Open Data Portals - A case study of Challenges and Opportunities

Maiara G. Flausino[1], Nádia P. Kozievitch[1], Keiko V. O. Fonseca[1], Eunice Liu[1]

[1]Department of Informatics Federal University of Technology Curitiba, Brazil

{maiaraflausino,nadiap,keiko,euniceliu}@utfpr.edu.br

Abstract. *Open data portals are one of the largest cluster of open data initiatives, and their features and impact encompasses a range of topics (including culture, mobility, and the environment), employing various search methods (such as indexes and filters), utilizing diverse forms of accessibility (via APIs and files), along with heterogeneous data (spatial, structured, unstructured, proprietary, vectorized). This paper evaluates two Brazilian open data portals, using as criteria data access, dataset characteristics, metadata and dataset schema.*

Resumo. *Os portais de dados abertos são um dos maiores grupos de iniciativas de dados abertos, e suas características e impacto abrangem uma variedade de tópicos (incluindo cultura, mobilidade e meio ambiente), empregando vários métodos de pesquisa (como índices e filtros), utilizando diversas formas de acessibilidade (via APIs e arquivos), juntamente com dados heterogêneos (espaciais, estruturados, não estruturados, proprietários, vetorizados). Este artigo avalia dois portais de dados abertos Brasileiros, usando como critério acesso a dados, características do dataset, metadados e esquema do dataset.*

1. Introduction

The principle of knowledge commons (conception of open data) was initially theorized by Robert King Merton in 1942[1], highlighting the benefits of accessible scientific data. It was only in 1995, through a document from a United States scientific agency, that the term "open data" was formally introduced, addressing the dissemination of geophysical and environmental information (Chignard, 2013). The global trend of making data accessible, especially at the government level, resulted in the emergence of the movement known as "open data" (Murray-Rust, 2008; Sayão and Sales, 2014), which later expanded to other areas, including the availability of data research and scientific production.

Within the global south, where the open data infrastructure does not satisfy existing definitions and principles created by open government experts[2], open data portals are the third largest cluster of open data initiatives, together with legislatures, elections and elected officials, and budget and spending information. Generally these open data portals (ODP) are from government, CSO and academic institutions (Chan and Kin-sing, 2015). The investigation of ODP features and their impact encompasses a range of topics (including culture, mobility, and the environment), employing various search methods (such as indexes and filters), utilizing diverse forms of accessibility (via APIs and files), adopting different file formats (CSVs, JSON), and incorporating a

[1] https://www.paristechreview.com/2013/03/29/brief-history-open-data/

[2] https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

variety of Geographic Information System (GIS) file formats (such as SHAPEFILES). Additionally, they leverage online visualization tools (such as Tableau, GIS maps, and charts) along with their respective metadata.

In this direction, this work presents an evaluation of two Brazilian data portals, using as criteria data access, dataset characteristics, metadata and dataset schema (methodology previously presented in 2016 (Oliveira et al. 2016)). The text of this work is organized as follows: Section 2 presents the related work, and Section 3 presents the data analysis. Section 4 presents the conclusions of this work, along with future work.

2. Related Work

According to the Open Knowledge Foundation[3], data is considered accessible when it is readily available to be used, reused and shared by anyone, without excessive impositions. In general, this involves disclosing the data in a non-restrictive format and under the terms of a license that, at most, requires acknowledgment of the original source and sharing under equivalent conditions. In this sense, according to Tim Berners Lee's proposal[4], the data is categorized within 5 stars: (1) 1 Star - Open License (OP): Data can be found on the internet under an open license; (2) 2 Stars - Readable (RE): The data is available in a structured form; (3) 3 Stars - Open Format (OF): Data is available on in a non-proprietary format; (4) 4 Stars - Uniform Resource Name (URI): The data complies with standards established by the World Wide Web Consortium (W3C); and (5) 5 Stars - Linked Data (LD): Data is linked to other data to provide broader context.

Many cities are adopting measures to make open data available. The challenge lies in how to extract meaning from such a vast amount of data, considering that this data is intricate and may involve geographical and temporal elements (Ferreira et al., 2013). When it comes to correlating multiple data sources, ensuring integrity and consistency is crucial. Maintaining coherence between these diverse sources of information is a significant challenge, particularly regarding discrepancies in the vocabulary used in the data, the absence of common identifiers in different datasets, lack of information, and other factors. According to Beluzo (2015), the following challenges were identified: data inconsistencies, which affect the integrity and quality of information. Problems related to the incompatibility between datasets from different sources were also observed, pertaining to integration, and the lack of adequate documentation of datasets, reflecting challenges in metadata management.

