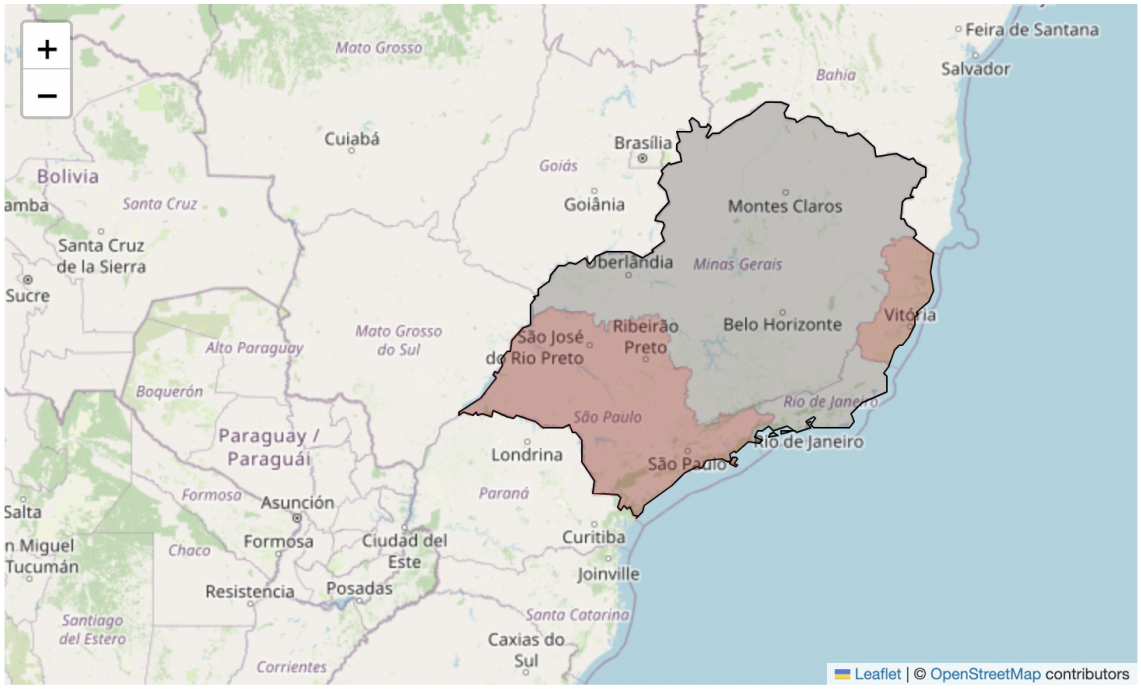


Figura 5.26: estados que possuem dados pessoais expostos, Região Sudeste do Brasil. Fonte: autor.

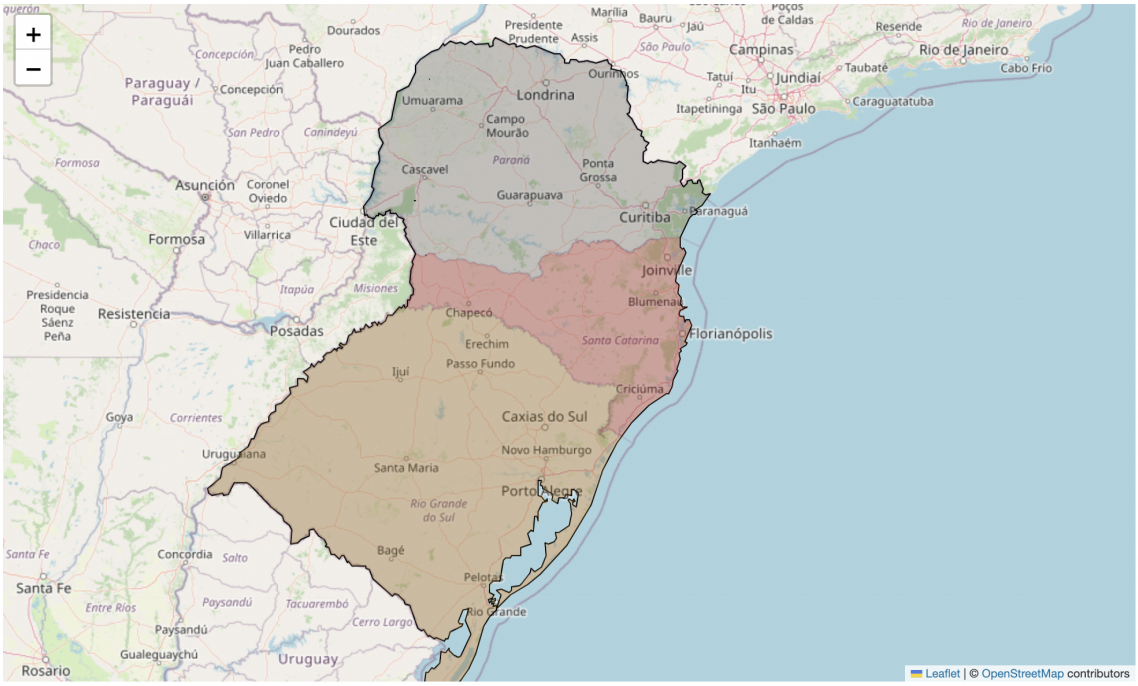


Figura 5.27: estados que possuem dados pessoais expostos, Região Sul do Brasil. Fonte: autor.

