TransInformação

ORIGINAL

Editora
Valéria dos Santos Gouveia Martins

Apoio
Universidade Federal de Minas Gerais (Projeto nº: 30201*42) e Conselho Nacional de Desenvolvimento Científico e Tecnológico.

Conflito de interesses
Os autores declaram que não há conflito de interesses.

Recebido
16 jan. 2024

Versão final
27 set. 2024

Aprovado
8 out. 2024

Catálogo de dados para descoberta e recuperação de dados abertos: uma solução baseada em APIs governamentais do Brasil

Data catalog for discovering and retrieving open data: a solution based on government APIs from Brasil

Patrícia Nascimento Silva[1], Gabriel Vieira Pereira da Silva[2], Douglas Rodrigues Fernandes Filho[2]

[1] Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Programa de Pós-Graduação em Gestão e Organização do Conhecimento. Belo Horizonte, MG, Brasil. Correspondência para/*Correspondence to*: P. NASCIMENTO SILVA. *E-mail*: <patricians@ufmg.br>.

[2] Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. Belo Horizonte, MG, Brasil.

Como citar este artigo/*How to cite this article*: Nascimento Silva, P.; Silva, G. V. P.; Fernandes Filho, D. R. Catálogo de dados para descoberta e recuperação de dados abertos: uma solução baseada em APIs governamentais do Brasil. *Transinformação*, v. 36, e2410694, 2024. https://doi.org/10.1590/2318-0889202436e2410694

Resumo

Com a evolução tecnológica e o surgimento de novos suportes, ambientes e ferramentas, as fontes de informação sofreram alterações nas formas de acesso e comunicação, sendo automatizadas por *Application Programming Interface* que permitem a representação, organização e recuperação de dados para diversos domínios. O catálogo, instrumento longínquo de organização e preservação em acervos, também sofreu modificações contemplando uma visão ampliada dos dados e metadados nas organizações. Unir esses recursos para descoberta e recuperação em fontes volumosas, valiosas, pouco padronizadas e dispersas como os dados abertos publicados pelos governos é uma solução plausível e necessária. O objetivo deste estudo foi modelar um catálogo de dados para as *Application Programming Interfaces* de dados governamentais abertos do Brasil. Caracterizada como descritiva e aplicada, a pesquisa foi dividida em três etapas, uma para cada objetivo específico, executadas entre 2023 e 2024. Na primeira etapa, o mapeamento das *Application Programming Interfaces*, realizado em estudo anterior, foi utilizado para estudo e identificação dos domínios, fontes de informação e ativos. Na segunda etapa foram definidos metadados para representação das *Application Programming Interfaces*, que considerou componentes padrões da arquitetura, o padrão e-PMG e um vocabulário controlado do governo para classificação. A terceira etapa consistiu na modelagem do catálogo de dados, por meio de um banco de dados relacional, com oito tabelas validadas com dados reais. O catálogo de dados permitiu descobrir metadados para recuperar conjuntos de dados abertos, fomentando o reúso e a interoperabilidade com outras aplicações, além disso, a solução é reprodutível para outros serviços e domínios de dados.

Palavras-chave: API. Catálogo de dados. Dados governamentais abertos. Fontes de informação. Metadados. Recuperação de informação.

Abstract

With technological evolution and the emergence of new media, environments and tools, information sources have under gone changes in the ways they are accessed and communicated, being automated by Application Programming Interfaces that allow data to be represented, organized and retrieved for various domains. The catalog, a long-standing tool for organizing and preserving collections, hás also under gone changes, taking on board na expanded vision of data and metadata in organizations. Combining these resources for Discovery and retrieval in voluminous, valuable, poorly standardizedanddispersedsourcessuch as open data published by governments is a plausible and necessary solution. The aim of this study was to model a data catalog for open government data Application Programming Interfaces of Brazil. Characterized as descriptive and applied, there search was divided in to three stages, one for each specific objective, carried out between 2023 and 2024. In the first stage, the mapping of Application Programming Interfaces, carried out in a previous study, was used to study and identify domains, information sources and assets. In the second stage, metadata was defined to representthe Application Programming Interfaces, takingin to account standard components of the architecture, the e-PMG standard and a controlled government vocabulary for classification. The third stage consisted of modeling the data catalog using a relational database with eight tables validated with real data. The data catalog made it possible to discover metadata to retrieve open data sets, encouraging reuse and interoperability with other applications, and the solutionis reproducible for other services and data domains.

Keywords: *API. Data catalog. Open government data. Information sources. Metadata. Information retrieval.*

Introdução

No contexto fundamental da Biblioteconomia e da Ciência da Informação, a maneira como os dados são organizados definem como eles podem ser pesquisados. Esta realidade se aplica a diversos suportes, ambientes e ferramentas informacionais e ganha destaque na "Era dos Dados" com o catálogo de dados corporativos (*Data Catalog*) que permite uma visão abrangente de todos os dados existentes em uma organização, subsidiando a descoberta de dados, a recuperação de informação e futuramente a criação de grafos de conhecimento (Olesen-Bagneux, 2023).

Com o avanço das tecnologias da informação e comunicação e principalmente da web, a disponibilização de dados, processáveis por máquinas, na *internet*, por meio de *Application Program Interface* (API), tornou-se uma fonte de informação da atualidade. Esta fonte de informação na web é utilizada por muitas empresas e governos para publicar e principalmente para trocar dados entre sistemas.

A interoperabilidade entre essas novas fontes de informação é possível graças a arquitetura de microsserviços que utiliza as APIs para implementar o acesso por máquina padronizado, com definições compartilhadas em suas documentações, para uma interação fácil e eficaz (Fowler, 2019; Saudate, 2021).

O movimento de acesso aberto, iniciado em 2001, durante a conferência *Budapest Open Access Iniciative*, fomentou o acesso aberto aos dados de pesquisa (Budapest Open Access Initiative, ©2024) e foi empreendido em outras áreas como a governamental. O termo *open government data* foi criado em 2007, em uma conferência em Sebastopol, Califórnia, para definir os princípios relacionados (Open Gov Data, 2007) e caracterizar os dados abertos sobre o governo, ou Dados Governamentais Abertos (DGA) (Albano, 2014). Essa nova tipologia de dados ganhou muito destaque, sendo considerado um fenômeno político e socioeconômico, promovendo envolvimento cívico e estimulando inovações no setor público e em diversas áreas (Nikiforova *et al.*, 2023).