Moreover, there are various restrictions when it comes to leveraging open data (Janssen et al., 2012), including issues related to technology, metadata, and standardization, or establishing data standards, and closer collaboration among the various stakeholders involved (Przeybilovicz et al., 2017). In this sense, Kono (2016) presents a synthesis of some challenges associated with open data, such as data integrity and accuracy, geographic mapping, and integration of information. But in parallel, there are several open initiatives, such as The Research Data Alliance[5], advocating for open

[3] https://okfn.org/en/

[4] https://5stardata.info/pt-BR/

[5] https://www.rd-alliance.org/

data sharing, and encouraging the use of open data for various purposes, such as transparency, innovation, research, and informed decision-making. The European Commission[6], for example, consider 4 indicators for open data maturity: (1) the level of development of national policies promoting open data, (2) the features and data made available on national data portals, (3) the quality of the metadata on national data portals, and (4) initiatives to monitor the reuse and impact of open data. A google scholar search on July, 26th, 2024 ordered by date with the keywords "open data portal" at abstracts[7] brought 66 results some of them referring to municipalities or government data portals, both subjects of interest of our work. Table 1 highlights some examples extracted of the search results that includes ODP of municipalities and/or government as main subject of study.

Table 1. Examples of ODPs

Document title	Focus	Link
Open data portal for smaller municipality	simple functional application for managing and visualizing open data based on Czech catalog of Open Data	https://is.muni.cz/th/fqn3z/DP_469406_Archive.pdf
Analysis of the Open Data Landscape in Mexico	challenges and identification of repositories and characteristics	https://sedici.unlp.edu.ar/handle/10915/166294
Exploring Estonia's Open Government Data Development as a Journey towards Excellence: Unveiling the Progress of Local Governments in Open Data Provision	explore the barriers and enablers of municipal OGD.	https://dl.acm.org/doi/abs/10.1145/3657054.3657161
Assessing the readability of open data portals: a case study of Open Data Pakistan	readability in open data publication policies and guidelines	http://jice.um.edu.my/index.php/MJLIS/article/view/48035

Other data portals have been studied by the literature, such as the city of Rzeszów - Poland (Duncan et al., 2020), Trikala - Greece (Moustaka et al., 2021), Häme - Finland (Jussila et al., 2019), London - England (Gupta et al., 2020), Taiwan (Shibuya et al., 2022), Norway (Ibrahim, 2022), Germany (Lämmel et al., 2020) and China (Chen, 2019). Brazilian ODPs (Curitiba and Natal) were previously analyzed in 2022 (Kozievitch et al. 2022), using, by instance, the same methodology (Oliveira et al. 2016). Several challenges were listed, such as the different formats, lack of Open Data principles, proprietary or non-structured data format (such as PDF), along with problems of open license specification, standards in metadata, vocabularies, documentation about changes and data over time, and feedback from customers.

3. Data Analysis: Curitiba and Natal

Brazil is one of the leaders of the Open Data initiative in South America[8], and one of the founders of the Open Government Partnership[9]. Nevertheless, a structured

[6] https://data.europa.eu/en/publications/open-data-maturity/2023

[7] https://scholar.google.com/scholar?start=0&q=%22open+data+portal%22&hl=pt-BR&scisbd=1&as_sdt=0,5

[8] https://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-LatinAmericaReport.pdf

[9] https://www.opengovpartnership.org/

methodology is still missing for ODPs evaluation[10]. Oliveira et al. 2016, for example, considers the following items: Data Access, Dataset Characteristics, Metadata and Dataset Schema. The CKAN platform is predominantly employed by the majority of Brazilian (ODPs), following the guidelines of the CKAN Domain Model[11]. Essentially, this model encompasses a dataset, core metadata, limitless additional metadata, relationships, resources, revisions (recording all changes), and task statuses.

This section evaluates Curitiba and Natal ODPs, using Oliveira et. al 2016 criterias. The data was manually extracted from both portals, within the first semester of 2024.

Curitiba has been following the open data movement, providing data on urban mobility through its city hall[12], Instituto de Pesquisa e Planejamento Urbano[13] and Urbanização de Curitiba[14]. Open Data in Curitiba resulted from a decree[15] and its ODP was established in 2014.