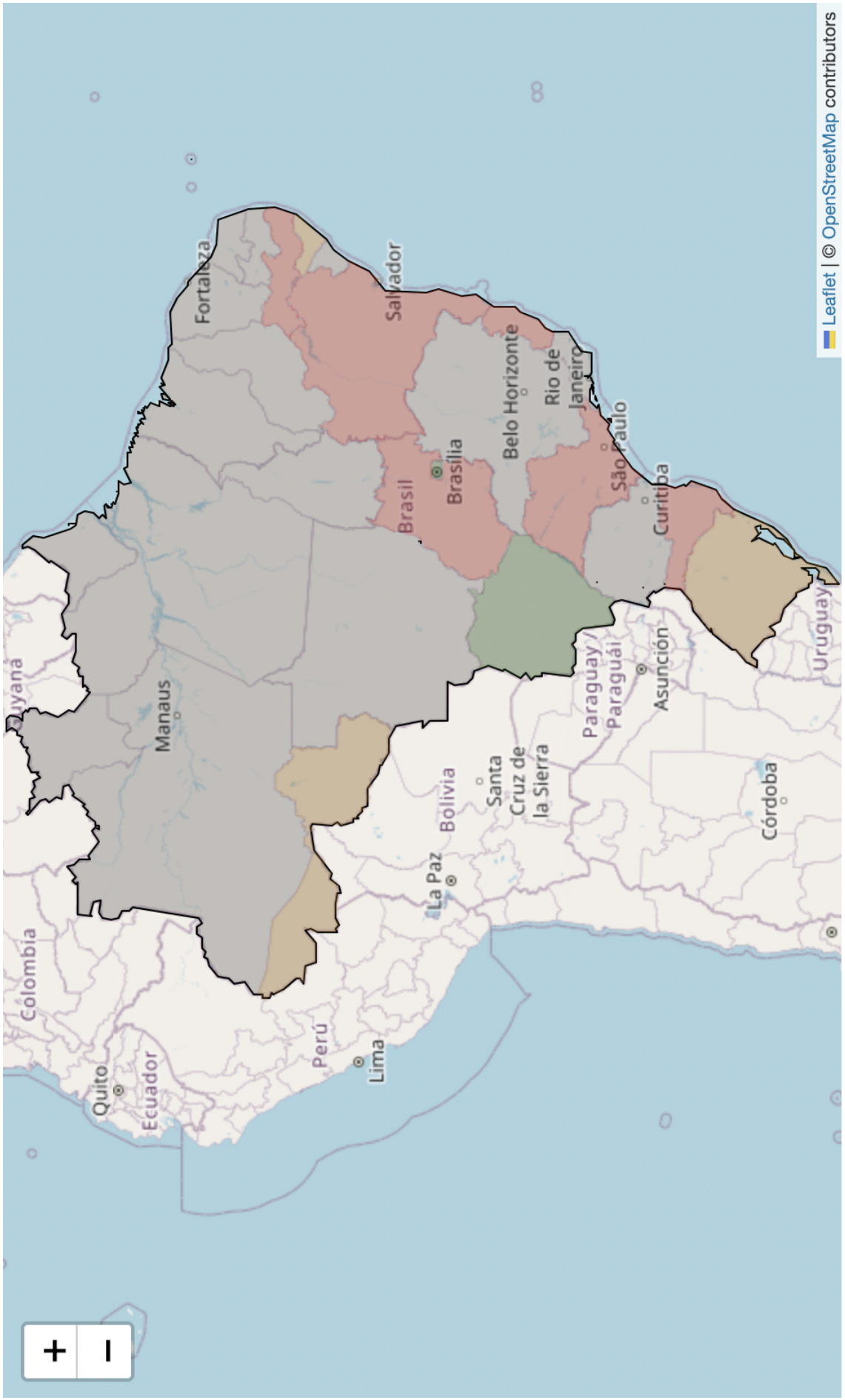


Figura 5.28: análise de conformidade a LGDP nos estados do Brasil. Fonte: autor.

6 CONCLUSÕES

O módulo M1 buscou entender o panorama atual de dados aberto no Brasil, respondendo à pergunta de pesquisa *Qual é o estado dos dados abertos no Brasil?*. Para responder tal pergunta, uma segunda questão foi levantada: *Quais plataformas ou framework de exposição de dados abertos existem?*. A partir das respostas à primeira pergunta, foi percebido que menos da metade dos estados brasileiros (42%), ou 11 dos 26, possuem portais de dados abertos utilizando uma padronização amplamente aceita, como CKAN. Como mencionado, isso dificulta estudos sistemáticos de dados abertos, pois é inviável desenvolver um protótipo individual para cada objeto de estudo.

Além da análise individual de cada estado que expõe os dados com CKAN, foram analisados, também, os órgãos governamentais que representam as dimensões conforme a Tabela 5.5 da OKFN. Assim, foi percebido que as oportunidades de melhoria estão em concordância com outros estudos, como o de (BENO et al., 2017), sendo elas: (1) dados desatualizados; (2) ausência de metadados para suportar a utilização dos dados; (3) especificação de licença de uso imprecisa, dificultando o amparo legal de utilização dos dados; (4) falta de padronização na exposição dos dados abertos em portais, ou seja, heterogeneidade entre portais de dados abertos. Os pontos onde houve boas métricas foram: (1) dados disponíveis em massa, tendo suas rotas HTTP retornando sucesso, pois, das 7 dimensões analisadas, 5 estavam com porcentagem de disponibilidade acima de 80%; (2) dados legíveis por máquina, com a maioria das dimensões analisadas tendo os dados disponíveis no formato CSV (legível por máquina).

Já o módulo M2 buscou entender se os dados expostos em portais brasileiros de dados abertos estavam em conformidade com a lei geral de proteção de dados (LGPD), respondendo às perguntas de pesquisa *Quais plataformas ou framework de exposição de dados abertos existentes possuem meios de verificação de conformidade com leis de proteção de dados?* e *Qual metodologia de exposição de dados abertos existentes possui meios de verificação de conformidade com a LGPD?*. Na literatura foi verificado que existem frameworks de exposição de dados abertos, inclusive propondo melhorias nas lacunas do CKAN — framework utilizado neste trabalho. No entanto, após uma busca do código fonte das soluções a fim de reproduzi-las, nenhum repositório de códigos foi encontrado, sendo este um dos motivos para utilização do CKAN. Ainda na literatura, foi feita uma busca por frameworks de verificação de conformidade com leis de proteção de dados pelo mundo, como o GDPR, em dados abertos. Porém, nenhum framework de conformidade com leis de proteção de dados em *Dados Abertos Governamentais* foi encontrado. Deste modo, os módulos M1 e M2 agem em complemento, e buscam atender a necessidade percebida durante a leitura da literatura.

Os experimentos do módulo M2 mostram que em alguns estados do Brasil, como Salvador, São Paulo e Santa Catarina possuem organizações que cadastram os datasets se

preocupando com a proteção de dados, no entanto, nem todos os datasets em seus portais estão em conformidade com a lei, uma vez que, durante o experimento, foram encontrados dados pessoais expostos, e sem o devido tratamento durante o experimento. Alguns estados demonstraram estar em conformidade, pois os módulos não verificaram CPFs expostos nas massas de dados analisadas, no entanto, grande parte dos estados do Brasil, como toda região Norte, parte da região Nordeste, estados importantes para a econonomia do país, como Rio de Janeiro, Paraná e Minas Gerais, não possuíam, na data dos experimentos, portais de dados padronizados com exposição através de frameworks amplamente utilizados, como CKAN. Deste modo, foi perceptível a necessidade de fomento e incentivo sobre a importância de padronizar a abertura dos dados e, ao fazê-lo atentar-se para a proteção de dados pessoais.

