Um estudo de caso sobre Portais de Dados Abertos

A case study for Open Data Portals

Matheus Biscaya Gutierrez
matheusgutierrez@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil

Nádia Puchalski Kozievitch
nadiap@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil

RESUMO

O Brasil está implementando 11 itens do *Open Government Partnership*, incluindo governo local aberto, dados abertos, ciência aberta, mudanças climáticas e água, transparência legislativa e controle social para a política nutricional. Dados abertos geralmente estão associados ao engajamento cívico, administração responsável e interação com cidadãos. Por outro lado, faltam evidências claras tanto para aqueles responsáveis por coletar, publicar e processar os dados, como para aqueles dispostos a contribuir com novas perspectivas e ideias para o desenvolvimento da cidade. Além disso, usabilidade, acessibilidade e integração ainda são listados como desafios na área. Este artigo tem como objetivo fornecer uma análise exploratória inicial de portais de dados abertos brasileiros e suas características, metodologias, frameworks e desafios.

PALAVRAS-CHAVE: Portais de Dados Abertos. Dados Espaciais. Análise Exploratória.

ABSTRACT

Brazil is working towards implementing 11 points of the *Open Government Partnership*, which includes local government transparency, open data, open science, climate change and water resources, legislative transparency and community oversight of nutritional policies. The concept of open data is commonly associated with civic engagement, responsible administration and citizenry integration. On the other hand, there is a lack of clear evidence for both those who are responsible for collecting, publishing and processing data, and those who are willing to contribute with new ideas and perspectives on how to further develop their cities. This paper aims to provide an exploratory analysis on existing brazilian open data portals and examine their characteristics, methods, frameworks and challenges.

KEYWORDS: Open Data Portals. Spatial Data. Exploratory Analysis.

INTRODUÇÃO

Apesar das mudanças resultantes das transições políticas, o Brasil continua comprometido com os esforços do governo aberto [Steibel 2020]. Um *Open Government Partnership* (OGP) é uma parceria global que reúne governo e líderes da sociedade para criar planos de ação que tornem os governos mais inclusivos, responsivos e responsáveis, com foco em temas como um ecossistema de dados abertos, ciência aberta, controle social e feedback do cidadão, transparência em reparos de desastres ambientais, legislatura aberta, transparência fundiária, clima e recursos hídricos e liberdade de informação. Na mesma direção o plano de ação da Infraestrutura Nacional de Dados Abertos (INDA) para o biênio 2021-2022 [CGINDA 2021] já prevê iniciativas na direção de realizar pesquisas sobre abertura e reuso de dados, eventos de cocriação, revisão das bases, aprimorar o framework de dados abertos do governo federal, orientação e capacitação do uso dos dados, e catálogo de bases de dados federais, entre outros.

Ainda nesta direção, do ponto de vista espacial, outras iniciativas surgem, como o Plano de Dados Abertos do IBGE 2022 [IBGE 2020], que já inclui a área de mapas no Portal, o Banco de Metadados Estatísticos do IBGE, o Catálogo de Metadados Geoespaciais, o Banco de Tabelas Estatísticas do SIDRA e a inclusão de canais alternativos de comunicação, como o Nomes (nomes da população recenseada pelo Censo Demográfico de 2010).

Cidades como Paris, Nova Iorque e Moscou têm buscado disponibilizar os dados da cidade através de datasets em portais de dados abertos (ODP). No Brasil, as cidades de Porto Alegre, São Paulo, Natal, Recife, Alagoas e Distrito Federal têm disponibilizado dados de diferentes temas em seus ODPs. Estas cidades têm explorado diferentes temas (como cultura, mobilidade e meio-ambiente), diferentes tipos de buscar (como índices, filtros), diferentes tipos de disponibilização (APIs, arquivos), diferentes formatos (CSVs, JSON, EXCEL), diferentes formatos de GIS (SHAPEFILES, GEOJSON, KML) juntamente com visualização online (Tableau, GIS maps e charts) e metadados. Percebe-se que as soluções de implementação de dados abertos variam de cidade para cidade, mesmo se forem apenas examinadas cidades brasileiras. Este fato chamou a atenção de vários países para a necessidade de padronização, e um exemplo resultante é a ISO/DIS 37110 de dados abertos para cidades inteligentes [International Standard Organization 2020].