Regarding access, Curitiba ODP has a proprietary platform[16], with no API available, and within the first semester of 2014, had 15 different data themes to be downloaded, along with an average of 55 different files. Regarding dataset characteristics, Curitiba ODP has a quantity of 9109 files (8891 being CSV), with an average of 153 GB base (with 987 MB being CSV), having the most outdated data from 2017 and the oldest data from 1934. From all these files, in particular, 12 CSV ones were analyzed. Regarding license types, Curitiba ODP uses the Creative Commons Attribution (everyone from individual creators to large institutions a standardized way to grant the public permission to use their creative work under copyright law). The average distribution of data formats has the following: 1) 99,24% as CSV; 2)0,41% as XLSX; 3) 0,12% as XML, 4) 0,04% as XLS; 5) 0,02% as PDF, and 6) 0,54% as other formats. Brazilian portals present a different history of dataset updates, but in particular, Curitiba ODP has an average of a three months update period. In the Brazilian context, there is no formal standard to describe metadata for ODPs. However, if the portal uses the CKAN platform, probably they will use the CKAN Domain Model[17]. This is not available in Curitiba, since it uses a proprietary platform.

Regarding access, Natal ODP[18] uses CKAN platform, with an API available, with 19 different data themes to be downloaded. Regarding dataset characteristics, Natal ODP has a quantity of 256 files (83 being CSV), with an average of 108 MB base (with 389 Kb being CSV), having the most outdated data from 2018 and the oldest data from 1958. From all these files, in particular, 8 CSV ones were analyzed. Regarding license types, Natal ODP uses the Creative Commons Attribution (everyone from individual creators to large institutions a standardized way to grant the public permission to use their creative work under copyright law). The average distribution of data formats has

[10] https://www.oecd-ilibrary.org/open-government-data_5k46bj4f03s7.pdf

[11] http://docs.ckan.org/en/ckan-1.8/domain-model.html

[12] https://www.curitiba.pr.gov.br/dadosabertos/

[13] ippuc.org.br

[14] http://urbs.curitiba.pr.gov.br/

[15] http://multimidia.curitiba.pr.gov.br/2014/00147194.pdf

[16] https://www.ici.curitiba.org.br/

[17] http://docs.ckan.org/en/ckan-1.8/domain-model.html

[18] http://dados.natal.br/

the following: 1) 33% as CSV; 2)24% as PDF; 3) 0% as XML, 4) 0% as XLS; 5) 43% as other formats.

Table 2. Brazilian legal framework for Open Data

Scope	Name	Description
Federal (Brazil)	General Data Protection Law[19]	Processing of personal data in digital media by individuals or public/private organizations
Federal (Brazil)	Federal Executive Branch Open Data Policy[20]	Establishes the way in which the federal government will make its data available to ensure public access
Federal (Brazil)	Access to Information Law[21]	Regulates citizens' constitutional right to access public information
Federal (Brazil)	Internet Civil Rights Framework[22]	Defines principles, guarantees and responsibilities for using the internet in Brazil
Federal (Brazil)	Digital Governance Strategy 2020 to 2023[23]	Provision of higher quality public policies and services, accessible at any time and place, at a lower cost to citizens
Federal (Brazil)	Steering Committee of the National Open Data Infrastructure[24]	Committee led by the CGU, composed of 9 public entities
Federal (Brazil)	Legal Framework for Science[25]	Stimulates scientific development and innovation
Federal (Brazil)	Biometric Data Law[26]	Manages science, research, technological training and innovation
Municipal (Curitiba)	Municipal Policy Decree[27]	Regulates the procedure for access to public information, classification and reclassification of confidential information.

The next analysis covers dataset schema. Random CSV files were chosen to this particular analysis. Considering the Syntax and Structural Rules Adherence, most CKAN-based portal datasets have a proper header (consisting of just one line). Curitiba ODP, in some cases, has a second line below the header (which ends up serving as an obstacle when analyzing the dataset), along with lacking headers. But Curitiba's case, where 100% of the headers of the ODP had all the columns declared within the metadata (all the files have a data dictionary). The duplicate data is low (7,81%). Regarding the field name convention, the CSV had 509 words in Portuguese language, 14 words in English, 14 words in both languages (Portuguese and English), and 245 words with no meaning as headers.

Natal ODP has the metadata (data dictionary) partially present for all the data themes (only the column data type), and all the files has just one line as header. The

[19] https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm
[20] https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm
[21] https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm
[22] https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm
[23] https://www.gov.br/governodigital/pt-br/EGD2020
[24] https://www.gov.br/participamaisbrasil/comite-gestor-da-infraestrutura-de-dados-abertos-
[25] https://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2016/Lei/L13243.htm
[26] https://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2016/Lei/L13243.htm
[27] http://mid-transito.curitiba.pr.gov.br/2018/8/pdf/00000038.pdf

duplicate data is lower than Curitiba (1,44%). Regarding the field name convention, the CSV had 143 words in Portuguese language, 0 words in English, 0 words in both languages (Portuguese and English), and 18 words with no meaning as headers.