As principais contribuições desta pesquisa foram:

- estudo da qualidade dos dados abertos no Brasil, resultante do módulo M1 deste trabalho;
- estudo de conformidade dos dados abertos em relação à leis de proteção de dados no Brasil, no caso a LGPD. Resultante do módulo M2 deste trabalho;
- desenvolvimento de framework reproduzível dos módulos M1 e M2, sendo o M2 o diferencial com relação as soluções existentes, pois verifica a conformidade com leis de proteção de dados;
- publicação de artigo na revista *Journal of Internet Services and Applications* (JISA), de qualis A2, disponivel em *https://doi.org/10.5753/jisa.2024.3980*.

6.1 Discussão

A massa de dados utilizada nos experimentos foi adquirida utilizando a API do CKAN. No entanto, alguns portais de dados estavam indisponíveis no momento da reprodução dos experimentos, ou, como no caso do portal brasileiro de dados abertos[1], que foi migrado e está permanentemente indisponível através do CKAN. À época dos experimentos do módulo M1, o portal brasileiro de dados abertos estava expondo seus dados através do CKAN, no entanto, por meados de agosto de 2023, o portal começou a ser migrado, fazendo com que a URL legada[2] (atualmente indisponível) precisasse ser utilizada na finalização nos experimentos do módulo M2. É possível comprovar isso através da Figura 6.1, que contém a resposta da *Controladoria Geral da União*[3] (CGU), após entrarmos em contato indagando sobre a implementação do CKAN no novo portal.

Cabe ressaltar que o exposto no parágrafo anterior corrobora uma das dificuldades de aquisição/consumo de dados abertos citadas por (BENO et al., 2017), e também percebida neste trabalho: a heterogeneidade da exposição de dados em portais de dados abertos. Um outro ponto negativo, neste cenário é, que desde a migração do portal brasileiro de dados abertos, só é possivel acessar os dados — através de APIs — fazendo cadastro e autenticação no portal, o que entra em conflito com o *livre acesso aos dados*, amplamente difundido pela *Open Knowledge Foundation*[4].

[1] https://dados.gov.br/home
[2] https://legado.dados.gov.br/
[3] https://www.gov.br/cgu/pt-br
[4] https://okfn.org/en/library/what-is-open/

C

CGU Suporte Dados Abertos <suporte.dadosabertos@cgu.gov.br>
Para: Você
Cc: Dados Abertos - CGU; Cláudio Geyer; Julio Anjos

Seg, 11/12/2023 10:49

Prezada Shirley, bom dia.

Devido a nova arquitetura do portal brasileiro de dados abertos não iremos disponibilizar a interface CKAN para consumo público, no entanto nós disponibilizamos uma API que permite consumir todos os dados públicos do portal.

Para instruções de como acessar a API basta visitar o link a seguir: https://dados.gov.br/dados/conteudo/como-acessar-a-api-do-portal-de-dados-abertos-com-o-perfil-de-consumidor

Qualquer dúvida entre em contato.

Atenciosamente

Equipe de Suporte do Portal de Dados Abertos/CGU
Núcleo de Dados Abertos
Coordenação-Geral de Transparência Ativa e Dados Abertos - CGDAT
Diretoria de Governo Aberto e Transparência - DGAT
Secretaria de Integridade Pública - SIP

CGU
Controladoria-Geral da União
www.cgu.gov.br

Figura 6.1: CGU sobre CKAN no novo portal brasileiro de dados abertos. Fonte: autor.

A qualidade dos dados interferiu diretamente nos resultados alcançados, como quando o atributo que o usuário deseja buscar nos recursos está mesclado em outras colunas com nome diferente. Isso acontece, por exemplo, com o termo LGPD. Na Figura 6.2, do conjunto de dados [5] disponível no portal de dados abertos do Distrito Federal[6], é possível verificar que na coluna *CPF/CNPJ* o algorítmo seria levado a predizer que o CPF está protegido, no entanto, neste mesmo recurso, quando a coluna *PERSONALIDADE* tem o valor *PESSOA JURÍDICA*, é possível perceber que a coluna *NOME TOMADOR* apresenta o que se assemelha a um CPF na linha correspondente.

6.2 Trabalhos futuros

O escopo do framework, atualmente, é a conformidade em dados pessoais, no entanto a LGPD prevê atuação também em dados pessoais sensíveis, que de acordo com a lei, são, dados sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dados referentes à saúde ou à vida sexual, e dados genéticos ou biométricos, quando vinculados a uma pessoa natural. Por exemplo, no portal de dados abertos do estado de São Paulo[7] existem recursos que expõem a combinação de dados pessoais, como nome completo (dados pessoais), com raça declarada e gênero (dados sensíveis), conforme a Figura 6.3, adapatada do *portal de Dados Abertos de São Paulo*.

Anteriormente, no estudo de (KOSINSKI; STILLWELL; GRAEPEL, 2013) foi visto que, através da análise de *likes* (mecanismo de associação positiva em conteúdos da rede social *Facebook*), foi possível alcançar uma precisão quase perfeita — nas palavras do autor — na predição de variáveis dicotômicas como etnia e gênero. Logo, existe, ainda, oportunidade de pesquisa e evolução sobre como dados pessoais sensíveis podem estar

[5] http://www.dados.df.gov.br/dataset/dados-do-prospera-df/resource/94a5c43c-1a1b-4dda-9579-e2cf288c4921