Com o passar dos anos, surgiram legislações e políticas específicas para publicação de dados abertos pelos órgãos governamentais. No Brasil, o marco inicial foi a Lei de Acesso à Informação (LAI), Lei Nº 12.527/2011, mas a Política de Dados Abertos foi oficializada pelo Decreto Nº 8.777/2016,

somente em 2016 (Brasil, 2011; Brasil, 2016a). A partir dessas novas regulações, o volume de DGA aumentou significativamente, assim como a diversidade de formatos e padrões utilizados, tornando-se uma tarefa cada vez mais difícil colocá-los em um contexto significativo para extrair informações valiosas (Kalampokis; Tambouris; Tarabanis, 2013).

Considerando a importância dos DGA, as APIs são recursos tecnológicos em potencial neste contexto informacional, apoiando o acesso, a organização e a recuperação de grandes volumes de dados para serem reutilizados em análises e novos produtos e serviços de informação (Silva, 2022). No Brasil, a recuperação de DGA em APIs é dificultada pela ausência de um padrão geral para publicação e pela falta de centralidade das informações que estão fragmentadas nos sites e portais dos órgãos que disponibilizam esse serviço (Silva, 2023).

É nesse contexto informacional que o catálogo de dados pode ser utilizado para otimizar a descoberta de dados e a recuperação de informação, permitindo a democratização dos DGA e seu reúso. O catálogo é um longínquo instrumento de guarda e permanência que está diretamente associado à organização e controle de um acervo (Pedrão, 2019), mas com a evolução dos suportes e novos ambientes de informação tem sofrido alterações em relação a sua concepção e implementação. Para Olesen-Bagneux (2023) o catálogo de dados é um mecanismo de pesquisa, sendo essencialmente um banco de dados contendo metadados extraídos de fontes de informação. Na atualidade, essa ferramenta tem sido utilizada junto à governança de dados para acompanhar todo o ciclo de vida dos dados e permitir pesquisar e encontrar efetivamente os dados de uma organização.

Diante da necessidade de recuperação dos DGA, no contexto das fontes de informação automatizadas, o catálogo de dados pode ser uma solução interessante para este domínio de investigação, ainda pouco explorado, mas com um grande potencial a ser revelado. Considerando todo esse cenário, o estudo parte das seguintes problemáticas: quais os tipos de dados existentes nas APIs de DGA? Como um catálogo de dados pode ser utilizado para descoberta de dados nas APIs de DGA do Brasil? O objetivo da pesquisa foi modelar um catálogo de dados para APIs de DGA no contexto brasileiro. Especificamente buscou-se: (i) analisar os dados e metadados das APIs de DGA brasileiras, (ii) definir os atributos e metadados para representação das APIs e (iii) modelar um catálogo de dados. Metodologicamente a pesquisa foi dividida em três etapas, uma para cada objetivo específico, e foi caracterizada como descritiva e aplicada.

Este estudo integra um projeto de pesquisa maior, intitulado: Observatório de Dados Abertos (https://observatoriodedadosabertos.eci.ufmg.br/) que reúne estudos sobre interoperabilidade, organização, recuperação e reúso de dados abertos do governo brasileiro. Esta pesquisa apresenta uma solução prática para modelagem de um catálogo de dados para um domínio específico, considerando os aspectos da organização e representação da informação, temáticas essenciais da Ciência da Informação que fazem intersecção com ferramentas tecnológicas da Ciência da Computação para alcançar um objetivo comum dos DGA que é a sua recuperação e o reúso pela sociedade. Estima-se que otimizar a representação, organização e recuperação dos dados abertos poderá viabilizar o acesso e uso dos dados por parte da sociedade e potencializar a criação de novos conhecimentos e oportunidades de inovação.

Conceitos e Fundamentação Teórica

A Recuperação de Informação apesar de ser considerada uma área recente, cujo termo foi criado em 1951 por Calvin Mooers, é um desafio vivenciado desde a antiguidade pela humanidade que vem organizando a informação para sua recuperação em diferentes suportes nas bibliotecas

(Baeza-Yates; Ribeiro-Neto, 2013). Com a rápida evolução da tecnologia da informação e comunicação e o surgimento de novos suportes informacionais e ambientes digitais, o desafio não foi cessado, e sim alterado de contexto.

O conceito e a noção de documento, muito discutida por autores clássicos da Documentação e da Ciência da Informação como Otlet e Briet ou como informação por Buckland, Capurro e Hjorland (Lara, 2010) é abrangente. No entendimento de Cunha e Cavalcanti (2008, p. 132) o documento é "[...] 2.4 Qualquer tipo de registro, sejam quais forem os seus dados ou configurações, quer se encontre em papel, pergaminho ou filme, ou em qualquer outro material". Essa definição admite incluir outros elementos como dados, informação e conhecimento, em diferentes suportes, considerando as características das fontes de informações atuais e contextos contemporâneos como os DGA.

Os DGA são compreendidos como um subdomínio dos dados abertos que conforme sua definição podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa, com a exigência de atribuição à fonte original e do compartilhamento pelas mesmas licenças em que foi apresentado (Open Knowledge, 2024). Quando esses dados abertos compreendem os dados públicos, provenientes de atividades ou bases governamentais, são denominados DGA. As iniciativas de DGA foram observadas em vários países a partir de 2009, em decorrência do movimento de acesso aberto e do governo aberto nos anos 2000 (United Nations Economics Commission for Europe, 2013; Budapest Open Access Initiative, ©2024). Seus princípios estabelecem recomendações sobre as melhores práticas de publicação na *internet* recomendando que os dados sejam completos, primários, atualizados, acessíveis, processáveis por máquina, com acesso não discriminatório, em formatos não proprietários e livres de licença (Open Gov Data, 2007). Essas recomendações envolvem a recuperação dos dados abertos, em fontes na web, de forma automatizada.

Desde o surgimento da *internet*, até então denominada ARPAnet, em 1969, nos Estados Unidos da América, o surgimento da World Wide Web, desenvolvida pelo programador inglês Tim Berners-Lee, em colaboração com o cientista da computação Robert Cailliau, em 1989, até os dias atuais, os ambientes têm se modificado consideravelmente. A cada nova evolução ou fase da web aumenta a quantidade de recursos, a interação dos usuários e as tecnologias, como a web semântica, para adicionar significado aos recursos e permitir a cooperação entre humanos e máquinas, e novos meios de armazenamento como o armazenamento em nuvem (Araújo; Fachin, 2015; Berners-Lee; Hendler; Lassila, 2001).

Em meio ao contexto da recuperação de informação e das evoluções da web, as APIs se caracterizam como uma forma de partilhar recursos entre sistemas, sendo também utilizadas como fontes de informação.