In comparison, the Brazilian ODPs present a substantial presence of words found in both English and Portuguese dictionaries, as well as a notable number of words lacking a discernible meaning (Oliveira et. al 2016).

Within the schemata size (number of fields in a header), there is no standard for Curitiba: the smallest one has four columns, and the biggest one has 213 columns, with an average size of 29,75 for a header. Considering the data sparseness, the number if null cells for Curitiba ODP is also low: 8, 75%. Natal ODP has the smallest schemata with 6 columns, and the biggest with 14 columns, with an average size of 11,87 for a header. Considering the data sparseness, the number if null cells for Natal ODP is also low: 1%.

In addition to Oliveira et. al 2016 parameters, this paper also considered additional characteristics: availability of dashboards, data preview, map view, related laws, availability in English, and accessibility. Considering Curitiba ODP, there are no availability of data visualization (dashboards, data preview, map view), or related laws, additional languages or items for accessibility. Natal ODP, at the other hand, allows the user to pre-visualize the data, or to use an API to download them. Map visualization is also available for the KML files. Natal ODP has the overall translation to several languages, but the dataset names, content, and data dictionary are still in english. Laws or items for accessibility are also not available.

In Curitiba, the spatial files can be found in another location (IPPUC - https://geocuritiba.ippuc.org.br/portal/apps/sites/#/geocuritiba), with maps, cartographic base, 3D app, historic maps, images, along with panels for accidents, urban equipments, and city structure for bicycles. Different layers can be added, printed, and files can be downloaded.The geodownload site has 56 different data subjects containing the shapefiles with the coordinate reference system SAD 69 UTM 22S (29192) for old files and SIRGAS 2000 (31982) to the newest ones. Spatial files from Natal are available at Semurb[28], listing 38 files. But visualization, coordinate reference system are not available.

In summary, compared to the previous analysis (which used the same methodology) in 2022 (Kozievitch et al. 2022), few has changed to 2024, within both open data portals. Natal, according to the ODI index[29] has position 15th, and has level four, while Curitiba is 4th within the ranking[30], Several recommendations previously listed in 2022 are still valid for 2024: 1) different vocabularies available within the data data (along with abbreviations and synonyms) could use Semantic Web strategies or meta tag in files; 2) linked data could be used to integrate raw data to pdf files, along with indicators and standard formats; 3) importance of metadata, open research, ethical and privacy considerations, and trustworthy data repositories; and 4) other questions arise, considering spatial data, such as projection options (users should be able to choose their projection), different formats (download data as KML, shapefile or via the API - SON, Geoservice, WMS), filter and search for data based on a geographic

[28] https://www.natal.rn.gov.br/semurb/geoinformacoes

[29] https://indicedadosabertos.ok.org.br/capital/natal-rn

[30] https://indicedadosabertos.ok.org.br/capital/curitiba-pr

location, data, and infrastructure investment.

Considering the legal framework, there are communities like the Open Data Group of Curitiba[31], which study and present guidelines, such as the study of legislation for open data, as illustrated in Table 2. Moreover, the data itself may face new challenges in its provision, such as transparency by design in open data portals. There are still other challenges, such as the interface design of ODPs, to facilitate access to information by different types of users (Liu et al. 2022). Functional and visual components such as search bar, horizontal navigation menu, category menu, summary graphs with data preview in different formats, icons and buttons for downloading files, sharing, search filters, maps, publication history among other design components, can improve the user experience when browsing and accessing data (Souza et al. 2023). Explanatory videos and referrals to external links with original sources of information also help data access. Analysis of the design interface of the ODPs listed the following fundamental summary information that could be presented in a synthesis card: category, title, responsible entity, open data summary, available file types, last update and number of accesses of the open data sets, graphs and feedback area. Accessibility qualifications such as font size, adequate contrast, responsiveness, use of alternative text, audio description and web standards are essential to be present.

4. Conclusion

The governance of disclosed data plays a vital role in supporting established services and seeking more sustainable solutions to address challenges such as urban population displacement, climate change resilience, and aging society. However, the use of open data and related portals are still not widely disseminated within the community. This paper presented a preliminary investigation conducted to identify scenarios and implications on two Brazilian open data portals, using as criteria data access, dataset characteristics, metadata and dataset schema. The paper also evaluated availability of dashboards, data preview, map view, related laws, availability in English, and accessibility. Compared to a previous analysis from 2022, few has changed in both open data portals. As future work, one can mention the expansion of the methodological approach, the computational analysis of the interface design integration, along with an analysis of other ODPs (national and international).