[6] http://www.dados.df.gov.br/

[7] http://dados.prefeitura.sp.gov.br/pt_PT/

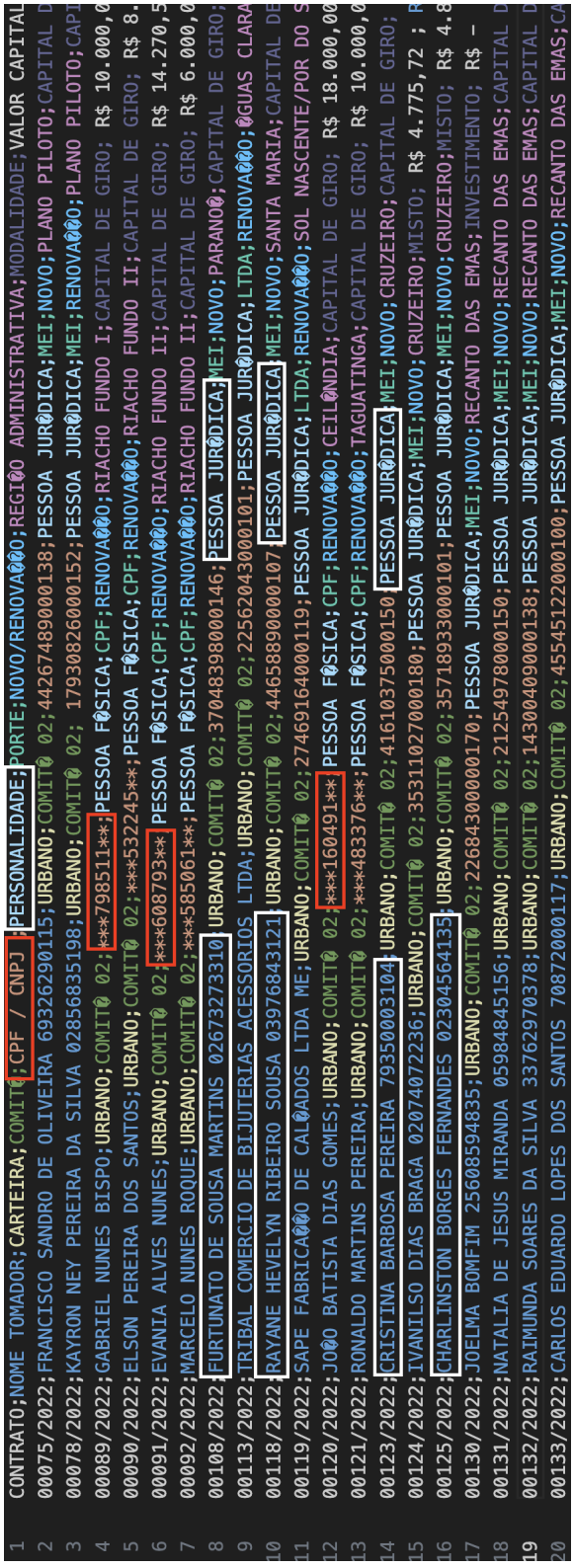
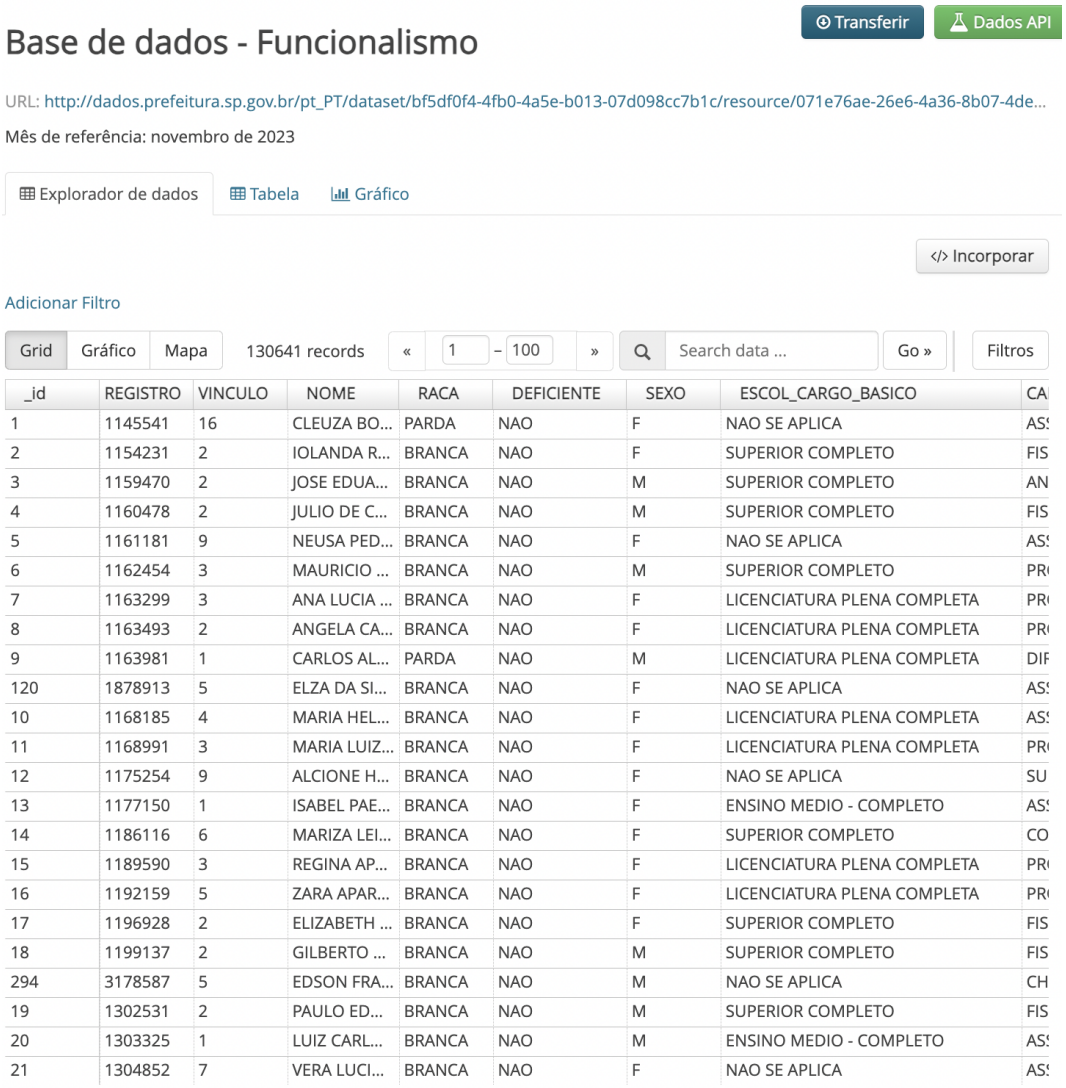


Figura 6.2: dados pessoais mesclados em recurso. Fonte: Portal de Dados Abertos do Distrito Federal.

sendo coletados, ou nesse caso gerados/sugeridos/inferidos, e quais seriam as implicações do não controle e da não fiscalização, ou conformidade, da proteção dos dados sensíveis aos usuários.



Base de dados - Funcionalismo

Transferir | Dados API

URL: http://dados.prefeitura.sp.gov.br/pt_PT/dataset/bf5df0f4-4fb0-4a5e-b013-07d098cc7b1c/resource/071e76ae-26e6-4a36-8b07-4de...