Uma API, ou uma interface de programação de aplicação, são padrões de programação para acesso a aplicações disponibilizadas na web. Elas podem ser consideradas uma forma de comunicação com outros sistemas que procura abstrair totalmente os detalhes da implementação subjacente, ou pelo menos o que for possível, permitindo ao usuário não familiarizado com determinado sistema recuperar as informações que necessita, de forma fácil e rápida. Uma API expande o público-alvo da aplicação e promove a interoperabilidade, mas para cumprir com esse objetivo, a API deve ser o mais autoexplicativa possível, implementar versões e ter uma boa documentação (Saudate, 2021).

Assim como as APIs, o catálogo de dados é um instrumento imerso ao fenômeno do *big data*, com grandes e heterogêneos volumes de dados, e está relacionado à descoberta de dados, ou seja, pesquisar e encontrar efetivamente os dados. Diferentemente da pesquisa nos dados, que

inclui o conteúdo, a pesquisa por dados inclui os metadados e pode ocorrer de forma aleatória ou ser estruturada. Quando estruturada é realizada de maneira formalizada em uma solução projetada para esse fim, como o catálogo de dados. A partir do catálogo, a descoberta de dados começa com a constatação que os dados existem, sem se concentrar no seu conteúdo (Olesen-Bagneux, 2023).

O catálogo de dados no contexto de uma API se concentra em apoiar a identificação de recursos da aplicação e a pesquisa por dados. No domínio dos DGA, isso significa permitir ao usuário identificar quais dados são disponibilizados para cada conjunto de dados abertos, para em um segundo momento possibilitar a recuperação de informação, ou seja, a pesquisa nos dados disponibilizados.

Procedimentos Metodológicos

A pesquisa foi caracterizada como descritiva, por descrever as características de uma população ou determinado fenômeno (Gil, 1999), neste estudo representado pelas APIs de DGA, e aplicada, por ter um interesse prático, onde os resultados serão utilizados imediatamente na solução de um problema real (Lakatos; Marconi, 2002), ou seja, a modelagem de um catálogo de dados. A investigação foi estruturada em três etapas: Etapa 1 - Estudo das APIs de DGA existentes no Brasil; Etapa 2 - Definição de metadados para representação das informações das APIs e Etapa 3 - Modelagem do catálogo de dados.

É importante ressaltar que este estudo considerou o objetivo do catálogo de dados, definido na literatura por Olesen-Bagneux (2023, p. 77), para: "saber que tipo de dados sua organização detém", esse objetivo vai de encontro com a problemática do estudo, no contexto dos DGA, que envolve saber os tipos de dados existentes nas APIs de DGA. No entanto, algumas adaptações foram necessárias para a modelagem deste catálogo de dados, visto que os dados não foram acessados diretamente em seus sistemas de origem e sim pelas APIs que disponibilizam os DGA de forma aberta para sociedade.

A Etapa 1 foi realizada com um dos artefatos do projeto que este estudo integra e mapeou a recuperação de dados abertos em APIs de DGA no Brasil. Uma pesquisa exploratória nos sites, instrumentos e documentos da política de dados abertos identificou as APIs de dados abertos dos órgãos do governo federal e mapeou, até 2023, 19 APIs de 14 órgãos que disponibilizam dados abertos (Nascimento Silva, 2023; Nascimento Silva; Da Silva, 2023). Para cada API foi observado seu conteúdo, a estrutura de dados, assim como os elementos que integram a API (rotas, parâmetros, retornos) para definição dos domínios, fontes de dados e ativos.

A Etapa 2 definiu uma estrutura padrão, a partir dos elementos que compõem as APIs, que deve ser considerada para representar e recuperar dados e informações no catálogo de dados. Foram considerados elementos para representação descritiva ou catalogação de dados abertos, envolvendo instrumentos do governo como o padrão de metadados e-PMG (https://www.gov.br/governodigital/pt-br/governanca-de-dados/PMGVersao1_1.pdf), um padrão específico para representação de informação no governo eletrônico, e da representação temática, a partir de um vocabulário controlado para classificação das APIs. De forma adicional, foram incluídos metadados técnicos para um maior detalhamento das APIs.

A Etapa 3 consistiu na modelagem do catálogo de dados, a partir dos resultados obtidos nas Etapas 1 e 2. A documentação da modelagem proposta foi o objeto em destaque nesta etapa que tem como objetivo relatar a arquitetura do catálogo de dados, assim como apresentar as definições, ações e procedimentos realizados. A Figura 1 apresenta um resumo das etapas, destacando os principais procedimentos, pré-requisitos utilizados, artefatos gerados e trabalhos futuros após a modelagem (implementação e consultas).

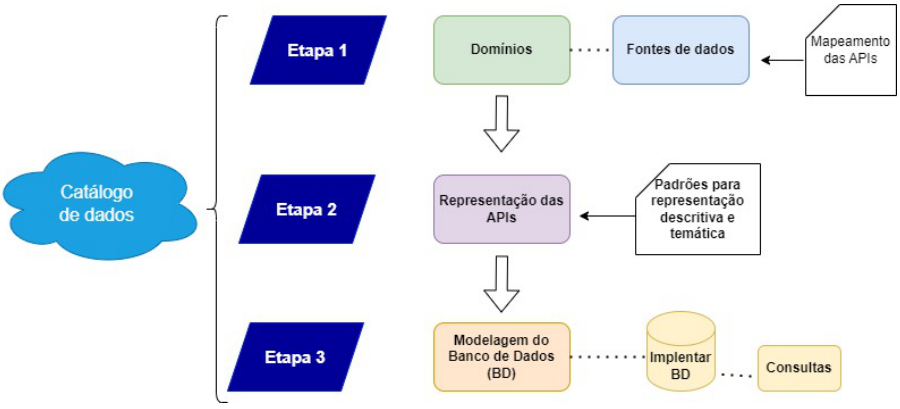

Figura 1 – Etapas metodológicas para implementação do catálogo de dados.
Fonte: Elaborada pelos autores (2024).

As três etapas foram iniciadas no segundo semestre de 2023 e concluídas em 2024. A equipe de trabalho envolveu inicialmente dois pesquisadores e a partir de outubro de 2023, três, de forma direta. Todas as atividades do projeto foram gerenciadas pela primeira autora e coordenadora do projeto, seguindo os eventos e artefatos do *framework* Scrum, com ciclos (*sprints*) semanais. O ambiente de trabalho (diretórios com documentações e modelagens, ferramentas de gestão e acompanhamento do projeto e salas de reunião online), foi construído em ferramentas web com de armazenamento em nuvem e utilizado por todos os integrantes para execução das atividades. O *framework* Scrum foi selecionado por ser uma metodologia ágil empregada em projetos em que não se conhece todos os requisitos, e mudanças podem ocorrer em seu desenvolvimento (Valente, 2020), situação identificada já na Etapa 1 ao identificar APIs heterogêneas e despadronizadas.