Embora a quantidade de dados abertos disponíveis online tenha aumentado, existem desafios que necessitam ser superados para que seja possível utilizá-los de forma plena. Existem problemas como: a inconsistência de disponibilidade de certos dados ou componentes espaciais, a diferença de formatos e nível de detalhes dos dados, baixa qualidade durante a integração dos dados em diferentes formatos, integração de dados georreferenciados em diferentes sistemas de referência e a integração de dados por diferentes provedores, contendo dados sobre o mesmo assunto, porém em diferentes níveis de detalhe ou precisão [Costa et al. 2017],[Simette et al. 2018],[Kozievitch et al. 2018]. Ainda é possível mencionar a falta de atualizações dos dados e a falta de integração automatizada.

Este artigo tem por objetivo realizar a análise exploratória para compreender os aspectos atuais da utilização de portais de dados abertos brasileiros, suas iniciativas e os desafios encontrados.

MATERIAIS E MÉTODOS

A Análise Exploratória de Dados (AED) é uma abordagem para navegar vários aspectos dos dados disponíveis com o objetivo de identificar o conhecimento latente, relações, proporções ou estruturas de dados [NIST/SEMATECH 2012], com o objetivo de sumarizar características. A abordagem não inferencial na análise de dados incentiva a perspectiva de abertura necessária para integrar novos domínios.

As Tabelas 1 e 2 ilustram características dos portais de dados abertos de seis cidades brasileiras. Apesar do INDA [CGINDA 2021] ter como objetivo um catálogo de bases de dados federais, ainda não há uma fonte única que liste estes ODPs nacionais. Note que elas diferem não só no tipo de interface, como nos temas adotados (*datasets*), tipos de dados disponibilizados, e tipo de acesso. Além disso, a maioria não possui integração nem a nível municipal (tendo, por exemplo, dados espaciais em portais diferentes e não integrados).

Tabela 1 – Características de portais de dados abertos brasileiros.

Capital	Interface	Datasets	Tipos	Acesso	Dashboard
Porto Alegre	CKAN	50	csv, pdf	Arquivo, API	Sim
São Paulo	CKAN	195	csv, xlsx, xls, pdf	Arquivo, API	Não
Belo Horizonte	CKAN	86	csv, pdf, json	Arquivo, API	Não
Natal	CKAN	18	csv, pdf, kml	Arquivo, API	Sim
Curitiba	Própria	26	csv, xlsx, pdf	Arquivo, Rsync	Não
Alagoas	CKAN	268	pdf, csv, xlsx, geojson, kml	Arquivo, API	Sim

Fonte: Autoria própria (2021).

Tabela 2 – Características de portais de dados abertos brasileiros.

Capital	Dicionários	Flow de atividade	Pré Visualização	Visualização em mapa	Múltiplos Formatos
Porto Alegre	Alguns	Sim	Sim	Sim	Alguns
São Paulo	Alguns	Sim	Sim	Sim	Alguns
Belo Horizonte	Alguns	Sim	Sim	Sim	Alguns
Natal	Alguns	Sim	Sim	Sim	Alguns
Curitiba	Alguns	Limitado	Não	Não	Não
Alagoas	Alguns	Sim	Sim	Sim	Sim

Fonte: Autoria própria (2021).

O portal de dados abertos do estado de Alagoas disponibiliza datasets de temas, inclusive de assuntos atuais extraordinários, como os dados da pandemia de covid-19 no estado. Os 268 datasets ofertados estão disponíveis em várias opções de formato. O portal disponibiliza vários *dashboards* que exibem informações relacionadas a atual pandemia, as contas do estado, a situação dos empregos e a situação dos órgão e serviços presentes no estado, como mostra a Figura 1. O portal de Alagoas foi construído utilizando a plataforma CKAN e dispõe de um controle de histórico de atividades, que monitora quais arquivos foram modificados e por quem.

O portal de dados do município de Natal disponibiliza datasets relacionados apenas a mobilidade urbana, mantendo também informações relacionadas a demanda pelos serviços de transporte público durante a pandemia de covid-19. Os dados da mobilidade urbana estão disponíveis em formato KML, CSV e PDF. O portal utiliza a ferramenta Power BI para exibir as informações da atividade do transporte público. Assim como o portal do estado de Alagoas, o portal do município de Natal utiliza da plataforma CKAN e possui controle de histórico de atividades.

RESULTADOS E DISCUSSÕES

Embora várias cidades disponibilizem seus próprios ODPs, a maioria não observa os princípios de dados abertos, fazendo uso de formatos proprietários ou não estruturados (como PDFs), negligenciando o licenciamento e a exigência de credenciais (normalmente necessários apenas para dados dinâmicos). Outros problemas encontrados são a segregação de dados em diferentes sítios (pertencentes à união, ao estado, ou a municípios), e a duplicação de dados.