Mês de referência: novembro de 2023

Explorador de dados | Tabela | Gráfico

Incorporar

Adicionar Filtro

Grid | Gráfico | Mapa | 130641 records | « | 1 – 100 | » | Search data ... | Go » | Filtros

_id	REGISTRO	VINCULO	NOME	RACA	DEFICIENTE	SEXO	ESCOL_CARGO_BASICO	CA
1	1145541	16	CLEUZA BO...	PARDA	NAO	F	NAO SE APLICA	AS
2	1154231	2	IOLANDA R...	BRANCA	NAO	F	SUPERIOR COMPLETO	FIS
3	1159470	2	JOSE EDUA...	BRANCA	NAO	M	SUPERIOR COMPLETO	AN
4	1160478	2	JULIO DE C...	BRANCA	NAO	M	SUPERIOR COMPLETO	FIS
5	1161181	9	NEUSA PED...	BRANCA	NAO	F	NAO SE APLICA	AS
6	1162454	3	MAURICIO ...	BRANCA	NAO	M	SUPERIOR COMPLETO	PR
7	1163299	3	ANA LUCIA ...	BRANCA	NAO	F	LICENCIATURA PLENA COMPLETA	PR
8	1163493	2	ANGELA CA...	BRANCA	NAO	F	LICENCIATURA PLENA COMPLETA	PR
9	1163981	1	CARLOS AL...	PARDA	NAO	M	LICENCIATURA PLENA COMPLETA	DIF
120	1878913	5	ELZA DA SI...	BRANCA	NAO	F	NAO SE APLICA	AS
10	1168185	4	MARIA HEL...	BRANCA	NAO	F	LICENCIATURA PLENA COMPLETA	AS
11	1168991	3	MARIA LUIZ...	BRANCA	NAO	F	LICENCIATURA PLENA COMPLETA	PR
12	1175254	9	ALCIONE H...	BRANCA	NAO	F	NAO SE APLICA	SU
13	1177150	1	ISABEL PAE...	BRANCA	NAO	F	ENSINO MEDIO - COMPLETO	AS
14	1186116	6	MARIZA LEI...	BRANCA	NAO	F	SUPERIOR COMPLETO	CO
15	1189590	3	REGINA AP...	BRANCA	NAO	F	LICENCIATURA PLENA COMPLETA	PR
16	1192159	5	ZARA APAR...	BRANCA	NAO	F	LICENCIATURA PLENA COMPLETA	PR
17	1196928	2	ELIZABETH ...	BRANCA	NAO	F	SUPERIOR COMPLETO	FIS
18	1199137	2	GILBERTO ...	BRANCA	NAO	M	SUPERIOR COMPLETO	FIS
294	3178587	5	EDSON FRA...	BRANCA	NAO	M	NAO SE APLICA	CH
19	1302531	2	PAULO ED...	BRANCA	NAO	M	SUPERIOR COMPLETO	FIS
20	1303325	1	LUIZ CARL...	BRANCA	NAO	M	ENSINO MEDIO - COMPLETO	AS
21	1304852	7	VERA LUCI...	BRANCA	NAO	F	NAO SE APLICA	AS

Figura 6.3: combinação de dados pessoais e sensíveis no portal de dados abertos de São Paulo. Fonte: Portal de Dados Abertos de São Paulo

Existe, também, oportunidade de evolução no framework para evitar, preventivamente, que dados pessoais sejam expostos em inconformidade com leis de proteção de dados (não apenas em dados abertos), uma vez que o trabalho atual buscou entender o quanto os dados abertos já publicados estariam em conformidade. Logo, existe interessante oportunidade de atuar na prevenção e fiscalização da exposição de dados pessoais, a exemplo de (FERREIRA et al., 2023), no escopo da saúde, (ALAMRI; JAVED; MARGARIA, 2021), no escopo de aplicações Web, e o trabalho de (GUAMáN; DEL ALAMO; CAIZA, 2021), que avalia sistematicamente a conformidade de aplicativos móveis, todos no escopo de proteção de dados pessoais em conformidade com o Regulamento Geral sobre a Proteção de Dados da Europa (GDPR).

REFERÊNCIAS

ABIDI, S. et al. A Web Service Security Governance Approach Based on Dedicated Micro-services. **Procedia Computer Science**, [S.l.], v.159, p.372–386, 2019. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 23rd International Conference KES2019.

AHMAD, H.; AUJLA, G. S. GDPR compliance verification through a user-centric blockchain approach in multi-cloud environment. **Computers and Electrical Engineering**, [S.l.], v.109, p.108747, 2023.

AHN, S. et al. A Fuzzy Logic Based Machine Learning Tool for Supporting Big Data Business Analytics in Complex Artificial Intelligence Environments. In: IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE), 2019. **Anais...** [S.l.: s.n.], 2019. p.1–6.

ALAMRI, B.; JAVED, I. T.; MARGARIA, T. A GDPR-Compliant Framework for IoT-Based Personal Health Records Using Blockchain. In: IFIP INTERNATIONAL CONFERENCE ON NEW TECHNOLOGIES, MOBILITY AND SECURITY (NTMS), 2021. **Anais...** [S.l.: s.n.], 2021. p.1–5.

ALI HASSAN, M. I.; TWINOMURINZI, H. A Systematic Literature Review of Open Government Data Research: challenges, opportunities and gaps. In: OPEN INNOVATIONS CONFERENCE (OI), 2018. **Anais...** [S.l.: s.n.], 2018. p.299–304.

ALMUSALAMI, A. et al. AffordAD: a user friendly tool for estimating housing affordability in abu dhabi. In: ANNUAL UNDERGRADUATE RESEARCH CONFERENCE ON APPLIED COMPUTING (URC), 2022. **Anais...** [S.l.: s.n.], 2022. p.1–6.

Anaconda (Python distribution). [S.l.]: Wikimedia, 2023.

ANGELI, A. TRANSPARÊNCIA E ACESSO À INFORMAÇÃO: quem É o cidadÃo que demanda a abertura de informaÇÕes pÚblicas no brasil? **Revista Eletrônica de Ciência Política**, [S.l.], v.7, n.2, 2016.

ARBEX, A. M. G. **Como os dados abertos podem revolucionar as cidades**. 2020.

ARCHENAA, J.; ANITA, E. M. A Survey of Big Data Analytics in Healthcare and Government. **Procedia Computer Science**, [S.l.], v.50, p.408–413, 2015. Big Data, Cloud and Computing Challenges.

ASSEMBLY, U. G. The right to privacy in the digital age : resolution / adopted by the general assembly. **United Nations Digital Library**, [S.l.], 2016. https://digitallibrary.un.org/record/858023?ln=enrecord-files-collapse-header.

AYDIN, S.; AYDIN, M. N. Ontology-based data acquisition model development for agricultural open data platforms and implementation of OWL2MVC tool. **Computers and Electronics in Agriculture**, [S.l.], v.175, p.105589, 2020.

BACHTIAR, A.; SUHARDI; MUHAMAD, W. Literature Review of Open Government Data. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY SYSTEMS AND INNOVATION (ICITSI), 2020. **Anais...** [S.l.: s.n.], 2020. p.329–334.