Acknowledgments

The authors thank the Curitiba City Hall, IPPUC and UTFPR.

References

Beluzo, J. R., Ideo: Integrador de dados da execução orçamentária brasileira. 2015. 126 f. Dissertação (Mestrado em Ciências) - USP, São Paulo, 2015.

Chan J. K. , Kin-sing, J. “The Social Impact of Open Data,” 2015. (Online). Available: http://www.opendataresearch.org/dl/symposium2015/odrs2015-paper20.pdf

Chen, X. The Development Trend And Practical Innovation Of Smart Cities Under The Integration Of New Technologies. FEM, V. 6, P. 485-502, 2019.

Chignard, S. A brief history of open data. 2013. available at https://www.paristechreview.com/2013/03/29/brief-history-open-data/

[31] https://www.youtube.com/watch?app=desktop&v=MtJOu9BZub8

Duncan,G., Dymora, P., Koczkodaj, W. W., Kowal, B., Mazurek, M. and Strzałka,D. "Open Government issues and opportunity: a case study based on a medium-sized city in Poland," 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), London, UK, 2020, pp. 563-568.

Ferreira, N. et al. Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. IEEE Trans. on Visua. and Computer Graphics, v. 19, n. 12, p. 2149–2158, dez. 2013.

Gupta, A. et al. An Orchestration Approach To Smart City Data Ecosystems. Technological Forecasting And Social Change, V. 153, P. 119929, 2020.

Hossain, M. A. et al. Data-Driven Innovation Development: An Empirical Analysis Of The Antecedents Using Pls-Sem And Fsqca. Annals Of Operations Research, 2022.

Ibrahim, A. M. Municipal Platforms: An Investigative Case Study From A Norwegian Municipality. In: Icegov '21. New York: Association For Computing Machinery, P. 275-284, 2022.

Janssen, M. et al. Benefits, adoption barriers and myths of open data and open government. Information Systems Management, v. 29, p. 258–268, 2012.

Jussila, J. et al. Open Data And Open Source Enabling Smart City Development: A Case Study In Häme Region. TIM Review, V. 9, P. 25-34, 2019.

Kono, F. A. M. Um modelo de representação computacional baseado em conceitos de crescimento urbano associados a alvarás e primitivas em banco de dados espacial. 2016. 143 f. Mestrado em Computação Aplicada - UTFPR, Paraná, 2016.

Kozievitch, N. P., et al. Assessment Of Open Data Portals: A Brazilian Case Study. In: 2022 Ieee International Smart Cities Conference (Isc2), P. 1-7, Pafos, Cyprus, 2022.

Lämmel, P. et al. Metadata Harvesting And Quality Assurance Within Open Urban Platforms. Journal Of Data And Information Quality, V. 12, N. 4, Article 22, 20 Pages, December 2020.

Liu, E., et al. Acessibilidade digital do canal de governança pública 'Central 156' de Curitiba, Brasil. In: Congresso Brasileiro de Pesquisa e Desenvolvimento em Design, P&D Design 2022, 2022, São Paulo. Anais do Congresso Brasileiro de Pesquisa e Desenvolvimento em Design P&D Design, 2022. p. 1-18.

Moustaka, V. et al. Urban Data Dynamics: A Systematic Benchmarking Framework To Integrate Crowdsourcing And Smart Cities' Standardization. Sustainability, V. 13, P. 8553, 2021.

Murray-Rust, P. Open data in science. Serials Review, v. 34, n. 1, p. 52-64, 2008.

Oliveira, M. I. S., et al. Open Government Data Portal Analysis: The Brazilian Case. In: dg.o ’16. New York, NY, USA: ACM, 2016, pp. 415–424

Przeybilovicz, E. et al.. Identifying Essential Organizational Characteristics for Smart Urban Governance. In: dg.o '17. ACM, New York, NY, USA, 2017. p. 416–425.

Sayão, L. F.; Sales, L. F. Dados abertos de pesquisa: ampliando o conceito de acesso livre. Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, v. 8, n. 2, p. 76-92, jun. 2014.

Shibuya, Y. et al. Do Open Data Impact Citizens’ Behavior? Assessing Face Mask Panic Buying Behaviors During The Covid-19 Pandemic. Scientific Reports, V. 12, 2022.

Sousa, F. K. ; et al . Portales de Datos Abiertos: un Análisis de Re-Diseño. In: Congreso Internacional de Investigación en Contabilidad y Empresa, 2023, Barcelona. 2023