Resultados e Discussão

Na Etapa 1 foi feita uma análise das 19 APIs mapeadas, a fim de identificar as áreas que estavam relacionadas e a estrutura básica comum para a maioria das APIs de dados abertos. Conforme Olesen-Bagneux (2023), a eficácia do catálogo de dados depende de como ele é estruturado e isso afeta diretamente a sua capacidade de pesquisa.

Assim, a partir dessa estrutura básica e do conteúdo identificado nas APIs foi definida a organização do catálogo de dados. Essa organização considerou a classificação das APIs em domínios, fontes de dados e ativos. A maioria das APIs eram bem distintas e envolviam diversos órgãos, desta forma foram definidos 11 domínios, conforme classificação do Vocabulário Controlado de Governo Eletrônico (VCGE), e 14 fontes de dados, em uma relação direta com os órgãos responsáveis, conforme apresentado no Quadro 1.

Para classificação das APIs e definição dos domínios foi escolhido o VCGE, um vocabulário controlado para indexar informações do governo federal brasileiro que abrange os assuntos dos domínios especificados no âmbito do governo e dos padrões vigentes de interoperabilidade (ePING, https://eping.governoeletronico.gov.br/) para o governo digital no Brasil (Brasil, 2016b). Optou-se por criar domínios mais abrangentes, visto que as APIs são bem diversas e apresentam conteúdos bem diversificados sobre uma área de assunto. Cada API foi considerada um ativo, visto que tinham características próprias. Destaca-se que nesta modelagem inicial foi considerado como ativo para o Instituto Brasileiro de Geografia e Estatística (IBGE) somente a API de dados agregados, por ser mais "conhecida", com dados de pesquisas e censos, e por sua estrutura manter a conformidade com as outras APIs disponibilizadas pelo órgão.

Quadro 1 – Organização de domínios, fontes de dados e ativos do catálogo de dados.

Domínio	Fonte dos dados	Ativo
Administração	Ministério da Economia	Portal Nacional de Contratações Públicas (PNPC)
	Câmara dos Deputados	Legislativo (Camara de deputados federal)
	Controladoria-Geral da União	*Application Programming Interface* (API) *Representational State Transfer* (REST) do Portal da Transparência do Governo Federal
		API REST do Portal de Dados Abertos
	Senado Federal	Senado
Assistência Hospitalar e Ambulatorial	Agência Nacional de Saúde	Operadoras *Application Programming Interface* (API)
Clima	Instituto Nacional de Pesquisas Espaciais	Previsão de Tempo em *Extensible Markup Language* (XML) – Centro de Previsão de Tempo e Estudos Climáticos do Instituto Nacional de Pesquisas Espaciais (CPTEC/INPE)
Compras governamentais	Ministério da Economia	API de Compras Governamentais
Cultura	Ministério da Cultura	Sistema de Apoio às Leis de Insentivo à Cultura (SALIC) propostas de projetos culturais a serem incentivados pelo Ministério da Cultura)
Economia e Finanças	Banco Central do Brasil (BACEN)	BACEN
	Ministerio da Fazenda	Dados Abertos de Custos com Depreciação do Governo Federal
Fiscalização do estado		Sistema de Análise da Dívida Pública, Operações de Crédito e Garantias da União, Estados e Municípios (Sapidem)
Fiscalização do Estado		Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro (SICONFI)
Pesquisa e Desenvolvimento	Instituto Brasileiro de Geografia e Estatística (IBGE)	Dados agregados do IBGE
	Instituto de Pesquisa Econômica Aplicada	Atlas do Estado Brasileiro (IPEA)
		IPEA - DATA
Previdência social	Sistema de Informações dos Regimes Públicos de Previdência Social (CADPREV)	CADPREV
Saúde	Ministério da Saúde	Departamento de Monitoramento, Avaliação e Disseminação de Informações Estratégicas em Saúde (DEMAS)

Fonte: Elaborada pelos autores (2024).

Na Etapa 2, os elementos comuns à estrutura das APIs, observados na Etapa 1, foram analisados para definição de uma estrutura padrão. Conforme destacado por Kobashi (1996), para representar a informação contida em um documento, o profissional da informação precisa ter conhecimento da instituição que mantém o arquivo (estrutura e funcionamento), da essência documental (de que trata e para que fins foi criado o documento) e a quem se destina o uso. Dessa forma, além da estrutura comum observada nas APIs foram analisados os ativos, o contexto e a tipologia dos dados abertos disponibilizados.

A arquitetura de microsserviços, definida e padronizada por APIs, foi um fator decisivo para estabelecer os elementos para representação das APIs no catálogo. Com isso, foram considerados os elementos que compõem uma API. Os microsserviços adotam APIs como canais para atender aos usuários, sendo similar a uma "função" na programação, que é composta pelo nome da função, parâmetros de entrada e a saída a ser gerada (Ghani; Zakaria, 2018). Com isso, baseado nas características dessa arquitetura e do seu funcionamento, os elementos comuns que compõem as APIs, considerados para sua representação foram: *URL* padrão, *endpoint* ou rotas disponíveis, também conhecido como *path*, parâmetros disponíveis para cada *endpoint* e retorno do *endpoint* (metadados retornados) com um exemplo do objeto retornado pela API.

A Figura 2 apresenta a validação dos elementos definidos na API da Câmara dos Deputados que possui a URL padrão: https://dadosabertos.camara.leg.br/api/v2/ e os *endpoints*: (1) /blocos (2) /deputados (3) /eventos (4) /frentes. Cada *endpoint* admite parâmetros específicos e para cada parâmetro repassado haverá um retorno associado. É importante destacar que a URL padrão irá executar somente se possuir a indicação do *endpoint* e respectivos parâmetros, conforme apontado na documentação da API (https://dadosabertos.camara.leg.br/swagger/api.html).