BARI, A.; SAATCIOGLU, G. Emotion Artificial Intelligence Derived from Ensemble Learning. In: IEEE INTERNATIONAL CONFERENCE ON TRUST, SECURITY AND PRIVACY IN COMPUTING AND COMMUNICATIONS/ 12TH IEEE INTERNATIONAL CONFERENCE ON BIG DATA SCIENCE AND ENGINEERING (TRUSTCOM/BIGDATASE), 2018. **Anais...** [S.l.: s.n.], 2018. p.1763–1770.

BENO, M. et al. Perception of Key Barriers in Using and Publishing Open Data. **JeDEM - eJournal of eDemocracy and Open Government**, [S.l.], v.9, n.2, p.134–165, Dec. 2017.

BERNERS-LEE, T. **Linked Data, the Semantic Web.** 2006.

BERTOLINI, G. et al. **Governo Aberto**: transparência e dados abertos. [S.l.]: Escola Nacional de Administração Pública (Enap), 2022. http://repositorio.enap.gov.br/handle/1/772.

BIONI, B.; SILVA, P. d.; MARTINS, P. Intersecções e relações entre a Lei Geral de Proteção de Dados (LGPD) e a Lei de Acesso à Informação (LAI): análise contextual pela lente do direito de acesso. **Controladoria Geralmda União (CGU)**, [S.l.], 2022.

BIZER, C. et al. DBpedia - A crystallization point for the Web of Data. **Journal of Web Semantics**, [S.l.], v.7, n.3, p.154–165, 2009. The Web of Data.

BOUCHELOUCHE, K.; GHOMARI, A. R.; ZEMMOUCHI-GHOMARI, L. Enhanced analysis of Open Government Data: proposed metrics for improving data quality assessment. In: INTERNATIONAL SYMPOSIUM ON INFORMATICS AND ITS APPLICATIONS (ISIA), 2022. **Anais...** [S.l.: s.n.], 2022. p.1–6.

CABEZUELO, A. S. Using open data repositories and geolocation to create value-added services for tourism. In: INTERNATIONAL CONFERENCE INFORMATION VISUALISATION (IV), 2020. **Anais...** [S.l.: s.n.], 2020. p.126–131.

CAMENISCH, J.; FISCHER-HüBNER, S.; RANNENBERG, K. **Privacy and Identity Management for Life**. [S.l.]: Springer Berlin, Heidelberg, 2011. v.1.

CHANG, Y.-H.; JANG, H.-C. Traffic Flow Forecast for Traffic with Forecastable Sporadic Events. In: TWELFTH INTERNATIONAL CONFERENCE ON UBI-MEDIA COMPUTING (UBI-MEDIA), 2019. **Anais...** [S.l.: s.n.], 2019. p.145–150.

CHEN, L.; JAKUBOWICZ, J. Inferring bike trip patterns from bike sharing system open data. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2015. **Anais...** [S.l.: s.n.], 2015. p.2898–2900.

CKANGUIDE. **CKAN User guide**. 2023.

COMANDè, G.; SCHNEIDER, G. Can the GDPR make data flow for research easier? Yes it can, by differentiating! A careful reading of the GDPR shows how EU data protection law leaves open some significant flexibilities for data protection-sound research activities. **Computer Law Security Review**, [S.l.], v.41, p.105539, 2021.

CONSOLI, S. et al. Cultural gems linked open data: mapping culture and intangible heritage in european cities. **Data in Brief**, [S.l.], v.49, p.109375, 2023.

CRUZ, B. A. B. LEI DE ACESSO À INFORMAÇÃO COMO MECANISMO DE CONTROLE SOCIAL SOBRE POLÍTICAS PÚBLICAS E COMBATE À CORRUPÇÃO. **Conferência Internacional de Comissários de Acesso à Informação - ICIC 2021**, [S.l.], 2022.

DOMINGUE, J. et al. **The Future Internet**. [S.l.]: Springer, 2011. v.1.

DOROBăț, I. C.; POSEA, V. Open Data Indicator: an accumulative methodology for measuring the quality of open government data. In: INTERNATIONAL CONFERENCE ON ELECTRONICS, COMPUTERS AND ARTIFICIAL INTELLIGENCE (ECAI), 2021. **Anais...** [S.l.: s.n.], 2021. p.1–4.

DUAN, Y. et al. Everything as a Service (XaaS) on the Cloud: origins, current and future trends. In: IEEE 8TH INTERNATIONAL CONFERENCE ON CLOUD COMPUTING, 2015. **Anais...** [S.l.: s.n.], 2015. p.621–628.

DYBA, T.; DINGSOYR, T.; HANSSEN, G. K. Applying Systematic Reviews to Diverse Study Types: an experience report. In: FIRST INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT (ESEM 2007). **Anais...** [S.l.: s.n.], 2007. p.225–234.

ESCOBAR, P. et al. Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary. **Computer Standards Interfaces**, [S.l.], v.68, p.103378, 2020.

FARIHA, A. et al. Mining Frequent Patterns from Human Interactions in Meetings Using Directed Acyclic Graphs. In: ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, Berlin, Heidelberg. **Anais...** Springer Berlin Heidelberg, 2013. p.38–49.

FAZZINGA, B.; GALASSI, A.; TORRONI, P. A privacy-preserving dialogue system based on argumentation. **Intelligent Systems with Applications**, [S.l.], v.16, p.200113, 2022.

FEDERAL, M. P. **O que é a LGPD?** https://www.mpf.mp.br/servicos/lgpd/o-que-e-a-lgpd.

FERREIRA, M. et al. RuleKeeper: gdpr-aware personal data compliance for web frameworks. In: IEEE SYMPOSIUM ON SECURITY AND PRIVACY (SP), 2023. **Anais...** [S.l.: s.n.], 2023. p.2817–2834.

Flask (web framework). [S.l.]: Wikimedia, 2023.

FLEINER, R. Linking of Open Government Data. In: IEEE 12TH INTERNATIONAL SYMPOSIUM ON APPLIED COMPUTATIONAL INTELLIGENCE AND INFORMATICS (SACI), 2018. **Anais...** [S.l.: s.n.], 2018. p.1–5.

FORGÓ, N. et al. An ethico-legal framework for social data science. **International Journal of Data Science and Analytics**, [S.l.], v.11, n.4, p.377–390, May 2021.

FOTOPOULOU, E. et al. Linked Data Analytics in Interdisciplinary Studies: the health impact of air pollution in urban areas. **IEEE Access**, [S.l.], v.4, p.149–164, 2016.

FOUNDATION, O. K. **What is open?** 2023.

FRANCESCONI, E.; GOVERNATORI, G. Patterns for legal compliance checking in a decidable framework of linked open data. **Artificial Intelligence and Law**, [S.l.], v.31, n.3, p.445–464, Sep 2023.