Levando em consideração apenas dados relacionados à mobilidade urbana, como mencionado em [Vila et al. 2016], existem diversos problemas, como a utilização de diferentes sistemas de coordenadas e formatos de arquivos. Formatos CSV, Excel, Json e Shapefiles (em alguns casos, APIs e formato KML) são preferidos em ODPs. É recomendado prover metadados, visualização de dados, ou atualizações que lidem com as mudanças que possam vir a ocorrer na formatação dos dados (juntamente com dados históricos). Ao lidar com dados espaciais, surgem novas exigências, como opções de projeção (usuários devem ser capazes de poder escolher a projeção dos dados), diferentes formatos (usuários podem preferir obter os dados em formato KML, Shapefile, ou através de API - JSON, Geoserviços, WMS), filtro e busca de dados baseado em localização geográfica, visualização de dados pelo navegador, entre outros.

Em paralelo, a chave para encontrar e compreender dados é o metadado. Neste caso, um conjunto de elementos de metadados deve ser associado a cada ativo para que eles possam ser usados de acordo com os princípios FAIR[1] (os dados devem ser Localizáveis, Acessíveis, Interoperáveis e Reutilizáveis - do inglês ***Findable, Accessible, Interoperable and Reusable***). Atualmente, existem quatro padrões de metadados genéricos que são amplamente usados, *Dublin Core* (DC[2]), Vocabulário do Catálogo de Dados (DCAT[3]), *DataCite*[4] e *Schema.org*[5]. Para um registro mais longo de padrões de metadados, há ainda o Catálogo de Padrões de Metadados[6] ou o *FAIRsharing*, endossado por *Research Data Alliance*[7]. Em paralelo, os metadados de proveniência (em conjunto com os metadados históricos) são a base para avaliar a qualidade e confiabilidade dos dados abertos. Eles (em conjunto com ontologia de proveniência) impactam diretamente problemas de semântica e baixo nível de interoperabilidade. Atualmente, os portais OGD fornecem proveniência por meio de elementos de metadados gerais, como criador, provedor, data de criação, data de publicação e tempo emitido.

Dados antigos podem ser problemáticos para tentar tomar decisões informadas. Portanto, é importante que os dados em um portal de dados abertos sejam atualizados regularmente. Uma opção é usar um feed publicado (por exemplo, RSS) ou API, em vez de arquivos estáticos para download. Além de permitir que usuários utilizem o dado mais atual, as possíveis atualizações já são refletidas no usuário final. Uma outra vantagem desta opção é a não duplicação de dados em dois locais, evitando problemas quando mais atualizações forem necessárias no futuro. No caso do portal utilizar arquivos, a manutenção dos links dos mesmos deve se manter atualizada (tanto o arquivo atual quanto o histórico).

A falta de padronização se torna um problema para nomes de rua. Um exemplo são os dados abertos da Receita Federal, que possui endereços cujo o nome acaba sendo registrado de diferentes formas. Este

1 https://www.nature.com/articles/sdata201618
2 https://dublincore.org/specifications/dublin-core/
3 https://www.w3.org/TR/vocab-dcat-2/
4 https://datacite.org/
5 https://schema.org/
6 https://rdamsc.bath.ac.uk/
7 https://www.rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases.html

problema é causado por erros de digitação, o uso de abreviações ou pelo erro no momento de identificar o tipo da via. Se o dado for cruzado com os dados abertos de Curitiba, surgem novos problemas, como nomes diferentes se referindo a uma mesma empresa. Soluções para esta situação incluem a utilização de processos de *entity matching*, e a utilização de informações de CEP.

Os diferentes vocabulários utilizados pelos dados (juntamente com abreviações e sinônimos) podem utilizar estratégias de Web Semântica [Berners-Lee et al. 2001] para otimizar a integração. Outra possível solução para integração dos dados com outras informações (como dados não estruturados) é a inclusão de tags em arquivos contendo informações relevantes. Um exemplo da utilização de tags é a inclusão de informações temporais no formato ”doc: published time”, obedecendo o protocolo *Open Graph*. Também é possível utilizar o vocabulário controlado proposto pelo Governo Federal através do projeto VCGE[8], ou o metadado padrão proposto pelo Governo Eletrônico através do e-PMG[9].