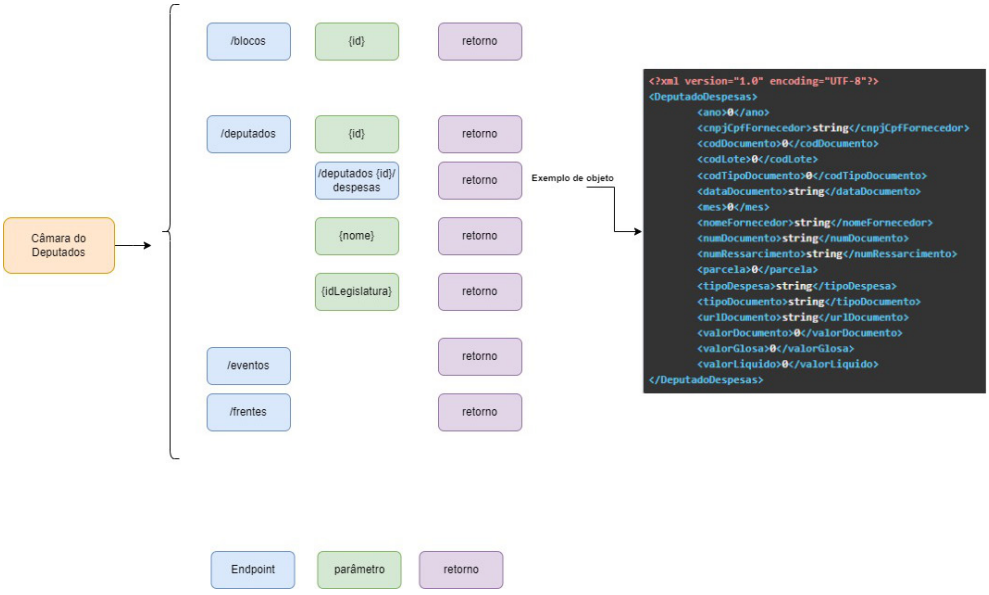

Figura 2 – Representação da Application Programming Interface – Exemplo Câmara dos Deputados.
Fonte: Elaborada pelos autores (2024).

Conforme destacado por Ferneda (2012), um Sistema de Recuperação de Informação (SRI) deve representar o conteúdo dos documentos do corpus e apresentá-los ao usuário de uma forma que permita a rápida seleção dos itens que satisfazem total ou parcialmente a necessidade de informação. Fazendo uma associação com as APIs, esta necessidade que geralmente é formalizada por uma expressão de busca e formulada pelo usuário, em um SRI convencional, agora está padronizada e definida com *endpoints* e parâmetros, estruturados no catálogo de dados. Com isso, considerando ainda a analogia com um SRI, a operação de tradução (representação da expressão de busca) que é comparada com a representação dos documentos no corpus, pela função de busca, no modelo de recuperação de informação, tornou-se uma requisição, uma operação simples e direta na API.

A partir dessas definições sobre a recuperação de informação e características observadas, foram mapeados todos os metadados necessários para descrever e representar as APIs, contendo a descrição da estrutura e caracterização das fontes. O padrão de metadados do governo eletrônico (e-PMG) foi considerado para caracterização dos componentes, contudo não havia uma correspondência direta com todos os elementos das APIs. Apesar de as APIs governamentais não utilizarem um padrão de metadados, essa condição não impediu a modelagem do catálogo, que considerou recursos comuns e mais gerais. Destaca-se que foram realizados vários testes, com dados já mapeados nas APIs, para validar se os metadados definidos atendiam as APIs mais diversas.

O uso apropriado do padrão de metadados assegura uma descrição normalizada e como consequência uma representação de qualidade que facilita o intercâmbio de informações e a recuperação da informação, promovendo a interoperabilidade (Alves, 2010). Nesse sentido, estabelecer um padrão de metadados para o desenvolvimento de uma API é essencial para promover a sua interoperabilidade com outros serviços, permitir a integração com outros *frameworks*, inclusive semânticos, catálogos e bases de dados. Além disso, essa padronização poderá contribuir com a governança de dados nas instituições e no contexto dos dados abertos, com a conformidade regulatória, implementada por legislações e pela Política de Dados Abertos do governo federal brasileiro.

Com os resultados das etapas 1 e 2 foi modelado em um banco de dados relacional para armazenamento dos metadados, ou seja, a modelagem propriamente dita do catálogo de dados

para as APIs (Etapa 3). Conforme Olesen-Bagneux (2023, p. 25) um catálogo de dados é "[...] essencialmente um banco de dados contendo metadados que foram enviados ou extraídos de fontes de dados no cenário de TI de uma determinada empresa".

Para modelagem do catálogo de dados foram criadas oito tabelas: 'apis', 'institutions', 'api_endpoints', 'endpoint_parameters', 'parameters', 'api_responses', 'tags' e 'endpoint_tags', conforme detalhado no Quadro 2. O diagrama completo com todas as tabelas e relacionamentos foi apresentado na Figura 3.

Quadro 2 – Detalhamento das tabelas do banco de dados.

1 de 2

Atributo	Tipo	Descrição
Tabela apis		
id	INT PRIMARY KEY NOT NULL	Identificador (ID) da Application Programming Interface (API), número que será o identificador da API.
name	VARCHAR(100) NOT NULL	Nome da API conforme indicação do órgão.
description	VARCHAR(300) NOT NULL	Descrição breve da API, considerando seu objetivo principal.
base_url	VARCHAR(500) NOT NULL	URL padrão para acesso aos diferentes enpointsda API.
documentation	VARCHAR(500) NOT NULL	URL da página de documentação.
response_type	VARCHAR(10) NOT NULL	Extensão do tipo de retorno, como: *JavaScript Object Notation* (JSON), *Extensible Markup Language* (XML), *Comma-separated values* (CSV), etc.
colection_date	DATE NOT NULL	Data da última coleta dos dados realizada.
authentication	CHAR(1), NOT NULL	Indica se a API precisa ou não de autenticação para ser utilizada, valor' S' ou 'N'.
version	VARCHAR(10)	Versão da API, caso tenha essa informação disponível.
institution_id	INT NOT NULL	Identificador do órgão federal responsável pela API.
last_api_update_date	DATE	Data da última atualização da API, de acordo com a documentação.
classification	VARCHAR(60)	Classificação da API de acordo com o Vocabulário Controlado do Governo Eletrônico (VCGE).
Tabela institutions		
id	INT PRIMARY KEY NOT NULL	Id da instituição, número que será o identificador do órgão federal responsável por alguma API.
name	VARCHAR(100) NOT NULL	Nome do orgão federal responsável por alguma API.
acronym	VARCHAR(10) NOT NULL	Sigla do órgão federal.
Tabela api_endpoints		
id	INT AUTO_INCREMENT PRIMARY KEY NOT NULL	ID do endpoint, número que identifica um endpoint da API.
api_id	INT NOT NULL	ID da API relacionada.
name	VARCHAR(100) NOT NULL	Nome do Endpoint.
short_name	VARCHAR(200) NOT NULL	Nome mais curto/ Nome mais explicativo.
url	VARCHAR(500) NOT NULL	Uniform Resource Locator (URL) do endpoint, parâmetros sem valores.
description	VARCHAR(300) NOT NULL	Descrição do que é retornado pelo endpoint.
example	VARCHAR(500) NOT NULL	Exemplo de URL com valores no lugar dos parâmetros.
api_responses_id	INT	ID do exemplo do retorno relacionado.
Tabela endpoint_parameters		
id	INT PRIMARY KEY NOT NULL	ID da tabela, número que identifica um parâmetro de um endpoint específico.
endpoint_id	INT NOT NULL	ID do endpoint relacionado.
parameter_id	INT NOT NULL	ID do parâmetro referente ao endpoint relacionado.
required	CHAR(1) NOT NULL	Indica se o parâmetro é obrigatório ou não.
Tabela parameters		
id	INT PRIMARY KEY NOT NULL	ID do parâmetro, número que identifica o parâmetro.
name	VARCHAR(100) NOT NULL	Nome do parâmetro.
type	VARCHAR(20) NOT NULL	Tipo de dado do parâmetro, exemplo: VARCHAR, CHAR, INT, entre outros.
format	VARCHAR(20)	Caso informado, o formato deve ser informado como parâmetro.
description	VARCHAR(300)	Descrição sobre qual informação o parâmetro repassa.