FVG. **DAPP e Open Knowledge Brasil lançam Índice de Dados Abertos no país**. 2017.

GEBKA, E. et al. Generating Value with Open Government Data: beyond the programmer. In: INTERNATIONAL CONFERENCE ON RESEARCH CHALLENGES IN INFORMATION SCIENCE (RCIS), 2019. **Anais...** [S.l.: s.n.], 2019. p.1–2.

Git. [S.l.]: Wikimedia, 2023.

GOV.BR. **Política de Dados Abertos**. 2023.

GRADY, N. W. et al. Big Data: challenges, practices and technologies: nist big data public working group workshop at ieee big data 2014. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2014. **Anais...** [S.l.: s.n.], 2014. p.11–15.

GUAMáN, D. S.; DEL ALAMO, J. M.; CAIZA, J. C. GDPR Compliance Assessment for Cross-Border Personal Data Transfers in Android Apps. **IEEE Access**, [S.l.], v.9, p.15961–15982, 2021.

GUO, G. et al. Realistic Transport Simulation: tackling the small data challenge with open data. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2019. **Anais...** [S.l.: s.n.], 2019. p.4512–4519.

HARHOFF, D.; LAKHANI, K. R. The Empirical Scope of User Innovation. In: REVOLUTIONIZING INNOVATION: USERS, COMMUNITIES, AND OPEN INNOVATION. **Anais...** [S.l.: s.n.], 2016. p.67–87.

HO, H. Y. et al. Ranking hospitals' burn care capacity using cluster analysis on open government data. **Computer Methods and Programs in Biomedicine**, [S.l.], v.207, p.106166, 2021.

HYLAND, B. et al. **Best Practices for Publishing Linked Data**. 2014.

IM, S. et al. Keyword-Based SPARQL Query Generation System to Improve Semantic Tractability on LOD Cloud. In: EIGHTH INTERNATIONAL CONFERENCE ON INNOVATIVE MOBILE AND INTERNET SERVICES IN UBIQUITOUS COMPUTING, 2014. **Anais...** [S.l.: s.n.], 2014. p.102–109.

INDEX, O. D. **Open Data Index**. 2018.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, Adoption Barriers and Myths of Open Data and Open Government. **Information Systems Management**, [S.l.], v.29, n.4, p.258–268, 2012.

KHURSHID, M. M. et al. Modeling of Open Government Data for Public Sector Organizations Using the Potential Theories and Determinants—A Systematic Review. **Informatics**, [S.l.], v.7, n.3, 2020.

KIRSTEIN, F.; BOHLEN, V. IDS as a Foundation for Open Data Ecosystems. , Cham, p.225–240, 2022.

KIRSTEIN, F. et al. Piveau: a large-scale open data management platform based on semantic web technologies. In: THE SEMANTIC WEB, Cham. **Anais...** Springer International Publishing, 2020. p.648–664.

KIRSTEIN, F. et al. Ronda: real-time data provision, processing and publication for open data. In: ELECTRONIC GOVERNMENT, Cham. **Anais...** Springer International Publishing, 2021. p.165–177.

KOSINSKI, M.; STILLWELL, D.; GRAEPEL, T. Private traits and attributes are predictable from digital records of human behavior. **Proceedings of the National Academy of Sciences**, [S.l.], v.110, n.15, p.5802–5805, 2013.

LEUNG, C. K. Mathematical Model for Propagation of Influence in a Social Network. In: ENCYCLOPEDIA OF SOCIAL NETWORK ANALYSIS AND MINING, New York, NY. **Anais...** Springer New York, 2018. p.1261–1269.

LEUNG, C. K. et al. Big Data Science on COVID-19 Data. In: IEEE 14TH INTERNATIONAL CONFERENCE ON BIG DATA SCIENCE AND ENGINEERING (BIGDATASE), 2020. **Anais...** [S.l.: s.n.], 2020. p.14–21.

LEUNG, C. K. et al. An Innovative Fuzzy Logic-Based Machine Learning Algorithm for Supporting Predictive Analytics on Big Transportation Data. In: IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE), 2020. **Anais...** [S.l.: s.n.], 2020. p.1–8.

LEUNG, C. K.-S.; CARMICHAEL, C. L. FpVAT: a visual analytic tool for supporting frequent pattern mining. **SIGKDD Explor. Newsl.**, New York, NY, USA, v.11, n.2, p.39–48, may 2010.

LEUNG, C. K.-S.; MACKINNON, R. K.; TANBEER, S. K. Fast Algorithms for Frequent Itemset Mining from Uncertain Data. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 2014. **Anais...** [S.l.: s.n.], 2014. p.893–898.

LIN, B.-Y. et al. An intelligent Ama safety protection system based on smart IoT data and deep learning. In: SIXTH INTERNATIONAL SYMPOSIUM ON COMPUTER, CONSUMER AND CONTROL (IS3C), 2023. **Anais...** [S.l.: s.n.], 2023. p.122–125.

LOGAREZZI, L. **Guia prático da lei de acesso à informação [livro eletrônico]**. [S.l.: s.n.], 2016.

MACHADO, J. S. et al. Towards Open Data in Digital Education Platforms. In: IEEE 19TH INTERNATIONAL CONFERENCE ON ADVANCED LEARNING TECHNOLOGIES (ICALT), 2019. **Anais...** [S.l.: s.n.], 2019. v.2161-377X, p.209–211.

MARTíN-MONCUNILL, D.; ALONSO GAONA GARCíA, P.; RAJABI, E. Navigating through the Linking Open Data cloud datasets: preliminary ideas of a visual search tool based on its defined domains. In: WORKSHOP ON ENGINEERING APPLICATIONS - INTERNATIONAL CONGRESS ON ENGINEERING (WEA), 2015. **Anais...** [S.l.: s.n.], 2015. p.1–7.

Matplotlib. [S.l.]: Wikimedia, 2023.

MAYAUD, J. R.; TRAN, M.; NUTTALL, R. An urban data framework for assessing equity in cities: comparing accessibility to healthcare facilities in cascadia. **Computers, Environment and Urban Systems**, [S.l.], v.78, p.101401, 2019.

MCCRAE, J. P. **The Linked Open Data Cloud**. Accessed on October 19, 2023.

MCDERMOTT, P. Building open government. **Government Information Quarterly**, [S.l.], v.27, n.4, p.401–413, 2010. Special Issue: Open/Transparent Government.

MELLO, L. E. et al. **Opening Brazilian COVID-19 patient data to support world research on pandemics**. [S.l.]: Zenodo, 2020.

MONTASARI, R. The Potential Impacts of the National Security Uses of Big Data Predictive Analytics on Human Rights. In: **Countering Cyberterrorism**: the confluence of artificial intelligence, cyber forensics and digital policing in us and uk national cybersecurity. Cham: Springer International Publishing, 2023. p.115–137.