Como mencionado em [Fonseca et al. 2020], crises como a pandemia de covid-19 tem indicada a importância de metadados, pesquisa aberta (para que dados abertos, ciência aberta, a modernização de dados científicos, e os ciclos de vida de dados linkados possam contribuir para o descobrimento científico rápido e confiável), considerações éticas e de privacidade, repositórios confiáveis, e investimento de infraestrutura. Seguindo nesta direção, existem projetos como o *Research Data Alliance* (https://www.rd-alliance.org/), o *Open Data Institute* (https://theodi.org/) e até recomendações (como o texto inicial da UNESCO em *Open Science*[10]), que vem providenciando instruções e recomendações para aqueles que desejam publicar dados abertos.

CONCLUSÃO

Este artigo apresentou uma análise exploratória preliminar com o objetivo de identificar as situações e implicações que surgiram devido a disponibilização de portais de dados abertos. Além disso, este artigo também demonstra as discussões relacionadas ao tema. As características encontradas podem ser utilizadas em análises futuras para que seja possível otimizar o sistema de informação e melhorar a sua integração. Baseando-se nesta investigação inicial, é possível perceber que os portais de dados ainda estão aprendendo como integrar as diferentes fontes de dados e como explorá-los.

Este trabalho pode ser estendido para analisar e verificar características de outros portais brasileiros, incluindo otimizações para inclusão de ferramentas, integração com outros setores e maior participação do cidadão.

AGRADECIMENTOS

Os autores agradecem a Universidade Tecnológica Federal do Paraná (DIREC 01/2020), ao Instituto de Pesquisa e Planejamento Urbano de Curitiba (IPPUC) e à Prefeitura de Curitiba por compartilhar parte dos dados utilizados neste estudo.

8 https://www.gov.br/governodigital/pt-br/governanca-de-dados/vcge_2_1_0.pdf
9 https://www.gov.br/governodigital/pt-br/governanca-de-dados/PMGVersao1_1.pdf
10 https://unesdoc.unesco.org/ark:/48223/pf0000378381.locale=en

REFERÊNCIAS

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. **The Semantic Web**, Scientific American. v. 284, n. 5, p. 34-43.

CGINDA. **Infraestrutura Nacional de Dados Abertos - INDA Plano de ação 2021-2022.** Comitê Gestor da Infraestrutura Nacional de Dados Abertos. 2021. (https://dados.gov.br/wp/wp-content/uploads/2021/06/PLANO-DE-AC%CC%A7A%CC%83O- INDA-versaorevisada-jun-2021.pdf)

COSTA, G.; KOZIEVITCH, Nádia Puchalski; FONSECA, Keiko; GADDA, Tatiana; BERARDI, Rita. **Integração de dados de redutores de velocidade no transporte público de curitiba**, XIII Escola Reg. de Banco de Dados. p. 123-126. 2017.

FONSECA, Keiko V. O.; KOZIEVITCH, Nádia P.; BERARDI, Rita C. G.; SCHMEISKE, Oscar. **Information Technology Macro Trends Impacts on Cities: Guidelines for Urban Planners**, Springer International Publishing. p. 1-24. 2020.

IBGE. Planos de Dados Abertos Julho/2020 - Julho/2022. 2020. (https://www.ibge.gov.br/np_download/novoportal/documentos_institucionais/Plano_de_Dados_Abertos_IBGE_2020_2022_1arevisao.pdf)

INTERNATIONAL STANDARD ORGANIZATION. **International Standard Organization 37110 - Management guidelines of open data for smart cities and communities: Overview and general principles**. 2020.

KOZIEVITCH, Nádia; PARCIANELLO, Yussef; FONSECA, Keiko; ROSA, Marcelo; GADDA, Tatiana; MALUCELLI, Francisco. **Transportation: An overview from Open Data Approach**, 4th IEEE Intern. Smart Cities Conf. p. 1-8. 2018.

NIST/SEMATECH. **E-Handbook of Statistical Methods.** 2012. (http://www.itl.nist.gov/div898/handbook/)

SIMETTE, Gabriely; PARCIANELLO, Yussef; KOZIEVITCH, Nádia P.; FONSECA, Keiko Veronica Ono. **Análise da situação dos redutores de velocidade de Curitiba**, Anais da XIV Escola Regional de Banco de Dados. Porto Alegre. p. 123-126. 2018.

Steibel, Fabio. **Brazil Design Report 2018-2020**. Open Government Partnership. 2020. (https://www.opengovpartnership.org/documents/brazil-design-report-2018-2020/)

VILA, Juan; KOZIEVITCH, Nádia; FONSECA, Keiko; GADDA, Tatiana; ROSA, Marcelo; GOMES-JR; Luiz Celso; AKBAR, Monika. **Urban Mobility Challenges – An Exploratory Analysis of Public Transportation Data in Curitiba**, Revista de Informática Aplicada. v. 12. 2016.