Quadro 2 – Detalhamento das tabelas do banco de dados.

2 de 2

Atributo	Tipo	Descrição
Tabela tags		
id	INT PRIMARY KEY NOT NULL	ID da Tag, número que identifica a tag.
name	VARCHAR(100) NOT NULL	Nome da Tag.
Tabela endpoint_tags		
id	INT PRIMARY KEY NOT NULL	ID da tabela, número que identifica a tag de uma API.
endpoint_id	INT NOT NULL	ID do Endpoint relacionado.
tag_id	INT NOT NULL	ID da Tag que quero realizar a ligação com a API.
Tabela api_responses.		
id	INT PRIMARY KEY NOT NULL	ID do retorno, número que identifica o retorno, resposta da API.
description	VARCHAR(300)	Descrição do que é retornado pelo endpoint.
data dictionary	VARCHAR(400)	URL do dicionário de dados existente ou respectivo glossário.
response_example	JSON	Exemplo do arquivo json ou link do endpoint.

Fonte: Elaborada pelos autores (2024).

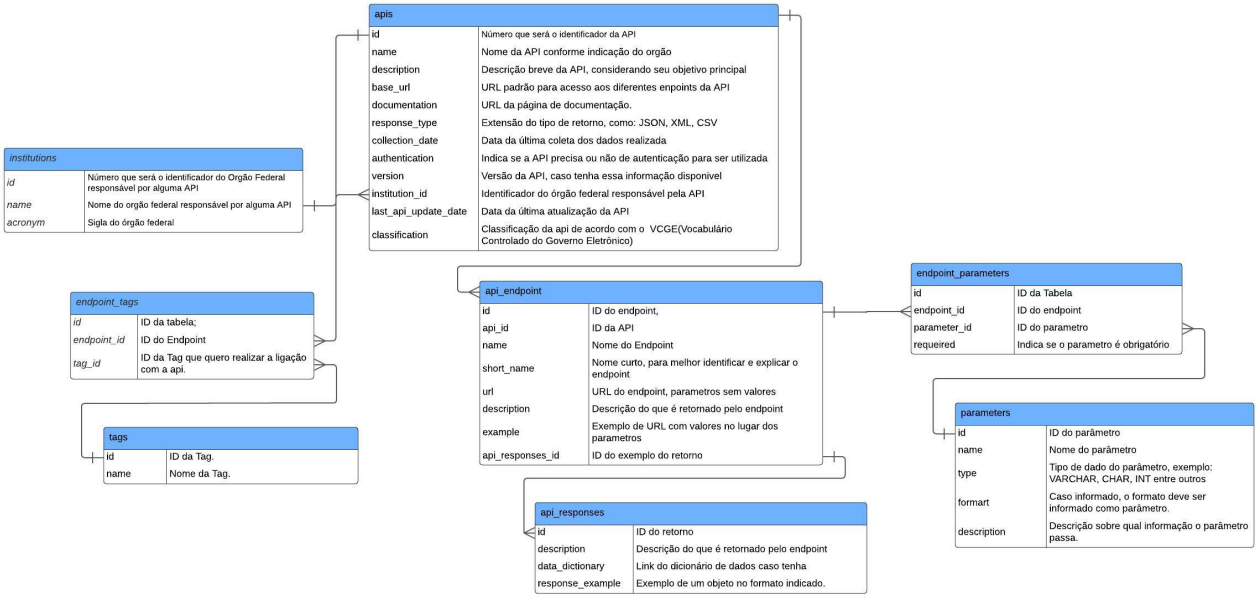

Figura 3 – Modelagem Entidade Relacionamento do banco de dados.
Fonte: Elaborada pelos autores (2024).

A nomenclatura do esquema seguiu o padrão *snake* case *patterns*. Neste padrão utiliza-se a separação das palavras com o caractere *underscore*, como por exemplo: 'api_endpoints'. Adotar o idioma inglês para as nomenclaturas foi uma decisão de *design* e boas práticas, uma vez que ele é amplamente utilizado entre desenvolvedores de sistemas e comandos de banco de dados, proporcionando uma maior consistência. A chave primária das tabelas, ou seja, o elemento que identifica unicamente cada linha foi incrementado sequencialmente, de forma automática, a cada inserção na tabela. Além dos atributos criados para descrição e detalhamento das APIs, foram criados atributos de controle para identificar, por exemplo, a exigência de autenticação em uma API.

Todas as tabelas criadas para o catálogo estão relacionadas e os relacionamentos foram pautados na composição padrão de uma API. Assim, uma API mantém uma conexão com N *endpoints*, os quais estão associados a N parâmetros. Cada *endpoint*, por sua vez, possui uma relação com um response (retorno) específico. As APIs foram classificadas conforme os assuntos do vocabulário VCGE e suas especificações. Destaca-se que devido às singularidades de cada API, os assuntos de VCGE ainda eram muito abrangentes, o que demandou a criação de *tags* mais

específicas para categorizar os *endpoints*. Assim, cada *endpoint* pode estar associado a diferentes tags, que são responsáveis por classificar tematicamente os recursos disponibilizados. Essas *tags* foram criadas a partir de uma análise de assunto realizada pelos autores, com os resultados da Etapa 1 e consultas realizadas nas APIs.