MURRAY-RUST, P. Open Data in Science. **Nature Precedings**, [S.l.], Jan 2008.

NOGUERAS-ISO, J. et al. Quality of Metadata in Open Data Portals. **IEEE Access**, [S.l.], v.9, p.60364–60382, 2021.

pandas (software). [S.l.]: Wikimedia, 2023.

PAREJA-LORA, A. et al. Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences: an introduction. In: DEVELOPMENT OF LINGUISTIC LINKED OPEN DATA RESOURCES FOR COLLABORATIVE DATA-INTENSIVE RESEARCH IN THE LANGUAGE SCIENCES. **Anais...** [S.l.: s.n.], 2019. p.ix–xxi.

PEDDI, P.; DASGUPTA, A.; GAIDHANE, V. H. Smart Irrigation Systems: soil monitoring and disease detection for precision agriculture. In: IEEE INTERNATIONAL IOT, ELECTRONICS AND MECHATRONICS CONFERENCE (IEMTRONICS), 2022. **Anais...** [S.l.: s.n.], 2022. p.1–7.

PETERS, M. A.; BRITEZ, R. G. **Open Education and Education for Openness**. [S.l.]: Sense Publishers, 2008. v.27.

PHILLIPS, B. UK further education sector journey to compliance with the general data protection regulation and the data protection act 2018. **Computer Law Security Review**, [S.l.], v.42, p.105586, 2021.

PIAO, C. et al. Privacy-preserving governmental data publishing: a fog-computing-based differential privacy approach. **Future Generation Computer Systems**, [S.l.], v.90, p.158–174, 2019.

Privacy and freedom. [S.l.: s.n.], 1969.

Project Jupyter. [S.l.]: Wikimedia, 2023.

Python (programming language). [S.l.]: Wikimedia, 2023.

QANBARI, S.; REKABSAZ, N.; DUSTDAR, S. Open Government Data as a Service (GoDaaS): big data platform for mobile app developers. In: INTERNATIONAL CONFERENCE ON FUTURE INTERNET OF THINGS AND CLOUD, 2015. **Anais...** [S.l.: s.n.], 2015. p.398–403.

RAHMATIKA, M. et al. An Open Government Data Maturity Model : a case study in bps-statistics indonesia. In: INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGY (ICOICT), 2019. **Anais...** [S.l.: s.n.], 2019. p.1–7.

RENZI, G. et al. A storytelling framework based on multimedia knowledge graph using linked open data and deep neural networks. **Multimedia Tools and Applications**, [S.l.], v.82, n.20, p.31625–31639, Aug 2023.

RHAHLA, M.; ALLEGUE, S.; ABDELLATIF, T. Guidelines for GDPR compliance in Big Data systems. **Journal of Information Security and Applications**, [S.l.], v.61, p.102896, 2021.

scikit-learn. [S.l.]: Wikimedia, 2023.

SILVA, P. et al. Using NLP and Machine Learning to Detect Data Privacy Violations. In: IEEE INFOCOM 2020 - IEEE CONFERENCE ON COMPUTER COMMUNICATIONS WORKSHOPS (INFOCOM WKSHPS). **Anais...** [S.l.: s.n.], 2020. p.972–977.

SOKOLOVSKA, A.; KOCAREV, L. Integrating Technical and Legal Concepts of Privacy. **IEEE Access**, [S.l.], v.6, p.26543–26557, 2018.

STF. **Lei Geral de Proteção de Dados (LGPD)**. https://portal.stf.jus.br/lgpd/.

SUBER, P. **Open Access**. [S.l.]: MIT Press essential knowledge, 2012.

SáNCHEZ-NIELSEN, E. et al. SuDaMa: sustainable open government data management framework for long-term publishing and consumption. **IEEE Access**, [S.l.], v.9, p.151841–151863, 2021.

TANG, R. et al. Open Government Data (OGD) sites and the sharing of country-specific real-time pandemic information: an investigation into covid-19 datasets available on worldwide ogds. **Information Processing Management**, [S.l.], v.60, n.6, p.103489, 2023.

TEMIZ, S. et al. Open data: lost opportunity or unrealized potential? **Technovation**, [S.l.], v.114, p.102535, 2022.

UBALDI, B. Open Government Data. , [S.l.], n.22, 2013.

UEDA, M. Licenses of Open Source Software and their Economic Values. In: SYMPOSIUM ON APPLICATIONS AND THE INTERNET WORKSHOPS (SAINT 2005 WORKSHOPS), 2005. **Anais...** [S.l.: s.n.], 2005. p.381–383.

UNDP. The impact of digital technology on human rights in Europe and Central Asia: trends and challenges related to data protection, artificial intelligence and other digital technology issues. **Istanbul: United Nations Development Programme**, [S.l.], 2023. https://www.undp.org/sites/g/files/zskgke326/files/2023-03/The

Visual Studio Code. [S.l.]: Wikimedia, 2023.

von Grafenstein, M.; JAKOBI, T.; STEVENS, G. Effective data protection by design through interdisciplinary research methods: the example of effective purpose specification by applying user-centred ux-design methods. **Computer Law Security Review**, [S.l.], v.46, p.105722, 2022.

WON, H. et al. An Advanced Open Data Platform for Integrated Support of Data Management, Distribution, and Analysis. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2021. **Anais...** [S.l.: s.n.], 2021. p.2058–2063.

WU, X. et al. Top 10 algorithms in data mining. **Knowledge and Information Systems**, [S.l.], v.14, n.1, p.1–37, Jan 2008.

ZAINAL, N. Z. et al. Open Government Data Use by Malaysian Researchers. Some empirical evidence. In: INTERNATIONAL CONFERENCE ON RESEARCH AND INNOVATION IN INFORMATION SYSTEMS (ICRIIS), 2019. **Anais...** [S.l.: s.n.], 2019. p.1–6.

ZHANG, A.; LV, N. Research on the Impact of Big Data Capabilities on Government's Smart Service Performance: empirical evidence from china. **IEEE Access**, [S.l.], v.9, p.50523–50537, 2021.

ZHANG, C.; YUE, P. Spatial grid based Open Government Data mining. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS), 2016. **Anais...** [S.l.: s.n.], 2016. p.192–193.

ÇALDAğ, M. T.; GöKALP, E. Understanding barriers affecting the adoption and usage of open access data in the context of organizations. **Data and Information Management**, [S.l.], p.100049, 2023.

ASSINATURAS

_____

Julio César Santos dos Anjos

_____

Prof. Dr. Cláudio Fernando Resin Geyer