Ao longo desta pesquisa foi possível observar que o catálogo de dados, neste estudo, também pode ser considerado um repositório de metadados, visto que seu objetivo é a pesquisa por dados em um banco de dados de referência. A partir dessa solução, um cidadão poderá saber quais tipos de informação determinada API retorna, respondendo a problemática deste estudo que envolve a descoberta de dados nas APIs de DGA do Brasil.

É importante destacar que muitas APIs ainda estão em desenvolvimento, apresentam instabilidades e sofrem modificações. Na documentação da API da Câmara, Câmara dos Deputados (2024), por exemplo, é informada a identificação da versão atual com a data da publicação e um alerta indicando que a versão "ainda é incompleta, sujeita a mudanças e que não substitui a versão original dos Dados Abertos", fazendo referência à página de publicação dos conjuntos e dados abertos do órgão.

Outro destaque é referente à documentação do VCGE, que estava disponibilizado no site do governo federal, mas em uma consulta, realizada em setembro de 2024, já não estava disponível, sendo acessado somente por um serviço de arquivo na *internet* (https://web.archive.org/web/20220124204600/https://www.gov.br/governodigital/pt-br/governanca-de-dados/vcge_2_1_0.pdf). A indisponibilidade de documentos oficiais é um problema que também se repete no acesso às APIs, visto que há alterações de versões, *links* e documentações, às vezes, sem nenhum tipo de comunicação aos usuários.

Considerando que o catálogo de dados terá uma interface (*front-end*) para os usuários realizarem pesquisas, serão criadas consultas para permitir executar buscas no catálogo de dados, por meio de filtros e pesquisa direta. Com isso, a entrada principal do catálogo de dados irá considerar a arquitetura de domínio criada e a representação das APIs será refletida na navegabilidade do catálogo de dados, permitindo aos usuários pesquisar por dados abertos de forma objetiva e padronizada. Os scripts para criação das tabelas e as consultas estão disponíveis no diretório do projeto (https://github.com/ObservatorioDadosAbertos/) e após conclusão da implementação do catálogo serão disponibilizados com acesso aberto.

Conclusão

A modelagem de um catálogo de dados é um dos artefatos do projeto de pesquisa "Observatório de Dados Abertos" que investiga temáticas recentes e inéditas sobre dados abertos com processamento automatizado, no contexto da representação, organização e recuperação de dados e informação, em interface com a Ciência da Informação e a Ciência da Computação. A partir de questionamentos sobre os dados existentes nas APIs de DGA e da função do catálogo de dados, a pesquisa teve como objetivo apresentar uma proposta inicial de modelagem de um catálogo de dados para APIs de DGA no contexto brasileiro.

Metodologicamente a construção do catálogo envolveu três etapas, a saber: Etapa 1 - Estudo das APIs de DGA existentes no Brasil; Etapa 2 – Definição de metadados para representação das informações das APIs e Etapa 3 – Modelagem do catálogo de dados. O mapeamento das APIs existentes foi realizado em uma pesquisa anterior e foi basilar para o estudo das APIs, permitindo a definição dos domínios, fontes de informação e ativos na Etapa 1. Na Etapa 2, a representação das APIs no catálogo de dados considerou os componentes padrões da arquitetura das APIs, utilizando,

quando possível, elementos do padrão de metadados e-PMG em conjunto com o VCGE para classificação temática das APIs.

Com base nos resultados das Etapas 1 e 2, a Etapa 3 consistiu na modelagem do catálogo de dados, por meio da criação de um banco de dados relacional com oito tabelas, refletindo a estrutura de uma API. A solução envolveu conceitos e técnicas para representação, organização e recuperação de dados nas APIs de DGA, bem como padrões de projeto de Sistemas de Informação e fundamentos de Engenharia de Software, considerando elementos da pesquisa por dados, ou seja, a descoberta dos metadados disponibilizados nestas APIs, que atualmente estão descentralizadas em portais e serviços de dados abertos nos órgãos do governo federal. Esta etapa permitiu a melhor compreensão do estudo, por todos os pesquisadores, sendo fundamental para manutenibilidade do catálogo de dados que estará em constante evolução para refletir as modificações das fontes de informação.

Em se tratando de APIs governamentais, um desafio constante é lidar com a instabilidade dos ambientes (servidores de aplicação, repositórios e ferramentas de documentação) e as modificações realizadas sem controle de versão. Essas questões afetam diretamente o catálogo que será alimentado com os dados das APIs oficiais dos órgãos do governo federal, podendo gerar retrabalho e manutenções não planejadas. Instabilidades em instrumentos do governo, como o VCGE, e a inexistência de padrões de representação e organização para o desenvolvimento de APIs nos órgãos federais também ampliaram os desafios, especialmente no desenvolvimento da modelagem.

Conclui-se que o catálogo de dados é um instrumento que permitiu organizar e armazenar os metadados das APIs de DGA e esta solução pode ser utilizada em outros domínios de dados. Destaca-se que apesar de a arquitetura do catálogo ter sido concluída, esse é um primeiro *release*. Assim, modificações não estão descartadas, visto que algumas APIs de DGA são bem específicas e podem sofrer alterações, assim como em qualquer projeto de *software*.

Estima-se que o catálogo de dados abertos permita aos cidadãos identificarem as APIs de DGA existentes no Brasil e descobrir metadados para recuperar conjuntos de dados, incentivando o reúso dos dados abertos e a interoperabilidade com outras aplicações. Como trabalhos futuros, consultas serão implementadas em um *front-end* do catálogo de dados, para pesquisa e interação dos usuários, e o banco de dados, "populado" com as APIs de DGA, será utilizado na implementação do catálogo de dados abertos.

Referências

Albano, C. S. *Dados governamentais abertos*: proposta de um modelo de produção e utilização de informações sob a ótica conceitual da cadeia de valor. 2014. Tese (Doutorado em Administração) – Universidade de São Paulo, São Paulo, 2014. Disponível em: http://www.teses.usp.br/teses/disponiveis/12/12139/tde-03062014-170642/. Acesso em: 8 jan. 2024.

Alves, R. C. V. *Metadados como elementos do processo de catalogação*. 2010. 132 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010. Disponível em: https://repositorio.unesp.br/handle/11449/103361. Acesso em: 26 set. 2024.

Araújo, N. C.; Fachin, J. Evolução das fontes de informação. *Biblos*, v. 29, n. 1, p. 81-96, 2015. Disponível em: https://periodicos.furg.br/biblos/article/view/5463. Acesso em: 19 abr. 2022.

Baeza-Yates, R.; Ribeiro-Neto, B. *Recuperação de Informação*: conceitos e tecnologia das máquinas de busca. Porto Alegre: Bookman, 2013.

Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Scientific American*, v. 284, n. 5, p. 34-43, 2001. Disponível em: http://www.scientificamerican.com/article.cfm?id=the-semantic-web. Acesso em: 10 jan. 2024.

Brasil. Lei nº 12.527 de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. *Diário Oficial da União*, Brasília, 18 nov. 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 10 jan. 2024.

Brasil. Lei nº 8.777, de 11 de maio de 2016. Institui a Política de Dados Abertos do Poder Executivo federal. *Diário Oficial da União*, Brasília, 11 maio. 2016a. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm. Acesso em: 10 jan. 2024.

Brasil. Ministério da Gestão e da Inovação em Serviços Públicos. *Vocabulário Controlado*: Vocabulário Controlado do Governo Eletrônico - VCGE, interface de comunicação, interoperabilidade, ePING. 2016b. Disponível em: https://www.gov.br/governodigital/pt-br/governanca-de-dados/vocabulario-controlado-do-governo-eletronico. Acesso em: 11 jan. 2024.

Budapest Open Access Initiative. *Read the Declaration*. Budapest: BOAI, ©2024. Disponível em: https://www.budapestopenaccessinitiative.org/read/. Acesso em: 10 jan. 2024.

Câmara dos Deputados. Dados Abertos. Brasília: Câmara dos Deputados, 2024. Disponível em: https://dadosabertos.camara.leg.br/swagger/api.html. Acesso em: 26 set. 2024.

Cunha, M.B. Cavalcanti, C.R.O. *Dicionário de Biblioteconomia e Arquivologia*. Brasilia: Briquet de Lemos/Livros, 2008.

Ferneda, E. *Introdução aos modelos computacionais de recuperação de informação*. Rio de Janeiro: Ciência Moderna, 2012.

Fowler, S. *Microserviços prontos para produção*: construindo sistemas padronizados em uma organização de engenharia de software. São Paulo: Novatec, 2019.

Gil, A.C. *Métodos e técnicas de pesquisa social*. 5. ed. São Paulo: Atlas, 1999.

Ghani, A.T.A.; Zakaria, M.S. Method for designing scalable microservice-based application systematically: A case study. *International Journal of Advanced Computer Science and Applications*, v. 9, n. 8, 2018. http://dx.doi.org/10.14569/IJACSA.2018.090817.

Kalampokis, E.; Tambouris, E.; Tarabanis, K. On Publishing Linked Open Government Data. *In*: Panhellenic Conference on Informatics, 17th., 2013, New York. *Proceedings* [...]. New York: Association for Computing Machinery, 2013.

Kobashi, N.Y. Análise documentária e representação da informação. *Informare*: *Cadernos do Programa de Pós-Graduação em Ciência da Informação*, v. 2, n. 2, p. 5-27, 1996.

Lakatos, E.M.; Marconi, M.A.*Técnicas de pesquisa*. 5. ed. São Paulo: Atlas, 2002.

Lara, M.L.G. Documento e significação na trajetória epistemológica da Ciência da Informação. *In*: Freitas, L.S.; Marcondes, C.H.; Rodrigues, A.C. *Documento*: gênese e contextos de uso. Niterói: Editora da UFF, 2010. p. 35-56.

Nascimento Silva, P. Recuperação de dados governamentais abertos em APIs: desafios no contexto brasileiro. *BiblioCanto*, v. 9, n. 2, 2023. Disponível em: https://periodicos.ufrn.br/bibliocanto/article/view/34226. Acesso em: 24 set. 2024.

Nascimento Silva, P.; Da Silva, G.V.P. Recuperação de dados abertos em APIs: consumindo dados agregados do IBGE. *Encontros Bibli*, v. 29, e96185, 2023. https://doi.org/10.5007/1518-2924.2024.e96185.

Nikiforova, A. *et al.* Towards high-value datasets determination for data-driven development: A systematic literature review. *In*: Electronic Government: IFIP WG 8.5 International Conference, EGOV 2023, 22nd., 2023, Budapest. *Proceedings* [...]. Berlin, Heidelberg: Springer-Verlag, 2023. https://doi.org/10.1007/978-3-031-41138-0_14.

Olesen-Bagneux, O. *O Catálogo de dados corporativo*: aprimore a descoberta de dados, garanta a governança de dados e impulsione a inovação. São Paulo: Novatec Editora, 2023.

Open Gov Data. *Eight principles of open government data*. Califórnia, 2007. Disponível em: https://public.resource.org/8_principles.html. Acesso em: 10 jan. 2024.

Open Knowledge. *The Open Definition*. London: Open Knowledge, 2024. Disponível em: http://opendefinition.org/. Acesso em: 10 jan. de 2024.

Pedrão, G. B. *A construção do catálogo Panizzi*: uma análise documental. 2019. Tese (Doutorado em Ciência da Informação) - Universidade Estadual Paulista, Marília, 2019.

Saudate, A. *APIs REST*: seus serviços prontos para o mundo real. São Paulo: Casa do Código, 2021.

Silva, P. N. Recuperação de dados em APIs: uma experiência prática no Twitter. *In*: Encontro Nacional de Pesquisa em Ciência da Informação, 22., 2022, Porto Alegre. *Anais eletrônicos* [...]. Porto Alegre: Ancib, 2022. Disponível em: https://ancib.org/enancib/index.php/enancib/xxiienancib/paper/viewFile/1007/728. Acesso em: 11 jan. 2024.

Silva, P. N. Observatório de dados governamentais abertos: acesso às APIs brasileiras. *Revista ACB*, v. 28, n. 1, p. 1-15, 2023. Disponível em: https://revista.acbsc.org.br/racb/article/view/2049. Acesso em: 11 jan. 2024.

United Nations Economic Commission for Europe. *Guidelines on Open Government Data for Citizen Engagement*. New York: United Nations, 2013. Department of Economic and Social Affairs Division for Public Administration and Development. Disponível em: https://aarhusclearinghouse.unece.org/resources/guidelines-open-government-data-citizen-engagement. Acesso em: 10 jan. 2024.

Valente, M. T. *Engenharia de software moderna*: princípios e práticas para desenvolvimento de software com produtividade. [*S. l.*]: Editora Independente, 2020.

Colaboradores

Conceitualização: P. NASCIMENTO SILVA; Curadoria de dados: P. NASCIMENTO SILVA, G. V. P. SILVA e D. R. FERNANDES FILHO. Investigação: P. NASCIMENTO SILVA, G. V. P. SILVA e D. R. FERNANDES FILHO. Administração de projetos: P. NASCIMENTO SILVA. Escrita – revisão & edição: P. NASCIMENTO SILVA.